# Biomolecular Correlation in Physical and Sequence Space

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Vom Fachbereich Biologie der Technischen Universität Darmstadt zur Erlangung des akademischen Grades eines *Doctor rerum naturalium* genehmigte Dissertation von

Dipl.-Bioinform. Franziska Hoffgaard aus Leipzig

Berichterstatter: Prof. Dr. Kay Hamacher

Mitberichterstatter: Prof. Dr. Gerhard Thiel

Eingereicht am: 01.07.2011

Mündliche Prüfung am: 24.08.2011

Darmstadt 2011 - D17

*Die Erinnerung ist ein Fenster,*
*durch das ich dich sehen kann,*
*wann immer ich will.*

# Summary

Investigating correlations is the key to understanding the nature of biological systems. In general, correlations describe the relationship between data sets or specific characteristics of data. To investigate correlations among and within biomolecules we discussed two complementary approaches to advance the understanding of evolution. Mutational dynamics can mainly be seen in the space of sequences whereas the altered phenotype is selected in the biophysical realm. By mutual information, an information-theoretical measure, we can identify potentially coevolving nucleotide or amino acid positions from a set of sequences combined into a multiple sequence alignment. In the biophysical realm, the mechanics of a biomolecule, which is important for its structure and function, is examined by various methods. Since molecular dynamics simulations and normal mode analysis are computationally expensive approaches, coarse-grained protein representations such as elastic network models have been developed. We used such protein models, particularly the Gaussian and the anisotropic network model, to jugde the importance of single residues or amino acid contacts on the dynamics of the biomolecule or distinct portions.

- In this thesis, we applied this analysis to distinct sets of hammerhead ribozyme sequences of type I and III to reveal coevolutionary hot spots shared among the different sequences. We observed a weaker coevolution of ribozymes originating from prokaryotes and eukaryotes compared to viroid sequences. Additionally, we obtained signals between helical stems I and II which is well-known from experiments. However, we noticed a coevolutionary connection between stems I and III throughout all sets of sequences that have not been reported yet.

- We applied an established protocol to a structural model of the small viral potassium channel Kcv, where we deleted single contacts and measured the resulting change in dynamics using the Frobenius norm. Here, we observed a mechanical connection of N- and C-terminal residues, whereas the selectivity filter seems almost mechanically uncoupled to the rest of the channel. A similar study was performed for the acetylcholinesterase as well where we additionally correlated mechanical changes with coevolutionary information. By means of coarse-grained protein models, we proposed a protocol for the Kcv to identify the transition from a functional to a non-functional channel upon N-terminal deletions.

- Furthermore, we utilized reduced molecular models to derive amino acid specific interaction constants directly from a set of protein structures obtained from e.g. from molecular dynamics simulations. To this end, we examined the performance of three approaches to retrieve the input parameters from an artificially constructed system. As it turned out, semidefinite programming is an efficient method for this task and was employed for a realistic application as well.

# Zusammenfassung

Der Schlüssel zum Verständnis der Natur von biologischen Systemen ist die Untersuchung von Korrelationen. Im Allgemeinen dienen Korrelationen der Beschreibung von Zusammenhängen von Datensätzen oder gar von spezifischen Eigenschaften, die diesen Daten zugrundeliegen. Um Korrelationen von Biomolekülen zu untersuchen, haben wir zwei komplementäre Ansätze diskutiert, um das Verständnis von Evolution voranzutreiben. Der Sequenzraum eines Biomoleküls ist Mutationen unterworfen, wohingegen sich der veränderte Phänotyp im biophysikalischem Raum bewähren muss. Mit der Mutual Information, einem der Informationstheorie entstammendem Maß, ist es möglich für ein Sequenzalignment potentiell coevolvierende Nukleotid- oder Aminosäurepositionen zu identifizieren. Im biophysikalischen Raum ist besonders die Mechanik des Biomoleküls von Interesse, weil sie sowohl für die Struktur als auch für die Funktion eine wichtige Rolle spielt. Da Moleküldynamik-Simulationen und Normalmodenanalysen sehr rechenintensive Ansätze sind, um die Mechanik zu untersuchen, sind in den letzten Jahren reduzierte Proteinmodelle entwickelt worden. Wir verwenden solche reduzierten Modelle, im Besonderen Gauß und anisotropische Netzwerkmodelle, um Rückschlüsse auf die Bedeutung einzelner Residuen oder Aminosäurekontakte für the Dynamik des Moleküls oder einzelner Regionen zu ziehen.

- In dieser Arbeit haben wir eine solche Analyse für unterschiedliche Sequenzdatensätze von Hammerhead-Ribozymen von Typ I und III durchgeführt um evolutionär wichtige Verbindungen aufzudecken, die diesen Ribozymen gemein sind. Dabei haben wir schwächere coevolutionäre Signale für die prokaryotische und eukaryotische Ribozyme festgestellt im Vergleich mit viroiden Sequenzen. Weiterhin konnten wir Signale zwischen den Helices I und II auffinden, die bereits in experimentellen Studien nachgewiesen wurden. Allerdings stellten wir für alle Datensätze Coevolution zwischen den Helices I und III fest, für die es bisher keine Hinweise gab.

- Ein etabliertes Protokoll wurde auf das Strukturmodell des kleinen viralen Kaliumkanals Kcv angewendet, wobei einzelne Verbindungen gelöscht wurden und die resultierende Veränderung in der Mechanik mittels der Frobenius Norm gemessen wurde. Hierbei beobachteten wir eine Verbindung zwischen N- und C-terminalen Residuen, wohingegen der Selektivitätsfilter vom restlichen Kanal mechanisch entkoppelt zu sein scheint. Eine ähnliche Studie wurde für Acetylcholinesterase durchgeführt, wobei wir die mechanischen Veränderungen zusätzlich noch mit Coevolutionsdaten überlagerten. Mit Hilfe der reduzierten Proteinmodelle haben wir zusätzlich ein Protokoll vorgestellt, mit dem es möglich ist den Übergang von einem funktionalen zu einem nicht-funktionalen Kanal aufgrund N-terminaler Deletionen zu detektieren.

- Weiterhin wurden reduzierte Modelle verwendet, um aminosäurespezifische Interaktionsstärken direkt aus einer Menge von Strukturen, die zum Beispiel aus Moleküldynamik-Simulation gewonnen wurden, zu bestimmen. Aus diesem Grund haben wir für drei An-

sätze untersucht, inwiefern sie in der Lage sind, gegebene Parameter eines artifiziellen Systems zu rekonstruieren. Semidefinite programming hat sich als effiziente Methode für diese Aufgabe herausgestellt und wurde anschließend angewendet, um die Parameter für eine realistische Anwendung zu identifizieren.

# Contributions

During the research for this PhD thesis, I worked on several projects. Results of those have been published for. Here, I will describe my contributions to the publications.

1. Weil, P; Hoffgaard, F; Hamacher, K (2009) Estimating sufficient statistics in co-evolutionary analysis by mutual information. *Comp Biol Chem* 33:440.

   Here, I participated in devising the study, provided implementation support and was involved in writing the manuscript.

2. Boba, P; Weil, P; Hoffgaard, F; Hamacher, K (2010) Co-evolution in HIV enzymes. *Proc. of BIOINFORMATICS 2010*, A. Fred, J. Filipe, H. Gamboa (eds.), p. 39.

   For this contribution, I provided routines for the computational analysis and was involved in devising the study.

3. Pape, S; Hoffgaard, F; Hamacher, K (2010) Distance-dependent classification of amino acids by information theory. *Proteins* 78:2322.

   For this study, I prepared the structural data set to be used and was involved in devising the study. Additionally, I participated in writing the manuscript.

4. Hoffgaard, F; Weil, P; Hamacher, K (2010) BioPhysConnectoR: Connecting sequence information and biophysical models. *BMC Bioinformatics* 11:199.

   For the software that has been published along with this contribution, I did almost all of the coding. Furthermore, I parallelized implemented protocols in order to enhance the efficiency of the supplied code. In addition, I participated in writing the manuscript.

5. Gebhardt, M; Hoffgaard, F; Hamacher, K; Kast, SM; Moroni, A; Thiel, G (2011) Membrane anchoring and interaction between transmembrane domains is crucial for $K^+$ channel function. *J Biol Chem* 286:11299.

   In this study, I performed a network-based analysis using ANMs (see section 1.1) to judge the importance of the $\pi$-$\pi$ stacking interactions of residues F30 and H83 in the Kcv channel. Furthermore, I participated in writing the manuscript.

6. Strunk, T; Hamacher, K; Hoffgaard, F; Engelhardt, H; Zillig, MD; Faist, K; Wenzel, W; Pfeifer, F (2011) Structural model of the gas vesicle protein GvpA and analysis of GvpA mutants *in vivo*. *Mol Microbiol* 81:56.

   Here, I accomplished the molecular dynamics simulations to verify the stability of the predicted GvpA structure as well as the analysis of the simulated time series.

7. Weißgraeber, S; Hoffgaard, F; Hamacher, K (2011) Structure-based, biophysical annotation of molecular coevolution of acetylcholinesterase. *Proteins*, accepted.

   Here, I devised the study and provided implemented routines to perform parts of the analysis. Additionally, I participated in writing the manuscript.

8. Wächter, M; Hamacher, K; Hoffgaard, F; Widmer, S; Goesele, M (2011) Is your permutation algorithm unbiased for $n \neq 2^m$? *9th International Conference on Parallel Processing and Applied Mathematics (PPAM 2011)*, accepted.

   For this contribution, I devised parts of the study, co-advized a student and participated in writing the manuscript.

# Contents

# 1 Molecular Biophysics

## 1.1 Elastic Network Models

Proteins are essential biological macromolecules that participate in virtually all processes within a cell. During protein biosynthesis amino acids are connected along the protein backbone by peptide bonds. The linear sequence of amino acid residues is the primary structure of the protein. The secondary structure contains information about motifs, such as $\alpha$-helices or $\beta$-sheets. A protein is stabilized in its tertiary structure by non-local bonds of proximate residues, e.g. disulfide bonds or salt bridges. Some proteins assemble into complexes forming the quarternary structure. Despite the fact that several thousand protein structures have been elucidated and deposited in the Protein Data Bank (PDB) [Berman *et al.*, 2000], the understanding of protein folding and function is still insufficient. Proteins fold spontaneously within micro- to milliseconds during or after biosynthesis, which is contradictory to the time a sequential sampling of the protein's conformation space would take. Protein folding pathways, i.e. a funnel-shaped energy landscape [Bryngelson *et al.*, 1995; Dill & Chan, 1997; Leopold *et al.*, 1992] that is biased towards the native state, are a potential conclusion from the Levinthal paradox [Levinthal, 1968].

Among the leading theoretical methods to investigate protein structure properties as well as protein folding and stability are molecular dynamics simulations (MD) and normal mode analysis (NMA). MD simulations have been widely used to study the stability of modeled proteins and protein complexes [Strunk *et al.*, 2011] as well as to obtain insights into their function [Tayefeh *et al.*, 2007]. Proteins, ribonucleic/deoxyribonucleic acids (RNA/DNA) and respective complexes are placed in a simulation box that also contains solvents to provide a realistic scenario. Ions, ligands and other biomolecules may be included in the simulation as well. As can be seen in Fig. 1.1 current simulation efforts are below the millisecond scale, where protein folding takes place. Nevertheless, fast events like transport in ion channels can be observed.

Since internal motions of a protein are known to play an important role in its function, NMA was employed in numerous studies as a tool to directly investigate vibrational motions using

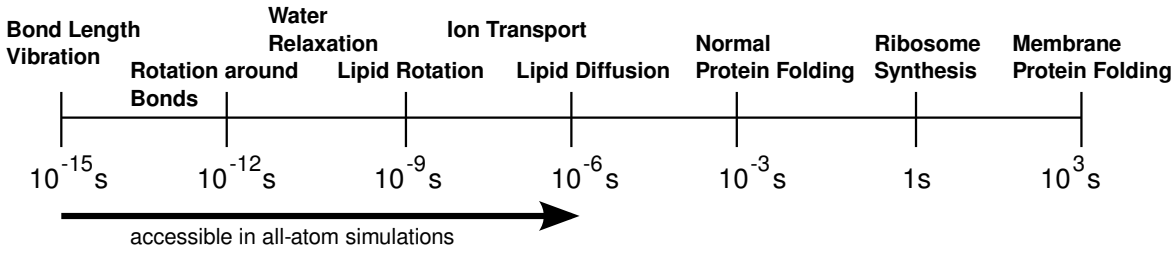



**Figure 1.1.:** Time scales of molecular dynamics simulations (adapted from Lindahl [2008]).

harmonic approximation. Amongst others, hinge and shear motions necessary for catalytic reactions were identified [Brooks & Karplus, 1985; Levitt *et al.*, 1985]. Although NMA is less time-consuming than MD simulations, it requires the protein to be thoroughly energy minimized. Due to computational limitations, energy minimization and NMA itself are often applied to the protein *in vacuo* which may lead to undesired side effects.

As an alternative, Tirion [1996] proposed to replace the complex potentials that are used in MD simulations and NMA approaches by a pairwise Hookean potential which is controlled by a single parameter. Surprisingly, despite its simplicity the potential was able to reproduce complex vibrational properties of biomolecules represented by the slow elastic modes. Theoretical temperature factors, also referred to as B-factors or Debye-Waller factors, showed very good agreement with the experimentally obtained crystallographic temperature factors which indicates no need for additional parametrization of the potential.

Based on Tirion's seminal work, Gaussian network models (GNM) have been developed by Bahar *et al.* [1997]. Here, a protein in its folded state is considered equivalent to an elastic network. In this coarse-grained approach, $C_\alpha$ atoms are identified with the junctions of the respective network and assumed to undergo Gaussian distributed fluctuations around equilibrium position. Interactions between closely located $C_\alpha$ pairs are modeled as harmonic springs with a single parameter $\gamma$ of the Hookean pairwise potential. Equilibrium correlations between fluctuations $\Delta R_i$ and $\Delta R_j$ of two $C_\alpha$ atoms $i$ and $j$ are defined as [Haliloglu *et al.*, 1997]:

$$\langle \Delta R_i \cdot \Delta R_j \rangle = \frac{k_B T}{\gamma} [\Gamma^{-1}]_{ij} \tag{1.1}$$

where $k_B$ and $T$ are the Boltzmann constant and the absolute temperature, respectively. $\Gamma$ is the symmetric Kirchhoff matrix, the subscript $ij$ indicates the $ij$th element. The Kirchhoff matrix describes the connectivity of the junctions of the elastic network, namely the $C_\alpha$ atoms, and is derived as follows:

$$\Gamma_{ij} = \begin{cases} -1 & \text{if } i \neq j \text{ and } R_{ij} \leq r_c \\ 0 & \text{if } i \neq j \text{ and } R_{ij} > r_c \\ -\sum_{i,i\neq j} \Gamma_{ij} & \text{if } i = j \end{cases} \tag{1.2}$$

$R_{ij} = |R_i - R_j|$ denotes the spatial separation of the $C_\alpha$ atoms of residues $i$ and $j$ in the native state. The range of non-bonded interactions is bounded by the cutoff distance $r_c$, e.g. $r_c = 7$ Å is a reasonable choice to include nearby residue pairs within a first interaction shell. Fig. 1.2 illustrates the effect of the cutoff distance on the number of considered interactions for the bovine pancreatic trypsin inhibitor (BPTI) [Wlodawer *et al.*, 1987].

Since the Kirchhoff matrix is positive semidefinite, i.e. the smallest eigenvalue is equal to zero due to a single degree of freedom, the Moore-Penrose pseudoinverse [Moore, 1920; Penrose, 1955] needs to be used in Eq. 1.1. It can be expressed as:

$$\Gamma^{-1} = U(\Lambda^{-1})U^T \tag{1.3}$$

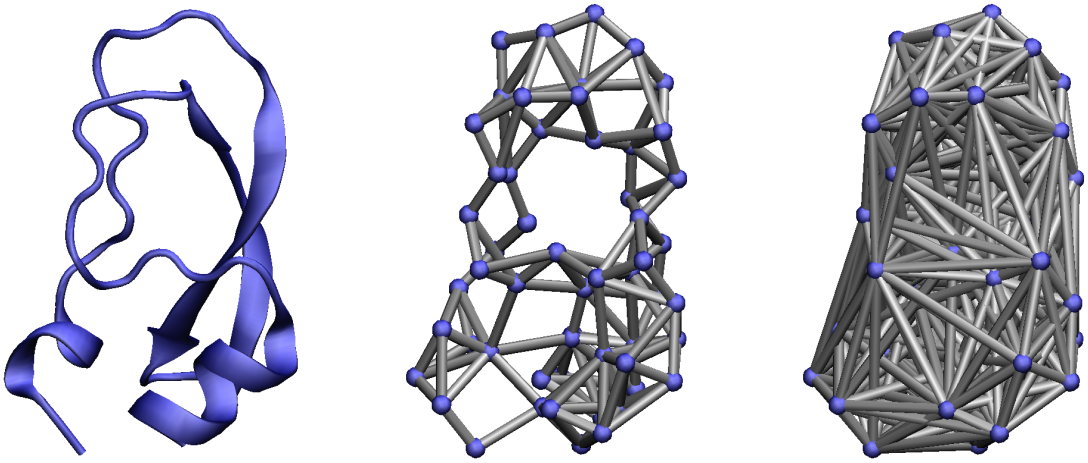


**Figure 1.2.:** The choice of the distance cutoff $r_c$ limits the range of interactions and is illustrated for the bovine pancreatic trypsin inhibitor [Wlodawer *et al.*, 1987] (left). Setting $r_c = 6$ Å and thus including only the most closely located residue pairs within a first interaction shell leads to a sparsely connected elastic network (middle). The network becomes more dense, if the distance cutoff is further increased. An interaction range of 13 Å contains also residue pairs of the second interaction shell (right). Graphics were generated using VMD [Humphrey *et al.*, 1996].

$\Lambda$ is a diagonal matrix containing the singular values $\lambda_i$ of $\Gamma$ and $U$ is an orthogonal matrix containing the respective singular vectors in each column. The inverse matrix can thus be decomposed as sum of contributions of the individual modes:

$$\Gamma^{-1} = \sum_{k=2}^{n} \frac{u_k u_u^T}{\lambda_k} \tag{1.4}$$

Mean square fluctuations of $C_\alpha$ atoms and, hence, theoretical B-factors can be derived from the diagonal entries of $\Gamma^{-1}$. Furthermore, the off-diagonal entries contain cross-correlations between the fluctuations of the $C_\alpha$ atoms. GNMs are capable to reproduce experimentally obtained temperature factors and, as a coarse-grained model, facilitate the analysis of slow vibrational modes of (large) macromolecules. Bahar *et al.* [1998] investigated the relation of slow and fast motions to function and stability. The slowest modes corresponding to small eigenvalues $\lambda_i$ are primarily associated with collective dynamics of the tertiary structure of a protein, and may thus (not exclusively) be regarded relevant for biological function. Residues active in the fastest modes are crucial for the structure or the stability of the protein in its native state. GNMs have been successfully employed in various studies [Bahar *et al.*, 1998, 1997; Demirel *et al.*, 1998; Erman & Dill, 2000; Haliloglu *et al.*, 1997] to elucidate dynamics of proteins and implications for biologically relevant features [Bahar *et al.*, 1999; Bahar & Jernigan, 1999; Erman, 2006; Shrivastava & Bahar, 2006].

Now, in GNMs all fluctuations are implicitly assumed to be isotropic, whereas in reality we are concerned with anisotropic motions. To this end, Atilgan *et al.* [2001] proposed anisotropic networks (ANM) as extension of GNMs to assess the directions of motions, which can directly be

relevant for biological mechanisms. The counterpart of the Kirchhoff matrix in GNMs as defined in Eq. 1.2 is the Hessian matrix $H$, which is a symmetric matrix with dimension $3N \times 3N$ for $N$ residues of a protein. The Hessian matrix is obtained as second derivative of the harmonic potential $V$ (see Eq. 1.7) that describes interactions within a distance $r_c$, a reasonable choice of the distance cutoff is $r_c = 13$ Å [Atilgan *et al.*, 2001; Hamacher & McCammon, 2006]. It is composed of $N \times N$ super elements of the following form:

$$H_{ij} = \begin{bmatrix} \partial^2 V/\partial X_i \partial X_j & \partial^2 V/\partial X_i \partial Y_j & \partial^2 V/\partial X_i \partial Z_j \\ \partial^2 V/\partial Y_i \partial X_j & \partial^2 V/\partial Y_i \partial Y_j & \partial^2 V/\partial Y_i \partial X_j \\ \partial^2 V/\partial Z_i \partial X_j & \partial^2 V/\partial Z_i \partial X_j & \partial^2 V/\partial Z_i \partial Z_j \end{bmatrix} \tag{1.5}$$

where $X_i$, $Y_i$ and $Z_i$ are the components of the displacement vector of residue $i$. Decomposing the Hessian matrix yields $3N - 6$ non-zero eigenvalues due to three rotational and three translational degrees of freedom. Its pseudoinverse $C$, also referred to as covariance matrix, is computed via singular value decomposition (SVD) [Press *et al.*, 1992] as sum of contributions of the individual modes similarly as the pseudoinverse of the Kirchhoff matrix (see Eq. 1.4).

$$C = \sum_{k=7}^{n} \frac{u_k u_u^T}{\lambda_k} \tag{1.6}$$

The covariance matrix contains information about correlated motions amongst residues and the respective directions. The B-factor of a residue $i$ is derived by summing up the diagonal elements of the respective super element $H_{ii}$. Again, theoretical temperature factors show striking resemblance to the crystallographic ones [Atilgan *et al.*, 2001; Doruker *et al.*, 2002].

Both GNM and ANM presented thus far do not invoke side chain specificity other than NMA or MD simulations based on empirical force fields. Hamacher & McCammon [2006] proposed a contact potential that accounts for amino acid type specific energies for interacting residue pairs. The potential $V$ of a biomolecule in terms of the extended ANM (eANM) is obtained as

$$V = \alpha a^{-2} \left[ \frac{a^2 K}{2} \sum_i (R_{i,i+1} - R^0_{i,i+1})^2 + \sum_{(i,j) \in C\prime} \kappa_{ij} (R_{ij} - R^0_{ij})^2 \right] \tag{1.7}$$

with $\alpha$ and $a$ being scaling constants. The potential consists of a term describing covalent contacts between direct neighbors and another term modeling non-covalent contacts of residues being closer than a certain distance cutoff $r_c$, interacting residue pairs are contained in the set $C\prime$. $K$ assigns a uniform interaction potential to all peptide bonds, whereas $\kappa_{ij}$ may include amino acid specificity. $R^0_{ij}$ is the distance of the $C_\alpha$ atoms of amino acid $i$ and $j$ in the native state of the protein. Obviously, the eANM, which includes the ANM as a special case, achieved better correlations with the experimentally obtained B-factors than the ANM without amino acid specific terms.

## 1.2 Extensions to Network Models

Coarse-grained approaches, such as elastic network models (ENM) of which GNM and ANM are special cases, have been developed to facilitate and accelerate the analysis of large biological systems. Dynamics of biological complexes consisting of a multitude of subunits, e.g. the ribosome, remained inaccessible due to limitations of computational time and resources. In the past, diverse coarse-grained protein models have been developed [Chu & Voth, 2007; Jeong *et al.*, 2006; Kundu *et al.*, 2002; Kuriyan & Weis, 1991; Lyman *et al.*, 2008; Maragakis & Karplus, 2005] based on different assumptions on structural, chemical and physical properties to tackle the problem of predicting protein dynamics. This requires optimization of model parameters and assumptions by still benefitting from the advantages of reduced protein descriptions.

In the following, we will discuss potential extensions of ENMs.

### 1.2.1 Alternative Contact Definitions - TanH

ENMs [Atilgan *et al.*, 2001; Haliloglu *et al.*, 1997; Hamacher & McCammon, 2006] as described in section 1.1 are a coarse-grained description of proteins. Residues are represented as a bead placed on the respective $C_\alpha$ position, and interactions between residues are modeled as harmonic springs between residues in contact. Contacts are defined using a distance cutoff $r_c$, i.e. two residues $i$ and $j$ are in contact if their spatial distance $R_{ij}$ is not larger than $r_c$. As result, we obtain a binary contact map CM with

$$\mathrm{CM}_{ij} = \begin{cases} 1 & \text{if } R_{ij} \leq r_c \\ 0 & \text{otherwise} \end{cases} \tag{1.8}$$

In ENMs such strict cutoff definitions are applied to crystal structures whose resolution is limited, nearly all of the about 63,000 experimentally derived structures deposited in the Protein Data Bank (PDB) [Berman *et al.*, 2000] are solved by X-ray or electron microscopy methods at a resolution of $\geq 1$ Å. Hence, a potential interaction can get lost as result of a minor displacement. Taking all interactions into account would solve this issue but neglect the benefit of a coarse-grained model. Therefore, more realistic contact definitions need to be considered.

Weighted-interaction ENMs [Hinsen *et al.*, 2000; Riccardi *et al.*, 2009] based on physical arguments have been introduced in past years. Hinsen *et al.* [2000] derived a force constant matrix for a $C_\alpha$ model from an all-atom model under the assumption that for any displacement of the $C_\alpha$ atoms the potential energy is minimized by the respective movement of the other atoms. Two distance categories are proposed to compute the distance dependent force constant $k$ to discriminate residue pairs along the backbone (being closer than 4 Å) and all other pairs.

$$k = \begin{cases} R_{ij} \cdot 8.6 \cdot 10^5 \text{ kJ mol}^{-1}\text{nm}^{-3} - 2.39 \cdot 10^5 \text{ kJ mol}^{-1}\text{nm}^{-2} & \text{for } R_{ij} < 4 \text{ Å} \\ R_{ij}^{-6} \cdot 128 \text{ kJ mol}^{-1}\text{nm}^4 & \text{for } R_{ij} \geq 4 \text{ Å} \end{cases} \tag{1.9}$$

In spite of the absolute scale, a good approximation of diverse potential energy surfaces (PES) was achieved by introducing a scaling factor. Hence, introducing distance dependent force constants for residue interactions that are modeled as harmonic springs may improve accuracy of coarse-grained protein models.

Yet another approach to investigate the flexibility and, hence, the dynamics of a protein was introduced by Halle [2002]. The study presented a distinct correlation between atomic mean square displacements, which are proportional to B-factors, and the number of non-covalent neighbors within a certain distance $r_c$. Avoiding time-consuming computations, e.g. matrix inversion or diagonalization, the so-called local density model is directly applicable to predict temperature factors. An extension of this model was proposed by Lin *et al.* [2008], the weighted contact number (WCN) model. Here, the number of contacts is weighted by the square of the reciprocal distance of contacting residues. Because of the fast decay of the prefactor, we are now concerned with a more simplified model that avoids the definition of a distance cutoff and is based on the distance of non-bonded residue pairs, only. Although no mechanical model or associated potential function is assumed, the WCN model can produce accurate B-factor profiles. Hence, taking interaction ranges into account may enhance the understanding of the mechanics of a protein [Hinsen, 2009].

Those results encouraged Yang *et al.* [2009] to invoke distance dependence for ENMs, referred to as parameter-free ENM (pfENM). On the assumption that all residues interact with each other the Kirchhoff (GNM, Eq. 1.2) and the Hessian (ANM, Eq. 1.5) matrix are inversely weighted by the respective squared distances of all residue pairs. Hence, the quite arbitrary value of a cutoff distance, whose optimal values vary for different proteins, is no longer necessary. pfENMs in comparison to ENMs allow a more accurate prediction of local fluctuations with respect to experimental data. Since pfENMs incorporate stronger long-range cohesion effects, discrete domains are not allowed to move sufficiently to capture larger conformational transitions that may occur in the context of biological function. If the power of the inverse distance dependence is increased up to 6 or 8, i.e. short-range interactions are strengthened, even conformational changes can be reproduced well.

Inspired by the presented studies in the field of ENMs concerning distance-weighted contact definitions, we investigated a further contact definition based on spatial separation of residues. ENMs invoking distance dependent force constants lead to major improvements in the prediction of local fluctuations in contrast to cutoff-based ENMs using a binary contact definition (see Eg. 1.8), but revealed shortcomings in modeling larger conformational transitions. Hence, we define a smooth contact function $f_c$, that weights interactions within a defined interval according to the distance of the corresponding residue, but long-range interaction may be excluded from consideration:

$$f_c = \frac{1}{2}(1 - \tanh(a \cdot R_{ij} - b)) \qquad \in [0,1] \tag{1.10}$$

The shape of the function is controlled by two parameters $a$ and $b$ (see Fig. 1.3). The idea is that interactions beyond the distance cutoff $r_c$ are not omitted, their influence is decreased depending on the separation distance $R_{ij}$ of the respective residues $i$ and $j$. Thus, we define a distance cutoff $r_c$ that describes the inflection point of the contact defining function $f_c$. Force constants of harmonic springs connecting residues separated by $R_{ij} \in [r_c - w, r_c + w]$ with

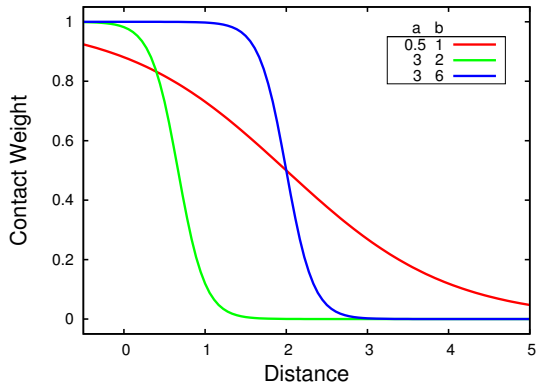



**Figure 1.3:** We use $f_c = \frac{1}{2}(1 - \tanh(a \cdot R_{ij} - b))$ as contact defining function. The influence of the parameters $a$ and $b$ as defined in Eq. 1.11 on the shape of the function is illustrated. Larger values of $a$ lead to a sharper transition, whereas $b/a$ defines the position of the inflection point.

a constant $w$ are weighted according to $f_c$. Hence, long-range interactions are included or omitted depending on the parameters of $f_c$. The parameters $a$ and $b$ can be derived from the definition of the inflection point of $f_c$ as $r_c = a/b$:

$$a = \frac{1}{w} \qquad b = a \cdot r_c \tag{1.11}$$

To avoid numerical issues, we set all contact values smaller than $10^{-9}$ to zero.

We defined a set of 74 proteins[1] that consist of a single chain each with chain length $l$ ranging from 31 to 639 with a mean of about 200 amino acids. For each protein, an ANM as described in section 1.1 is computed with the following contact definitions:

- Binary contact scheme with cutoff distance $r_c$, see Eq. 1.8 (*binary*)
- Inverse weighting scheme according to Yang *et al.* [2009] (*inverse*)
- Weighting according to $f_c$ with cutoff distance $r_c$ and interval $w$, see Eq. 1.10 (*tanh*)

Furthermore, we employed two weighting schemes for parametrization of interactions of contacting residue pairs.

- All covalent and non-covalent residue pairs are connected by harmonic springs with a unique force constant. Hence, weights for peptide bonds $K$ and each non-covalent interaction $\kappa_{ij}$ of amino acids of type $i$ and $j$ are set to $K = \kappa_{ij} = 1$ RT/Å$^2$.
- Covalent bonds are modeled as harmonic springs with a force constant of $K = 82$ RT/Å$^2$ as proposed by Hamacher & McCammon [2006]. Non-covalent interactions are modeled by the weighting scheme that was put forward by Miyazawa & Jernigan [1996] (MJ).

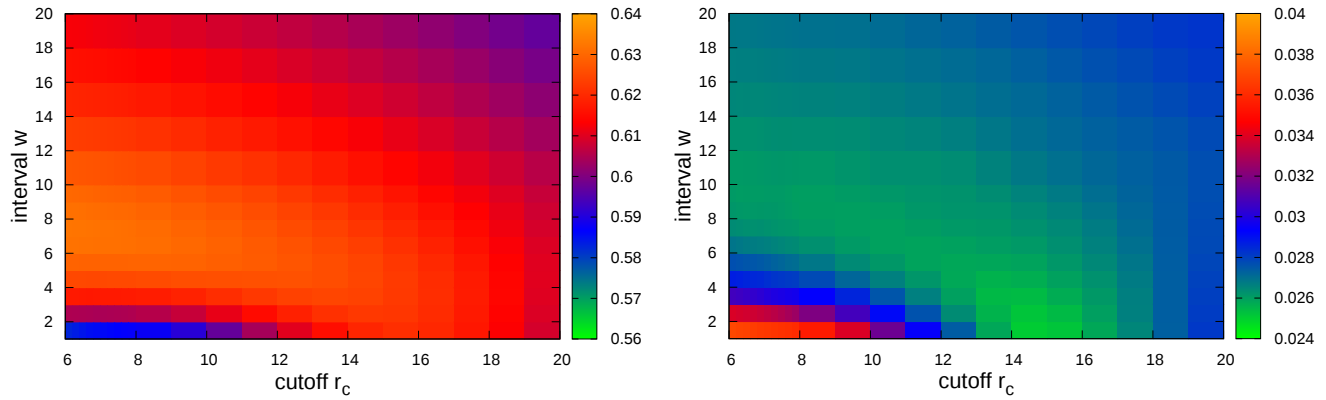
Hence, we are able to judge on the influence of sequence specific potentials. The quality of the contact modeling approaches is evaluated by correlating predicted to experimentally derived B-factors. Local fluctuations, described by B-factors, are derived from diagonal entries of the

[1] Protein Set A (see Appendix A.1)

| parameter | domain | increment |
|---|---|---|
| $r_c$ [Å] | 6 – 8 | 0.2 |
| | 9 – 30 | 1 |
| | 32 – 50 | 2 |

| parameter | domain | increment |
|---|---|---|
| $w$ [Å] | 0.2 – 1 | 0.2 |
| | 2 – 10 | 1 |
| | 12 – 30 | 2 |
| | 35 – 100 | 5 |

**Table 1.1.:** Variation of Parameters $r_c$ and $w$, that are used for contact definition. Note that $w \geq 0.2$ Å leads to valid results for all proteins only for a cutoff distance $r_c \geq 8$ Å.




**(a)** mean **(b)** variance

**Figure 1.4.:** Mean and variance of the Spearman correlation coefficient of computed and experimental B-factors for varying parameters of the contact defining function $f_c$ (see Eq. 1.10). All interactions are modeled with a unique force constant.

covariance matrix. For both cutoff-based approaches, the model parameters $r_c$ and $w$ are varied (see Tab. 1.1). Note that for small cutoff distances, i.e. $r_c < 8$ Å, we need to define $w \geq 0.4$ Å to avoid additional singularities, i.e. increasing the size of the null space.

Pearson and Spearman correlation coefficients of theoretical and experimental B-factors are computed for each $r_c$-$w$ combination. Additionally, mean and variance of the correlation coefficients of the protein set are determined. The results for the Spearman coefficient, which in contrast to the Pearson coefficient reveals non-linear relations as well, are shown in Fig. 1.4 for $r_c \in [6\text{ Å}, 20\text{ Å}]$ and $w \in [0.4\text{ Å}, 20\text{ Å}]$. Higher values of $r_c$ and $w$ lead to worse correlations (data not shown). Clearly, we can detect a band of parameter settings showing a high average correlation ($\approx 0.64$) by small variances. Interestingly, smaller cutoff distances ($r_c \in [6\text{ Å}, 10\text{ Å}]$) in combination with small to medium intervals ($w \in [6\text{ Å}, 15\text{ Å}]$) are to be preferred. The results for the Pearson correlation coefficient reveal the same trends (data not shown).

Fig. 1.5 shows the comparison of the Spearman correlations obtained for the different contact definitions. Since we have sampled various parameter settings for cutoff-based models, we determined those settings that show the highest Spearman correlations averaged over all proteins under consideration (see Tab. 1.2). The scatter plots which compare the prediction capabilities of the contact models have been generated using the obtained parameters. Points below the

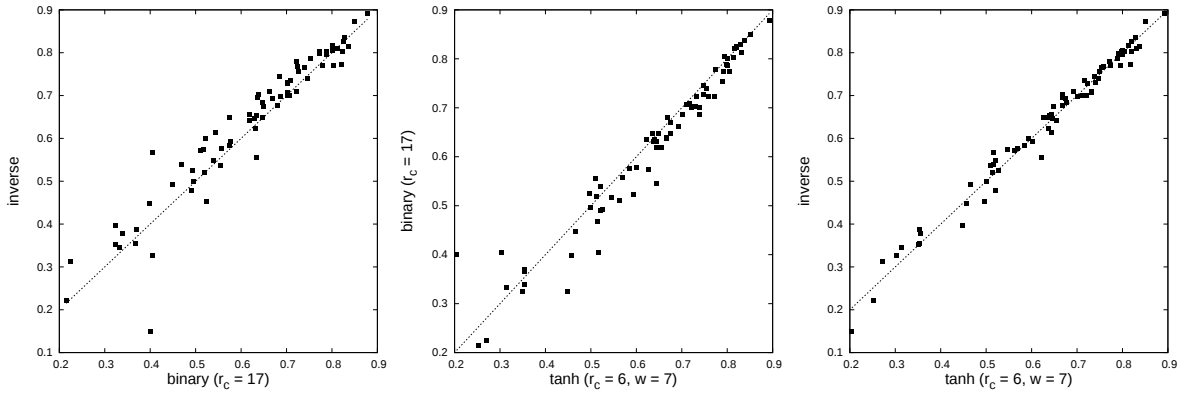



**Figure 1.5.:** Comparison of the diverse contact definitions. For each protein in the set the Spearman correlation coefficient of experimental and predicted B-factors has been computed. The results for the different contact definitions are plotted against each other. Results for the cutoff-based methods are shown for the parameter setting that performed best on average.

| binary model | tanh model | |
|---|---|---|
| $r_c = 17$ Å | $r_c = 6$ Å | $w = 7$ Å |

**Table 1.2.:** Parameter(s) of different contact models that show the highest Spearman correlation of predicted and experimental B-factors.

straight line indicate a better performance of the contact model plotted on the x-axis, otherwise the y-axis model achieved a better prediction. Both the contact weighting scheme proposed by Yang *et al.* [2009] and the contact model based on $f_c$ improve the ability of ANMs to reproduce B-factors.

The analysis was repeated for amino acid specific potentials as well. We observed worse performance in B-factor prediction for all contact schemes, i.e. the average correlation is reduced by 0.03–0.06. Invoking amino acid specificity, the tanh contact definition yielded the highest correlation on average (measured as Spearman correlation coefficient), whereas the binary contact matrix (Eq. 1.8) performed better than the scheme proposed by Yang *et al.* [2009] (data not shown).

In our study, we discussed three approaches how to define a contact between amino acids for an ANM of a protein (see section 1.1). The quality of ANMs is usually judged by the correlation of experimentally determined and predicted B-factors that describe local fluctuations of atoms. If we apply a unique parametrization to all interactions, i.e. we do not distinguish between covalent and any non-covalent contact, we found that defining contacts proportional to $1/R_{ij}^2$ and our proposed weighting according to $f_c$ yield similar correlations of predicted and experimental B-factors. Both methods are an improvement of the originally proposed binary contact scheme. Adding sequence specificity to the interaction potentials leads to decreased correlation coefficients for either method, and is, hence, not recommended. The results of our model can be improved, if we determine the best parameters for each protein other than using those parameters that performed best for the protein set on average as we did in this study.

### 1.2.2 B-Factors without a Mechanical Model

Apart from ENMs we delineated in previous sections, Shih *et al.* [2007] developed a simple method to derive protein dynamics directly from the tertiary structure of the protein without a mechanical model. Their work is grounded on the observation, that atoms which are buried inside the protein fluctuate less around their equilibrium position than atoms at the surface of the protein do. Obviously, there is a linear relationship between atomic fluctuations and squared distances from the protein's center of mass. Although the knowledge of the amino acid sequence is not required, Shih *et al.* [2007] yielded results for B-factors and correlation matrices that are in excellent agreement with experimental data (crystallography, NMA). In the following, we investigate if adding sequence specificity will improve the B-factor prediction results of the presented simplistic approach. For this purpose, we defined a set of 960 monomeric proteins[2], that are classified according to diverse SCOP (Structural Classification of Proteins) [Murzin *et al.*, 1995] categories.

According to Shih *et al.* [2007], the fluctuations of an atom $i$ are proportional to its spatial distance ($r_i^{\,p}$) from the center of the protein $p$ in native state. Thus, the B-factor of atom $i$ is obtained as follows:

$$B_i \sim b_i = \frac{3}{8\pi^2}(\vec{r}_i^{\,p})^2 \tag{1.12}$$

where $B_i$ and $b_i$ are experimental and theoretical B-factors, respectively. Again, we are concerned with a coarse-grained protein model that takes only $C_\alpha$ atoms representing their respective amino acids into account. The distance of residue $i$ to the center of the protein $p$ is defined as

$$(\vec{r}_i^{\,p})^2 = (\vec{x}_i - \vec{c}_0^{\,p})^T(\vec{x}_i - \vec{c}_0^{\,p}) \tag{1.13}$$

with $\vec{c}_0^{\,p}$ being the center of the protein. The center of the protein can be computed as center of mass $\vec{c}_0^{\,p} = \vec{x}_0^{\,p}$ or as geometric mean $\vec{c}_0^{\,p} = \vec{y}_0^{\,p}$. The center of mass is computed using the $C_\alpha$ coordinates and the masses of the respective amino acids $m_i$ taken from JenaLib [Reichert *et al.*, 2000].

$$\vec{x}_0^{\,p} = \frac{\sum_{i=1}^{N_p} m_i \vec{x}_i}{\sum_{i=1}^{N_p} m_i} \qquad \vec{y}_0^{\,p} = \frac{1}{N_p}\sum_{i=1}^{N_p} \vec{x}_i \tag{1.14}$$

$N_p$ is the number of residues in protein $p$. In the following, we will refer to the model proposed by Shih *et al.* [2007] as *ShihB*.

Below, we will present two approaches based on these definitions, that invoke sequence specificity.

[2] Protein Set B (see Appendix A.2)

**Approach 1 (*asShihB*): Adding amino acid specificity.** Here, we introduce amino acid specific scaling factors $a_{t(i)}$ for residue $i$ of type $t(i)$. Hence, B-factors are computed as:

$$b_i = \frac{3}{8\pi^2} a_{t(i)} (\vec{r}_i^{\,p})^2 \tag{1.15}$$

Furthermore, we define the error function $e$, which has to be minimized. This function is a measure of the distance between predicted and experimental B-factors $B_i^{\,p}$.

$$e = \sum_{p=1}^{P} \sum_{i=1}^{N_p} \left( B_i^{\,p} - \frac{3}{8\pi^2} a_{t(i)} (\vec{r}_i^{\,p})^2 \right)^2 \tag{1.16}$$

To this end, we define a protein specific function $I_p(t)$, which yields a set containing all indices of residues of type $t$.

$$I_p(t) : \{1, \ldots, 20\} \longrightarrow \{i | t(i) = t\} \tag{1.17}$$

Thus, the error function in Eq. 1.16 can be rewritten as

$$e = c + \sum_{p=1}^{P} \sum_{t=1}^{20} \sum_{i \in I_p(t)} \left[ a_{t^*}^2 ((\vec{r}_i^{\,p})^2)^2 - 2 a_{t^*} (\vec{r}_i^{\,p})^2 B_i^{\,p} \right] \tag{1.18}$$

$$e = c + \sum_{t=1}^{20} \left[ a_{t^*}^2 \alpha_t - 2 a_{t^*} \beta_t \right] \tag{1.19}$$

With

$$a_{t^*} = \frac{3}{8\pi^2} a_t \quad \text{and} \quad \alpha_t = \sum_{p=1}^{P} \sum_{i \in I_p(t)} ((\vec{r}_i^{\,p})^2)^2 \quad \text{and} \quad \beta_t = \sum_{p=1}^{P} \sum_{i \in I_p(t)} (\vec{r}_i^{\,p})^2 B_i^{\,p} \tag{1.20}$$

the error function is minimized by

$$a_t = \frac{8\pi^2}{3} \frac{\beta_t}{\alpha_t} \quad \forall\ t = 1, \ldots, 20 \tag{1.21}$$

Hence, we determine $a_t$ for a set of proteins that consist of a single chain, only. The scaling factors may be correlated to amino acid specific properties.

**Approach 2 (*psShihB*): Adding protein specific translation and rotation.** Shih *et al.* [2007] computated protein specific scaling and translation constants $S_p$ and $A_p$ to improve B-factor prediction. Similarly to Eq. 1.16, we defined a new error function $e_2$ as

$$e_2 = \sum_{p=1}^{P} \sum_{i=1}^{N_p} (B_i^{\,p} - (S_p a_{t(i)} (\vec{r}_i^{\,p})^2 + A_p))^2 \tag{1.22}$$

The protein specific scaling and translation constants are computed as follows:

$$S_p = \frac{\sum_{i=1}^{N_p} {B_i}^p {b_i}^p - \frac{1}{N_p}\left(\sum_{i=1}^{N_p} {B_i}^p\right)\left(\sum_{i=1}^{N_p} {B_i}^p\right)}{\sum_{i=1}^{N_p} {b_i}^p {b_i}^p - \frac{1}{N_p}\left(\sum_{i=1}^{N_p} {b_i}^p\right)\left(\sum_{i=1}^{N_p} {b_i}^p\right)} \tag{1.23}$$

$$A_p = \frac{1}{N_p}\left(\sum_{i=1}^{N_p} {B_i}^p - S_p \sum_{i=1}^{N_p} {b_i}^p\right) \tag{1.24}$$

where $N_p$ is the number of amino acids in protein $p$. Furthermore, ${B_i}^p$ and ${b_i}^p$ are the experimental and predicted B-factors of protein $p$, respectively. We extend *asShihB* by introducing a protein specific scaling factor $S_p$ as well as a translation constant $A_p$.

In total, we need to determine $2P + 20$ unknown values for $A_p$, $S_p$ and $a_t$ that minimize the error function $e_2$. For this purpose, the conjugate gradient method as implemented in the GNU Scientific Library (GSL) [Galassi, 2009] was employed. By this method, all equations are minimized at once, but, however, the results strongly depend on the start values (data not shown). Note that we did not obtained results for all proteins in our set.

With *asShihB* we derived amino acid specific scaling factors that minimized the deviation of theoretical predicted and experimental B-factors of a protein set. Tab. 1.3 shows the scaling factors that were computed using two different protein center definitions. Apart from minor deviations, we observed the same constants, which allows us to focus on the geometric definition for further investigations. Interestingly, the values for different amino acids do not vary much either.

We furthermore investigated the convergence of the scaling factors $a_t$ for increasing numbers of proteins. The results are shown exemplarily for amino acids methionine (M), glutamine (Q) and tyrosine (Y) in Fig. 1.6. We observe a similar behavior for the other amino acid types as well. Large fluctuations occur for smaller number of proteins (up to 100), but they do not

| | A | C | D | E | F | G | H | I | K | L |
|---|---|---|---|---|---|---|---|---|---|---|
| *gm* | 0.785 | 1.025 | 0.805 | 0.873 | 0.864 | 0.772 | 0.947 | 0.790 | 0.799 | 0.857 |
| *mw* | 0.785 | 1.025 | 0.805 | 0.876 | 0.865 | 0.770 | 0.947 | 0.789 | 0.799 | 0.857 |
| | M | N | P | Q | R | S | T | V | W | Y |
| *gm* | 0.949 | 0.774 | 0.817 | 0.842 | 0.835 | 0.777 | 0.720 | 0.761 | 0.714 | 0.741 |
| *mw* | 0.953 | 0.774 | 0.817 | 0.844 | 0.837 | 0.777 | 0.719 | 0.759 | 0.718 | 0.746 |

**Table 1.3.:** Amino acid specific constants $a_t$ that were obtained to improve the B-factor prediction results of the method proposed by Shih *et al.* [2007]. Results are shown for different center definitions: *mw* indicates results computed for the center of mass, whereas *gm* indicates results that are computed for the geometric mean.

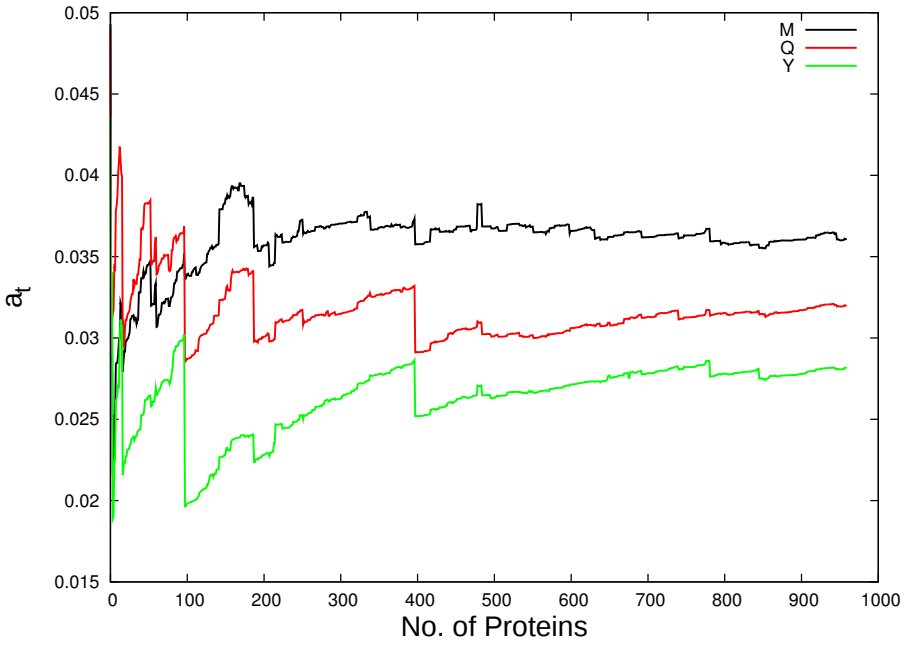



**Figure 1.6.:** Convergence plot for amino acid specific scaling factors as computed with *asShihB*. The results are presented for methionine (M), glutamine (Q) and tyrosine (Y).

vanish completely. Single proteins may change the results dramatically. This indicates that the derived amino acid specific constants strongly depend on the underlying protein set. Hence, the approach *asShihB* is applicable for homogeneous sets of proteins with similar properties and thus showing similar mechanics.

The prediction performance of *ShihB*, *asShihB* and *psShihB* is compared in Fig. 1.7 for the protein with PDB code 153L [Weaver *et al.*, 1995]. Clearly, the results of *ShihB* resemble the experimental B-factor the most. In a more general comparison, we compute the total deviation from experimental B-factors as well as the Spearman correlation coefficient for each modeling approach. We observe similar correlation coefficients for *ShihB* and *asShihB* averaged over all proteins ($\approx 0.52$), in contrast to a lower mean correlation of *psShihB* ($\approx 0.32$). Fig. 1.8 depicts the comparison of both *asShihB* and *psShihB* with *ShihB*. For each model, the sum of absolute deviations/errors $e$ of theoretical and experimental was computed. For each protein, we determined the difference of $\Delta e = e_0 - e_x$ with $x \in$ [*asShihB*,*psShihB*]; $e_0$ is the error obtained by *ShihB*. If $\Delta e > 0$, the method proposed by Shih *et al.* [2007] performs better than the respective extension by means of absolute differences. Both extensions are outperformed by the basic model. Note that only for the basic approach the scaling and translation terms are computed protein specific, whereas both extensions aimed to utilize general amino acid properties that can be found in all proteins to minimize the average deviation from experimental B-factors. The results for the proposed sequence and/or protein specific extensions may be improved for set of similar proteins to deduce mechanical features from structures.

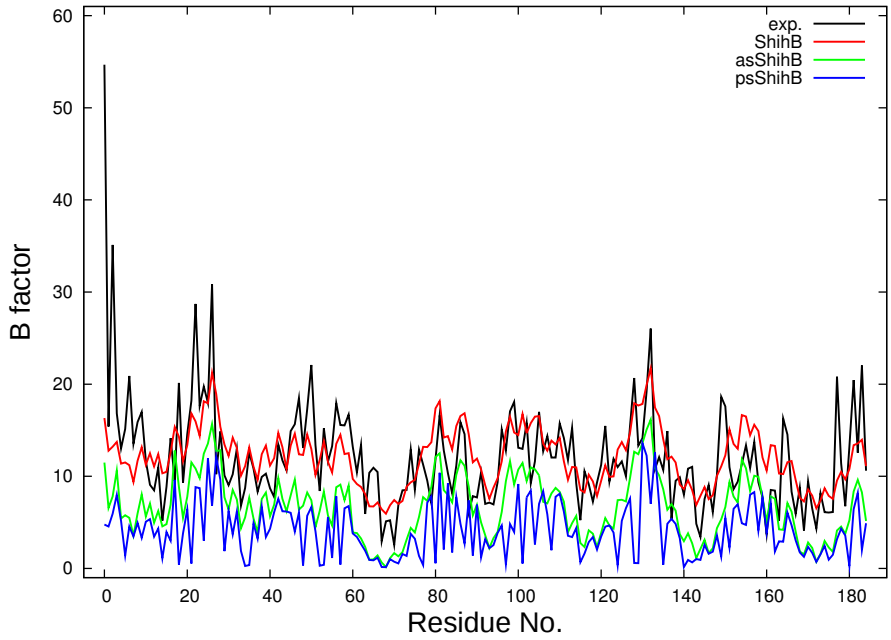
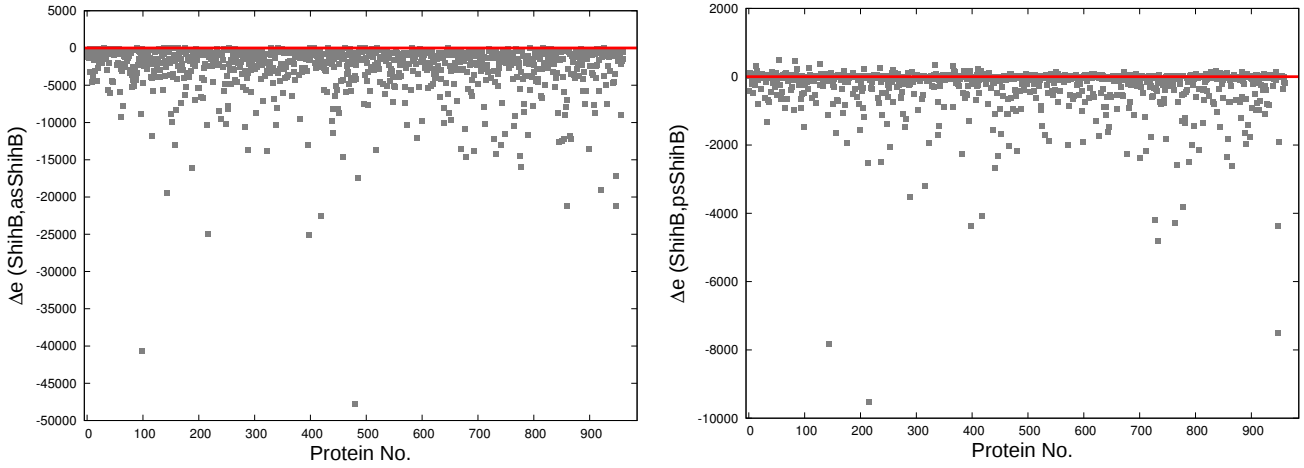



**Figure 1.7.:** For a single protein (PDB code: 153L [Weaver *et al.*, 1995]) the experimental B-factors (*exp.*) are compared to the predicted results of *ShihB*, *asShihB* and *psShihB*.




**(a)** *asShihB* **(b)** *psShihB*

**Figure 1.8.:** The difference of absolute deviations $\Delta e$ from theoretical B-factors are shown for both the *asShihB* and the *psShihB* model with respect to the basic method *ShihB* computed for each protein in our protein set. Predicted B-factors for proteins with $\Delta e < 0$ show less deviations from the experimental temperature factors for the extended than for the original model. $\Delta e = 0$, which is highlighted by the red line indicates no difference of both approaches in accuracy of B-factor prediction.

## 1.3 Parameter Fitting – Proof of Principle

As we already discussed in section 1.1, ENMs have been developed to facilitate the analysis of proteins and protein complexes. Dynamics of large RNA-protein complexes like the ribosome were hardly accessible without reduction schemes [Hamacher & McCammon, 2006; Trylska *et al.*, 2005]. Previous studies presented coarse-grained approaches for MD simulations and NMA based on ENMs [Li & Cui, 2002; Moritsugu & Smith, 2007] as well. Such reduced molecular models allow fast computations of mechanical characteristics of a protein, and, furthermore, a high-throughput screening of *in silico* generated mutants to evaluate the importance of single contacts or residues for protein dynamics [Hamacher, 2010]. Alongside the benefits of ENMs their limitations, that have been discussed inter alia by Soheilifard *et al.* [2008], need to be considered, as well. Numerous coarse-grained approaches have been proposed to investigate the mechanics of proteins, e.g. GNMs and ANMs [Atilgan *et al.*, 2001; Bahar *et al.*, 1997] plus extensions, as well as translation library screw models [Kundu *et al.*, 2002; Kuriyan & Weis, 1991]. Based on simplified assumptions, such reduced protein models approximate protein dynamics omitting for instance effects of electrostatics that are included implicitly at best.

To obtain reliable results, thorough parametrization has to be applied to ENMs. For example, with cutoff distances $r_c$ the range of interactions is determined which affects dynamical properties of residues. Short cutoff distances omit long-range interactions that have been detected to play a prominent role for potassium channels [Gazzarrini *et al.*, 2004]. For large $r_c$ values, distinct signals may get lost due to self-averaging over nearly all interactions. Both GNM and ANM (see section 1.1) do not take into account sidechain interactions since amino acids are modeled as beads placed on the respective $C_\alpha$ position. The parametrization of the interaction potentials affects the resulting mechanics of the protein. For topological studies, a homogeneous parametrization scheme that does not discriminate between bonded and non-bonded contacts would be preferred to determine how structure defines dynamics. For more detailed analysis, amino acid specific potentials can be employed to parametrize the corresponding force constants, as was put forward by Hamacher & McCammon [2006]. Hence, derived dynamical features depend on properties of amino acids as well. To this end, force constants of springs connecting residues within a chain, also called intrachain contacts, can be parametrized by knowledge-based MJ potentials [Miyazawa & Jernigan, 1996] that have been derived from contact frequencies and take into accounts various effects, such as electrostatics, hydrogen bonds, etc.. Interchain contacts, i.e. interacting residues of different subunits/chains of a protein, may be weighted according to values suggested by Keskin *et al.* [1998] (KE). The rationale behind the two-class weighting scheme is a differing statistic of contact frequencies within or among chains due to packing density and hydrophobic effects. Hamacher & McCammon [2006] suggested a force constant of $K = 82$ RT/Å$^2$ as parameter for covalent bonds in accordance to previous studies [Ming & Wall, 2005; Trylska *et al.*, 2005]. Since covalent bonds are far more rigid than non-covalent interaction, weights of non-covalent bonds are typically an order of magnitude smaller than those of peptide bonds (average MJ$\approx$ 3.2 RT/Å$^2$, average KE$\approx$ 3.5 RT/Å$^2$).

In the following, we discuss three approaches which are based on ANMs (see section 1.1) to extract amino acid specific interaction potentials directly from an ensemble of conformations of a common energy minimum structure sampled by MD simulations or other methods. Result-

ing parametrization schemes may enhance the understanding of mutual interactions of residues within proteins with respect to structure and dynamics. Prior to an application to "real" data, we perform a proof-of-principle analysis for the methods (presented in sections 1.3.1, 1.3.2 and 1.3.3). Therefore, we try to regain the (known) interaction potentials of artificially constructed data of BPTI and discuss the performance of the applied approaches. A procedure that fails to retrieve the interaction potentials of this artificial setting can be omitted from further consideration.

The multidimensional PES of a biomolecule comprises multiple minima that are separated by saddle points [Stillinger & Weber, 1984]. Vibrations within such energy basins can be captured by ENMs, but jumps among minima are beyond their scope, which presupposes thorough conformation sampling. In sections 1.4.1 and 1.4.2 we consider two proteins, BPTI and polyalanine (PA), and describe the respective sampling of data that will be used in section 1.4.3 to extract amino acid specific interaction potentials by application of the presented parameter fitting method(s).

### 1.3.1 Stochastic Tunneling

Stochastic algorithms are often applied if the optimization problem under consideration is not feasible for deterministic approaches. So-called Monte Carlo (MC) algorithms [Metropolis & Ulam, 1949] based on statistical thermodynamics have already successfully been applied to computationally difficult problems, e.g. protein folding [Li & Scheraga, 1987] and the traveling salesman problem [Černý, 1985]. In general, MC approaches are employed to minimize the energy of the system under consideration by reiterated stochastic generation of new system configurations. A transition of state $x_{i-1}$ of iteration $i-1$ to state $x_i$ of iteration $i$ is accepted if the Metropolis criterion [Metropolis *et al.*, 1953] is satisfied: either the energy $E_i$ is smaller than $E_{i-1}$, i.e. $\Delta E = E_i - E_{i-1} < 0$, or a random number $rnd \in [0,1]$ fulfills the condition $rnd \leq e^{-\beta \Delta E}$. The internal parameter $\beta$ regulates the convergence behavior of the algorithm; it is physically related to the temperature $T$, as $\beta = 1/k_B T$ with $k_B$ being the Boltzmann constant.

Systematically exploring the PES of a protein is exponentially difficult in the number of residues, as we are concerned with a huge number of low-energy conformations separated by high kinetic barriers. The existence of many local minima is referred to as multiple-minima-problem [Li & Scheraga, 1987]. Although, a consensus regarding the existence of a "funnel structure" for the dynamics of protein folding has emerged in the past [Bryngelson *et al.*, 1995; Dill & Chan, 1997; Leopold *et al.*, 1992], the search for global minima of the PES, which are assumed to correspond to native state configurations, can get trapped in local minima surrounded by high kinetic barriers. To tackle the difficulty of passing from an encountered ground state to another, stochastic tunneling (STUN) was introduced as an extension of MC algorithms with minimization [Hamacher, 2006; Hamacher & Wenzel, 1999; Wenzel & Hamacher, 1999]. The physical idea behind STUN is to allow tunneling of forbidden regions that have shown to be irrelevant for the low-energy properties of the problem. Hence, kinetic barriers between different local

| STUN transformation function $f_s$ |
|---|
| $f_s(E) = 1 - e^{-\gamma(E-E_b))}$ |
| $f_s(E) = \tanh(\gamma(E - E_b))$ |
| $f_s(E) = \ln(\gamma(E - E_b) + \sqrt{1+\gamma^2(E-E_b)^2})$ |

**Table 1.4.:** The presented functions have shown their potential of transforming the PES to allow tunneling [Hamacher, 2006]. $E$ denotes the current energy of the system, and $E_b$ the lowest energy encountered thus far. $\gamma$ is the "tunneling" parameter of the algorithm. Either function can be used in the MC-STUN approach.

minima are lowered and can be overcome more easily. The following non-linear transformation of the PES is used for STUN here:

$$f_s(E) = \sinh(\gamma(E - E_b)) \tag{1.25}$$

with $E_b$ being the lowest minimum encountered thus far; $\gamma$ is the "tunneling" parameter of the algorithm. Hamacher [2006] proposed additional functions to transform the PES of a protein (see Tab. 1.4) that can be used in order to improve the convergence of the algorithm depending on the problem. Convergence behavior may further be improved by adjusting the internal parameter $\beta$ during the simulation: if a specified threshold for the short-time moving average of $f_s$ is not exceeded, $\beta$ is reduced by some fixed factor, otherwise it is increased [Wenzel & Hamacher, 1999]. In addition, the probability to accept a transition from $x_{i-1}$ to $x_i$ of the original Metropolis criterion is modified for STUN as well. The newly generated configuration $x_i$ is accepted if either the resulting (transformed) energy $f_s(E_i)$ is lower than $f_s(E_{i-1})$ or the condition $rnd \leq e^{-\beta(f_s(E_i)-f_s(E_{i-1}))}$ is fulfilled for a random number $rnd \in [0,1]$.

### Fitting with MC-STUN

We apply MC-STUN to parametrize underlying amino acid interaction potentials of a Hessian or a covariance matrix which have been derived from experimental data, such as MD simulations, NMA, etc.. We assume that all (clustered) configurations fluctuate around a central structure $S$, a local minimum of the PES. We derive an ANM (see section 1.1) for this central structure with unknown interaction potentials to parametrize amino acid contacts. The parameters that best describe the data generating process are obtained if the "distance" of experimental and theoretical ANM data is minimal. As distance $d$ we define the Frobenius norm (FN) between the respective Hessian or covariance matrices $M^{\text{exp}}$ and $M^{\text{ANM}}$:

$$\text{FN}(M^{\text{exp}}, M^{\text{ANM}}) = \sqrt{\sum_{ij} \left( M_{ij}^{\text{exp}} - M_{ij}^{\text{ANM}} \right)^2} \tag{1.26}$$

In terms of MC algorithm, we do not minimize the energy $E$ for a protein system but the distance $d$ between experimental matrices $M^{\text{exp}}$ and theoretical ones $M^{\text{ANM}}$. For each encountered local minimum $d_b$, that represents the smallest distance between experiment and theory thus far, we obtain a set of local optimal interaction parameters $\rho_k$, $k = 1, \cdots, K$ with $K$ being the number

of interaction types. Different definitions of interaction types may invoke discrimination of covalent and non-covalent interactions up to amino acid specificity. In each iteration step $i$, each interaction parameter $\rho_k^{(i)}$ is stochastically modified by $\rho_k^{(i)} = \rho_k^{(i-1)} + 2\varepsilon(0.5 - r)$, where $r$ is a random number drawn from a uniform distribution in $[0, 1]$ and $\varepsilon$ describes the margin of variability of the parameters. Transitions are accepted if the Metropolis criterion is satisfied for the STUN-transformed distances. In summary, the pseudocode of the employed MC-STUN algorithm is shown in algorithm 1.

**Algorithm 1** Pseudocode for MC-STUN.

**for** $i = 1 \rightarrow N$ **do**
  stochastic modification of all parameters $\rho_k^{(i)} = \rho_k^{(i-1)} + 2\varepsilon(0.5 - r)$
  compute distance $d_i = \mathrm{FN}(M^{\mathrm{exp}}, M^{\mathrm{ANM},(i)})$
  compute distance difference $\Delta = f_s(d_i) - f_s(d_{i-1})$
  **if** $\Delta < 0$ or $rnd \leq e^{-\beta\Delta}$ **then**
    set $\rho_k^{(i)} = \rho_k^{(i)}$
  **else**
    set $\rho_k^{(i)} = \rho_k^{(i-1)}$
  **end if**
**end for**

In the following proof-of-principle analysis, we examine the capability and performance of MC-STUN to derive harmonic interaction potentials from an artificially constructed Hessian matrix its pseudoinverse, the covariance matrix, using ANMs (see section 1.1).

### MC-STUN : Proof of Principle

As input, we constructed a Hessian matrix $H$ (see Eq. 1.5) and its respective covariance matrix $C$ (see Eq. 1.6) for the bovine pancreatic trypsin inhibitor (BPTI) (PDB code 6PTI [Wlodawer *et al.*, 1987]) based on ANM theory. Interaction potentials of peptide bonds were parametrized with $K = 82$ RT/Å$^2$ as proposed by Hamacher & McCammon [2006]. Non-bonded interactions were modeled within a distance of 13 Å and parametrized by the knowledge-based MJ interaction potentials [Miyazawa & Jernigan, 1996]. We employed MC-STUN to regain the interaction parameters that have been used to construct those matrices.

To extract the interaction potentials, we again used an ANM setting the cutoff distance to $r_c = 13$ Å. We initialized force constants for non-bonded contacts with the arbitrary value of $\rho_k^{(0)} = 5$ RT/Å$^2$ for all amino acid interactions. Covalent bonds were initially parametrized with the potential $\rho_K^{(0)} = K = 82$ RT/Å$^2$. In total, we are concerned with 151 types of amino acid interactions within BPTI at $r_c = 13$ Å which is used both for construction and fitting. Interaction types that are not present, are omitted from the fitting procedure. Since the convergence of the fitting algorithm depends on its internal parameters, we varied those parameters $\varepsilon$, $\beta$ and $\gamma$ (see Tab. 1.5). We restricted the number of iterations to $10^8$ and $10^6$ for the fitting of the Hessian and the covariance matrix, respectively. With the resulting minimum distance $d_b$, we are able to compare different fittings of the same matrix. Smaller distances are related to better minima and, hence, the corresponding parameters are closer to those used for data generation.

| settings for Hessian fittings | |
|---|---|
| $\varepsilon$ | 0.1, 0.2, 0.3 |
| $\beta$ | 1, 2, 5, 10, 15, 20, 25, 50, 75, 80, 90, 100, 110, 125, 150, 175, 200, 225 |
| $\gamma$ | 0.001, 0.01, 0.05, 0.1, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.9 |

| settings for Inverse fittings | |
|---|---|
| $\varepsilon$ | 0.1, 0.2, 0.3, 0.4 |
| $\beta$ | 0.05, 0.1, 0.3, 0.5, 0.7, 0.9 |
| $\gamma$ | 0.1, 3, 10, 50, 100, 400, 1000, 1500, 2000 |

**Table 1.5.:** MC-STUN parameter settings. We varied the internal parameters $\varepsilon$, $\beta$ and $\gamma$ of MC-STUN to achieve convergence of the algorithm. $\varepsilon$ describes the margins of variability for the stochastic modification of fitted interaction potentials at the beginning of each iteration. $\beta$ is a variable of the Metropolis criterion that modulates the acceptance of "worse" results, and $\gamma$ controls the tunneling behavior of the algorithm.

As we know the underlying parametrization scheme, we can also evaluate the goodness-of-fit on an absolute scale. To this end, we computed correlation coefficients for both matrices and parameters as well as the deviations of parameter values.

### MC-STUN : Results

For the proof of principle, we investigated the capability of the extended MC-STUN algorithm to determine amino acid specific interaction potentials that have been used to construct both Hessian and covariance matrix based on ANM theory. To this end, the fittings were started with varying internal parameters $\varepsilon$, $\beta$ and $\gamma$. In Tab. 1.6 we list the algorithm parameters that performed best in terms of yielding minimum distances. In order to evaluate the quality of the fitted interaction strengths, we used FN (see Eq. 1.26) as distance measure. It computes the deviation of the matrix generated with fitted interaction potentials from the reference input matrix. We notice that the absolute value of the best distance $d_b$ that was encountered during simulation is not comparable for both types of matrices. This observation can be explained by differing domains of definition. Values of the Hessian matrix are derived directly from protein coordinates and interaction potentials, whereas the covariance matrix is computed as pseudoinverse from the Hessian (see Eq. 1.5 and 1.6). By this non-linear transformation, large values of the Hessian matrix lead to small values of the covariance matrix and vice versa. Hence, the resulting distance $d_b$ can be used to judge the quality of different fitting runs for the same problem only. For this reason, we computed further qualitative measures for each fitted set of interaction potentials: a) Pearson correlation of both resulting Hessian and covariance matrix with respect to the input data ($\mathrm{cor}_h$, $\mathrm{cor}_i$), b) Pearson correlation of derived and input interaction strengths ($\mathrm{cor}_\rho$) as well as c) absolute differences of corresponding potentials, thereby we distinguish non-bonded ($\Delta_\rho$) and bonded ($\Delta_{\mathrm{pep}}$) values. The results for the five best fittings of each matrix are shown in Tab. 1.6. We detect an evident correspondence between small distances and small deviations of the fitted interaction potentials leading to high correlation coefficients of the respective matrices.

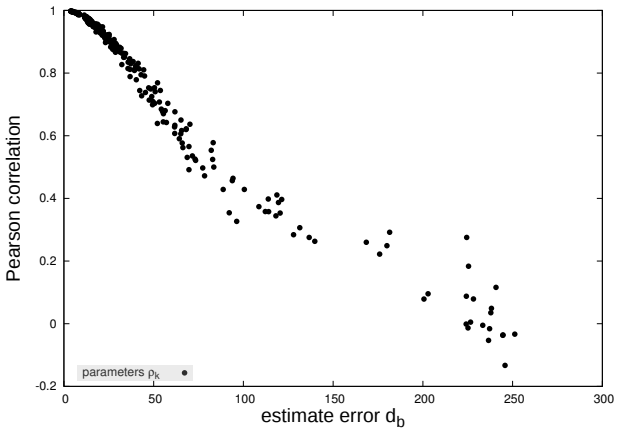
Obviously, nearly all of the best fittings invoke a small $\varepsilon = 0.1$, which defines the maximum variation of the interaction parameters $\rho_k$ at the beginning of each iteration step. Small varia-

| Hessian | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $\varepsilon$ | $\beta$ | $\gamma$ | $d_b$ | $\mathrm{cor}_h$ | $\mathrm{cor}_i$ | $\Delta_\rho$ | $\Delta_{\mathrm{pep}}$ | $\mathrm{cor}_\rho$ |
| 0.1 | 110 | 0.7 | 3.8002 | 0.9999 | 0.9999 | 11.84 | 0.002 | 0.9974 |
| 0.1 | 90 | 0.9 | 3.8722 | 0.9999 | 0.9999 | 10.76 | 0.039 | 0.9978 |
| 0.1 | 125 | 0.7 | 3.8976 | 0.9999 | 0.9999 | 12.37 | 0.009 | 0.9973 |
| 0.1 | 80 | 0.9 | 3.9135 | 0.9999 | 0.9999 | 12.84 | 0.015 | 0.9970 |
| 0.1 | 225 | 0.35 | 3.975 | 0.9999 | 0.9999 | 11.37 | 0.001 | 0.9973 |
| Inverse | | | | | | | | |
| $\varepsilon$ | $\beta$ | $\gamma$ | $d_b$ | $\mathrm{cor}_h$ | $\mathrm{cor}_i$ | $\Delta_\rho$ | $\Delta_{\mathrm{pep}}$ | $\mathrm{cor}_\rho$ |
| 0.1 | 10000 | 0.9 | 0.0148 | 0.9999 | 0.9999 | 36.58 | 1.48 | 0.9644 |
| 0.1 | 5000 | 0.9 | 0.0249 | 0.9998 | 0.9997 | 57.28 | 3.90 | 0.9342 |
| 0.1 | 5000 | 0.7 | 0.0292 | 0.9998 | 0.9996 | 80.85 | 2.13 | 0.8896 |
| 0.1 | 2000 | 0.9 | 0.0557 | 0.9994 | 0.9987 | 134.72 | 5.57 | 0.7558 |
| 0.2 | 2000 | 0.9 | 0.0577 | 0.9991 | 0.9986 | 144.52 | 3.65 | 0.6728 |

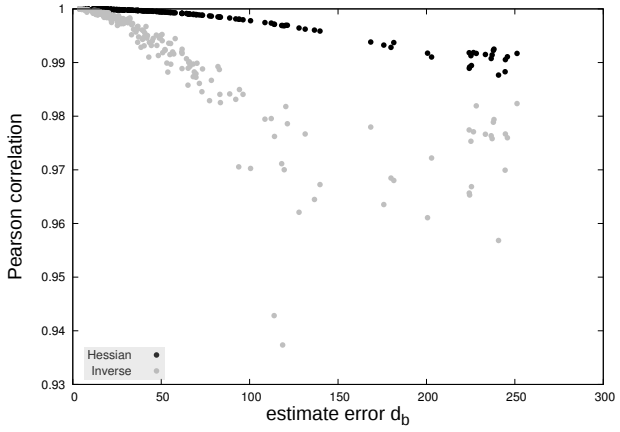
**Table 1.6.:** The results for MC-STUN based fittings of interaction potentials by usage of the constructed ANM-Hessian or its pseudoinverse (Eq. 1.5 and 1.6). For each variant, we listed the five best fittings only, i.e. those with smallest FN values $d_b$ for Hessian or covariance matrix depending on the input matrix type. Varying internal parameters $\varepsilon$, $\beta$ and $\gamma$ were used. To evaluate the quality of the derived interaction potentials, we computed absolute deviations for non-bonded ($\Delta_\rho$) and bonded ($\Delta_{\mathrm{pep}}$) interaction as well as the resulting correlation of interaction potentials ($\mathrm{cor}_\rho$), and the corresponding Hessian ($\mathrm{cor}_h$) and covariance matrices ($\mathrm{cor}_i$).

tions of the fitted values are preferred. However, further initial settings of interaction potentials should be examined for this finding as well, to exclude an artifact owing to the proximity of start to "real" values. Note that although the initial interaction potentials were chosen to be close to those used for the reference matrices, not all fitting routines were capable to yield appropriate results (see Fig. 1.9). What becomes evident especially for fittings in the space of the inverse matrix – i.e. the covariance matrix – is that the best results were obtained for high $\beta$ values in combination with high $\gamma$ values. The internal parameter $\gamma$ defines the tunneling behavior of the algorithm, i.e. for larger values kinetic barriers of the PES are easier to cross. A large value for $\beta$ enforces the algorithm to accept almost no distance increases, as it controls the probability to accept a transition in a state that is worse than the previous one. Hence, using the observed parameter settings, large areas of the PES are sampled; the acceptance of state transitions is strongly biased towards "better" states with a smaller energy or in our cases smaller differences of input and fitted matrix. The choice of optimal parameters for both Hessian and inverse based fittings is similar, but differs in the actual magnitude.

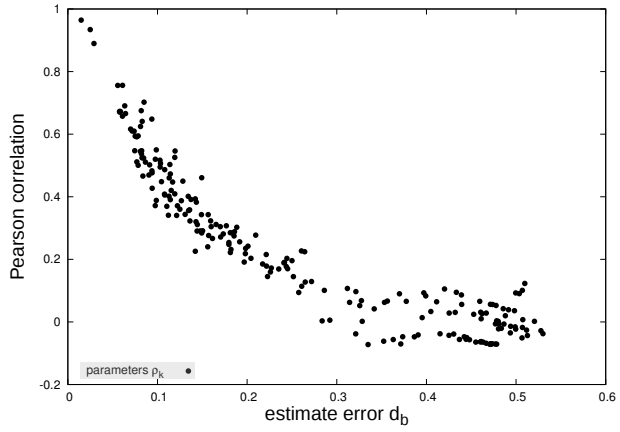
In Fig. 1.9 we show the results for all fittings with varying internal parameters. We compare the results for fittings of the Hessian matrix and its inverse matrix. Therefore, we plotted the final minimum distances $d_b$ with the Pearson correlation coefficient for the resulting parameters and matrices with respect to the input (artificially constructed) data. Interestingly, the correlation of both Hessian and covariance matrix is not sensitive to the actual parametrization of the interaction potentials. Even for the largest FN values, a correlation of 0.93 is obtained, for the Hessian



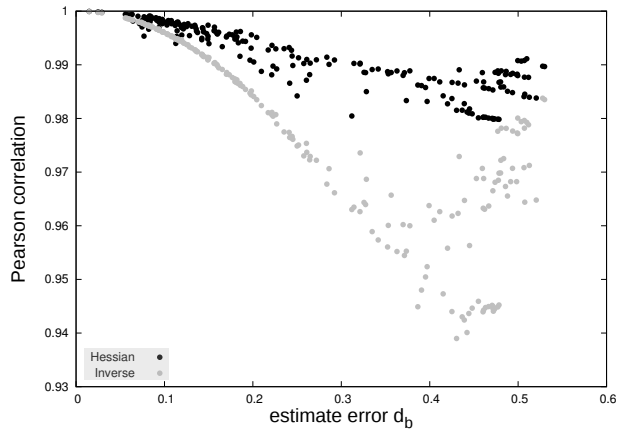

**(a)** Correlation of fitted and input interaction strengths for fittings based on the Hessian matrix.




**(b)** Correlation of Hessian and covariance matrices for the fitted interaction parameters obtained from Hessian fittings.




**(c)** Correlation of fitted and input interaction strengths for fittings based on the inverse Hessian.




**(d)** Correlation of Hessian and covariance matrices for the fitted interaction parameters obtained from covariance matrix fittings.

**Figure 1.9.:** Results for MC-STUN fittings based on the Hessian (a, b) and its inverse (c, d), which were performed to determine the underlying interaction potentials $\rho_k$. For each minimum distance $d_b$ that describes the deviation of the fitted from the original matrix, the correlation of fitted contact potentials (a, c) and the corresponding matrices (b, d) are plotted for each reference matrix type, respectively.

matrix we find an even higher correlation with 0.98. This results presumably from the large difference between the interaction potentials of covalent and non-covalent contacts, that may dominate the matrix structure. In addition, the matrix correlation results are influenced by the type of the reference matrix. For example, we detect less scattering for the Hessian than for the covariance correlation if we performed the fit with the Hessian as reference matrix, at least for smaller distances $d_b$. The same holds for the fitting of the pseudoinverse as well, even though the covariance matrix is more sensitive to misfitted interaction parameters than the Hessian. In general, we notice that the derivation of interaction potentials is easier using the Hessian matrix as reference system. Appropriate results showing high correlation with the original values were achieved for diverse parameter settings. In contrast, for the covariance matrix only three fitted sets of contact potentials showed a correlation larger than 0.8 with the original ones.

This comes as no surprise, since the interaction strengths are directly contained in the Hessian matrix, whereas the covariance matrix is obtained by a non-linear transformation. Hence, the force constants parametrizing springs of contacting residues are only implicitly contained in this matrix. Another aspect, that needs to be discussed in this context, is the runtime of the algorithm. In theory, the number of iterations that were performed should preferably be increased to assure better convergence and, therefore, better results. In the case of fitting covariance matrices, limitations arise from the protein size since in each iteration the pseudoinverse of the Hessian has to be computed, which increases the runtime dramatically. Fitting of covariance matrices may get infeasible due to resource restrictions even though we are concerned with a coarse-grained protein model. Furthermore, the initial search for optimal algorithm parameters needs to be performed for each fitting setup. Therefore short runs are simulated to determine those parameters, but their influence on long-time convergence behavior of the algorithm remains unclear. Nevertheless, we were able to demonstrate the capability of MC-STUN to derive interaction potentials of a given ANM based on both Hessian and covariance matrix.

Although we have proven that interaction potentials can be determined from a given Hessian or covariance matrix, we omit application of this method to derive interaction potentials from a set of experimental structures, derived from e.g. MD simulations, due to its shortcomings. We therefore decided to turn our attention to an alternative parametrization procedure which is described next.

### 1.3.2 Likelihood Based Methods

Maximum likelihood estimation (MLE) is a well-established procedure to determine the data generating parameters of stochastic processes. The idea is to identify the parameters of the underlying model by maximizing the probability to observe the given data set. Numerous studies have employed MLE based approaches: For example, Murshudov *et al.* [1997] presented an MLE based method for macromolecular structure refinement; it was also applied to DNA sequences to derive evolutionary trees [Felsenstein, 1981] or to estimate haplotype frequencies [Excoffier & Slatkin, 1995]. A major advantage over MC based methods (section 1.3.1) is the guarantee to identify the optimal parameters whenever the likelihood is maximized. In contrast, stochastic algorithms locally sample the whole parameter space and may never converge to a local optimum.

For a given data set $\vec{x} = (x_1, x_2, ..., x_n)$ with independent, identically distributed observations $x_i$ the likelihood is defined as the product of the respective probabilities given a process with the underlying set of parameters $\vec{\Theta}$.

$$P(\vec{x}|\vec{\Theta}) = \prod_i P(x_i|\vec{\Theta}) \tag{1.27}$$

Thus, the observations $\vec{x}$ have been generated by a process with the yet unknown model parameters $\vec{\Theta}$. The best estimate of those process parameters is found for the maximum of the

likelihood. Similarly, maximum *a posteriori* estimation (MAPE) is applicable as well, whenever prior knowledge is available. The posterior probability is derived from the Bayesian theorem:

$$P(\vec{\Theta}|\vec{x}) = \frac{P(\vec{x}|\vec{\Theta}) \cdot P(\vec{\Theta})}{P(\vec{x})} = \prod_i \frac{P(x_i|\vec{\Theta}) \cdot P(\vec{\Theta})}{P(x_i)} \tag{1.28}$$

Since the probability of the observed events $P(x_i)$ does not depend on the process parameters and can thus be considered as constant during $\vec{\Theta}$-optimization, it is omitted from the estimation routine. For convenience, the log-likelihood $\log P(\vec{x}|\vec{\Theta})$ is often used instead of the likelihood $P(\vec{x}|\vec{\Theta})$. This is justified as application of the logarithm changes only the absolute value but not the position of the maximum (or minimum). Hence, we apply MLE and MAPE by using the log-likelihood rather than the likelihood. The estimation of model parameters is similar for both MLE and MAPE, but the latter approach additionally invokes *a priori* information, and is formulated as:

$$\hat{\vec{\Theta}} = \arg\max_{\vec{\Theta}} \ \log P(\vec{x}|\vec{\Theta}) = \arg\max_{\vec{\Theta}} \ \sum_i \log P(x_i|\vec{\Theta}) \tag{1.29}$$

$$\hat{\vec{\Theta}} = \arg\max_{\vec{\Theta}} \ \log(P(\vec{x}|\vec{\Theta}) \cdot P(\vec{\Theta})) = \arg\max_{\vec{\Theta}} \ \sum_i \log P(x_i|\vec{\Theta}) + \log P(\vec{\Theta}) \tag{1.30}$$

Maximizing the likelihood with respect to the unknown parameters invokes computing of derivatives of the log-likelihood, and of the prior in case of MAPE, in terms of $\Theta_j$, i.e. for each parameter. The resulting equations are set to zero and and solved to obtain the estimates of $\Theta_j$.

### Fitting with MLE

In this study, we employ MLE/MAPE to determine amino acid specific interaction potentials that best describe the mechanics of a protein. Again, we make use of the coarse-grained ANM representation (see section 1.1) to derive those contact potentials. Such potentials are used to model the spring constants of contacting residues in a protein. The larger the respective force constant is the more rigid is the connection of the pair of amino acids. Here, we are concerned with a set of different experimental structures, e.g. from MD simulations, of a protein describing fluctuations around an equilibrium state. Estimating the interaction potentials of the fluctuating protein paves the way to understand its intrinsic biomechanical properties. To this end, we define the likelihood to observe a single protein configuration $\vec{x}$ as:

$$P(\vec{x}|\vec{\Theta}) = \frac{e^{-\frac{\beta}{2} \cdot \vec{x}^T \cdot V(\vec{\Theta}) \cdot \vec{x}}}{Z(\vec{\Theta})} \tag{1.31}$$

where $V(\Theta)$ denotes the Hessian matrix of an ENM (Eq. 1.5) and $Z$ the partition function, which is defined as follows:

$$Z(\vec{\Theta}) = \int e^{-\frac{\beta}{2} \cdot \vec{x}^T \cdot V(\vec{\Theta}) \cdot \vec{x}} d^{3N}x = \prod_{i=7}^{3N} \sqrt{\frac{2\pi}{\lambda_i}} \tag{1.32}$$

For the Hessian matrix, we find $3N-6$ non-zero eigenvalues $\lambda_i$, $N$ is the number of residues. In the following, we set $\beta = 1$. The likelihood of a single configuration $\vec{x}$ describes its probability given a protein model that parametrizes interactions between contacting amino acids by the potentials $\vec{\Theta}$. Note that $\vec{x}$ is a displacement vector that contains the absolute deviations of each amino acid from the assumed equilibrium state. Based on Eq. 1.31, we define the likelihood for a set of $S$ structures, that were sampled independently from the very probability distribution as:

$$P(\{\vec{x}_s\}|\vec{\Theta}) = \frac{e^{-\frac{1}{2}\sum_s^S \vec{x}_s^T \cdot V(\vec{\Theta}) \cdot \vec{x}_s}}{Z(\vec{\Theta})^S} \tag{1.33}$$

For the estimation of the underlying interaction potentials, we maximize the log-likelihood averaged over the set of configurations.

$$\log P(\{\vec{x}_s\}|\vec{\Theta}) = -\frac{1}{2S}\sum_s^S \vec{x}_s^T \cdot V(\vec{\Theta}) \cdot \vec{x}_s - \log Z(\vec{\Theta}) \tag{1.34}$$

In addition, we introduce prior information, e.g. by employing knowledge-based interaction potentials [Keskin *et al.*, 1998; Miyazawa & Jernigan, 1996], and, hence, perform an MAPE as well. Therefore, we assume the parameters to be drawn from a normal distribution that is characterized by an expectation value $\mu$ and a standard deviation $\sigma$. The logarithm of the *a priori* distribution is:

$$\log P(\Theta_i) = -\frac{1}{2}\log(2\pi\sigma_i^2) - \frac{1}{2}\frac{(\Theta_i - \mu_i)^2}{\sigma_i^2} \tag{1.35}$$

Combining the averaged log-likelihood (Eq. 1.34) and the logarithmic prior (Eq. 1.35), we get the log-posterior:

$$\log P(\vec{\Theta}|\{\vec{x}_s\}) = -\frac{1}{2S}\sum_s^S \vec{x}_s^T \cdot V(\vec{\Theta}) \cdot \vec{x}_s - \log Z(\vec{\Theta}) - \frac{1}{2}\sum_i \left(\log(2\pi\sigma_i^2) + \frac{(\Theta_i - \mu_i)^2}{\sigma_i^2}\right) \tag{1.36}$$

Maximizing the log-likelihood or the log-posterior requires the derivation of those probability functions with respect to all $\Theta_i$. The derivative of the partition function $Z$ (Eq. 1.32) cannot be determined analytically but must be computed numerically. Therefore, we investigate the smoothness of $Z$, i.e. a monotonous behavior even if interaction potentials are modified. To this end, it is sufficient to consider an effective partition function $z$ that is defined as:

$$z = \sum_{i=7}^{3N} \log \lambda_i \tag{1.37}$$

We determined the ANM-Hessian matrix for BPTI (PBD code 5PTI [Wlodawer *et al.*, 1984]) and computed the respective eigenvalues $\lambda_i$ via SVD. Interactions of contacting amino acids were parametrized by MJ interaction potentials [Miyazawa & Jernigan, 1996]. We varied the force constant that describes the strength of GLY-ARG contacts in the range $[0, 10]$ with a step size

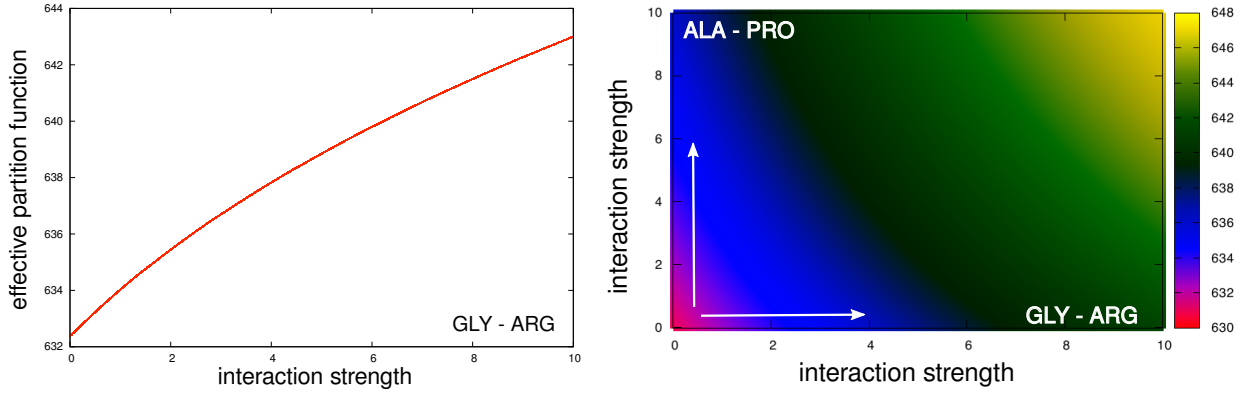



**Figure 1.10.:** The smoothness of the (effective) partition function (Eq. 1.37) was investigated by varying the interaction potential of one (a) or two (b) amino acid contact types.

of 0.1. Furthermore, we repeated the computation by varying GLY-ARG interaction strengths as described and additionally the potential for ALA-PRO contacts in the same range. Those types of amino acid contacts have been chosen since they can be found more than once in BPTI for a contact distance $r_c = 13$ Å. For both one and two modified interaction potentials, the effective partition function $z$ exhibits a monotonous behavior without any points of discontinuity (see Fig. 1.10). Thus, we convinced ourselves of the numerical stability of numerical derivatives of $Z$.

The derivatives of the logarithmic likelihood (Eq. 1.34) and posterior (Eq. 1.36) are computed and set to zero:

$$\frac{\partial}{\partial \Theta_i} \log P(\{\vec{x}_s\}|\vec{\Theta}) = -\frac{1}{2S}\sum_s \vec{x}_s^T \cdot V^{(i)}(\vec{\Theta}) \cdot \vec{x}_s - \frac{Z^{(i)}(\vec{\Theta})}{Z(\vec{\Theta})} \stackrel{!}{=} 0 \qquad (1.38)$$

$$\frac{\partial}{\partial \Theta_i} \log P(\vec{\Theta}|\{\vec{x}_s\}) = -\frac{1}{2S}\sum_s \vec{x}_s^T \cdot V^{(i)}(\vec{\Theta}) \cdot \vec{x}_s - \frac{Z^{(i)}(\vec{\Theta})}{Z(\vec{\Theta})} - \frac{(\Theta_i - \mu_i)}{\sigma_i^2} \stackrel{!}{=} 0 \qquad (1.39)$$

with $V^{(i)}(\vec{\Theta})$ being the derivative of the Hessian matrix in terms of $\Theta_i$. For the derivative of the partition function $Z$, we use the first order difference as a numerical approximation.

$$Z^{(i)}(\vec{\Theta}) \approx \frac{Z(\vec{\Theta}) - Z(\vec{\Theta} - h\vec{e}_i)}{h} \qquad (1.40)$$

where $\vec{e}_i$ is a unit vector, i.e. a vector containing zeros for each position, except the $i^{th}$ entry is one, and the constant $h$ approaches zero. A numerically stable computation of the term $Z^{(i)}(\vec{\Theta})/Z(\vec{\Theta})$ was obtained by using the formula and setting $h = 10^{-7}$:

$$\frac{Z^{(i)}}{Z} = \frac{1}{h}\left[ e^{\sum_{i=7}^{3N}\left(-\log\sqrt{\tilde{\lambda}_i}+\log\sqrt{\lambda_i}\right)} - 1 \right] \qquad (1.41)$$

We presented the mathematical framework, that is required to use likelihood based methods to estimate interaction potentials that describe the connectivity of amino acid types in a structure ensemble of a protein. The fitting procedure utilizes ANM theory (see section 1.1) for modeling the Hessian matrix and the corresponding eigenvalues. A set of structures described by their respective displacement vectors is required as input, other than for MC-STUN (see section 1.3.1) which was based on a Hessian or a covariance matrix that were derived from a multitude of protein configurations. Note that the runtime of MLE/MAPE fitting depends on the number of interaction types only and not on the number of snapshots. In the following, we perform a proof-of-principle analysis for the likelihood based approaches and discuss the capability to extract interaction potentials from a set of displacement vectors $\vec{x}$ drawn from a known probability distribution (see Eq. 1.31).

### MLE/MAPE : Proof of Principle

To validate the capability of the likelihood based fitting approaches to extract amino acid specific contact potentials, we created a set of snapshots with known interaction parameters. We computed an ANM-Hessian matrix parametrized with amino acid specific MJ potentials for residue pairs closer than a cutoff distance $r_c = 13$ Å. Peptide bonds were weighted by 82 RT/Å$^2$. Similar to the MC-STUN approach (section 1.3.1) that has proven the capability of regaining interaction potentials from both Hessian and covariance matrix, we applied the fitting to the protein BPTI (PDB code 6PTI [Wlodawer *et al.*, 1987]). Snapshots or rather displacement vectors $\vec{x}$ are distributed according to the probability distribution described by the Hessian matrix $V(\vec{\Theta})$ (see Eq. 1.31).

We applied a Cholesky decomposition with pivoting strategy to the covariance matrix $C$ yielding $C = L \cdot L^T$ to generate snapshots that are correlated according to $C$. Furthermore, a vector $\vec{y}$ comprising $3N$ elements drawn from a standard normal distribution $\mathcal{N}(\mu = 0, \sigma = 1)$ is computed and multiplied with the lower triangular matrix $L$ resulting in $\vec{x} = L \cdot \vec{y}$. By pivotal resorting, we obtain a displacement vector $\vec{x}$ which was drawn from the defined probability distribution. To evaluate the quality of snapshot generation, we determined the number of configurations that is necessary to reconstruct the initial correlation of amino acids pairs described by $C$.

The actual fitting routine was implemented using a multidimensional root-finding algorithm of the GNU Scientific Library (GSL) [Galassi, 2009]. The algorithm was not capable to find the root for all 151 interaction types at once (data not shown). To tackle this issue, interaction potentials were estimated successively. We, therefore, systematically grouped amino acids into disjunct groups as was suggested by Pape *et al.* [2010] for amino acid alphabet reduction (see Fig. 1.11, discussed in section 4.2). Starting with a single amino acid type and, thus, two possible types of interactions between any two residues (covalent vs. non-covalent), interaction potentials are retrieved from the set of displacement vectors. Note that peptide bonds are considered separately in each step. Subsequent iterations use randomly modified results from the previous tree level as starting point. Each value is stochastically changed within ±10%. A schematic view of the tree-based "fit-n-split" amino acid handling is presented in Fig. 1.11. Prior information needed for MAPE (Eq. 1.35) is obtained by averaging the MJ parameters according

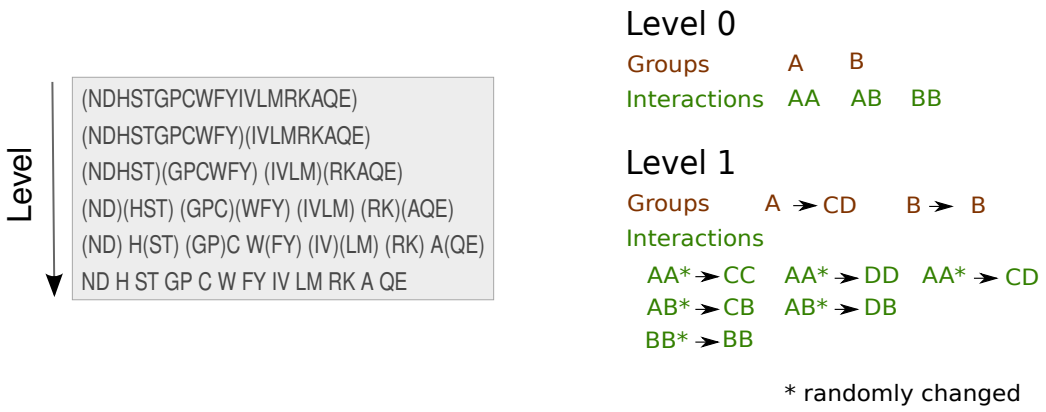



**Figure 1.11.:** On the left, the amino acid grouping scheme as proposed by Pape *et al.* [2010] is shown. The arrow indicates the order, in which the grouping is applied to the fitting of interaction potentials using MLE/MAPE. On the right, we illustrate how the splitting of interaction types is performed for subsequent levels.

to the amino acid alphabet of the current tree level yielding the expectation value $\mu$ of a normal distribution.

Primarily, we compared the performance of MLE (Eq. 1.34) and MAPE (Eq. 1.36) for a set of 10,000 snapshots that were generated as described before. For MAPE, we used a rather narrow prior based on a normal distribution with a standard deviation of $\pm 5\%$ of the expectation value which is computed from the (averaged) MJ potentials. In addition, a sensitivity analysis was performed for MAPE to detect the influence of the number of snapshots as well as the width of the prior, which is defined as standard deviation, on the quality of the fitted interaction parameters (see Tab. 1.7).

### MLE/MAPE : Results

Since MLE and MAPE require a set of independent snapshots drawn from the respective distribution, we used a Cholesky decomposition approach to generate those displacement vectors. Here, we examine how many snapshots are necessary to approach the original correlation among residues as described by the covariance matrix $C$ which was used for snapshot generation. To this end, we computed the correlation of the displacement vectors, i.e. the fluctuation of amino acids. In Fig. 1.12 we present a scatterplot comparing entries of the initial correlation matrix with the entries of correlation matrices that were computed for 10, 100, 1,000 and 10,000 snapshots. Obviously, the resemblance of both data sets increases with the number of snapshots. To quantify this observation, we computed the Pearson correlation coefficient between initial residue correlation and the resulting correlation derived for varying numbers of snapshots. For about 500 independent displacement vectors, we detect a correlation of larger than 0.9. Hence, a set of about 500 structures captures more than 90% of the dynamics of a protein.

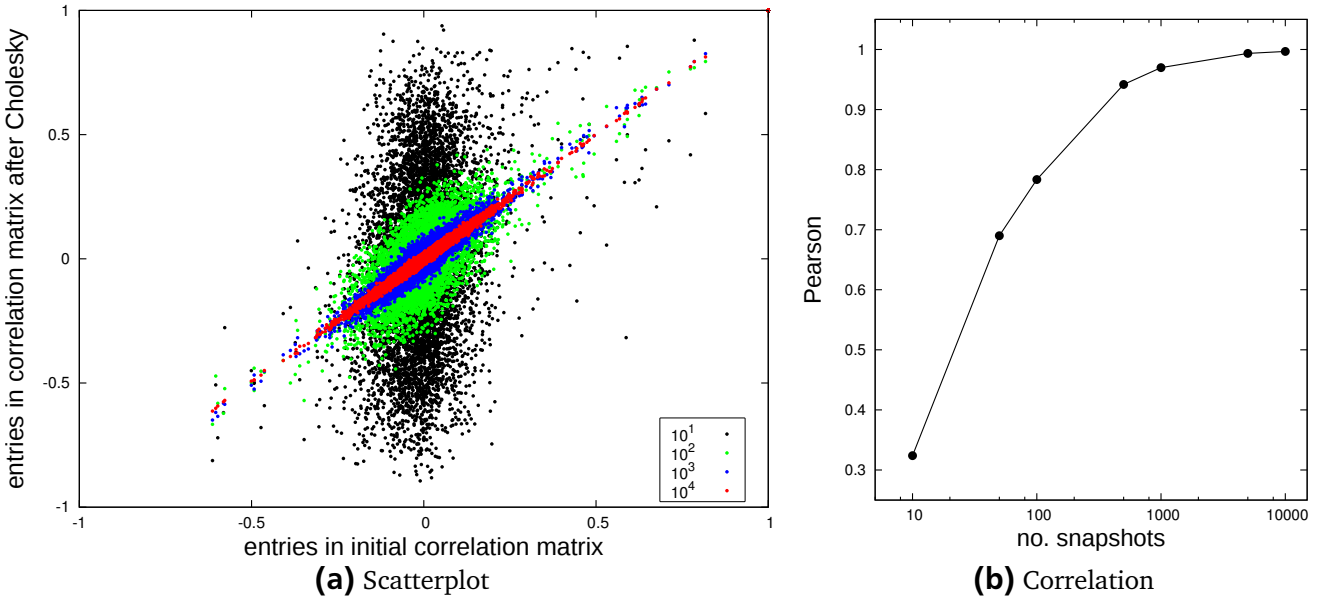
To verify the concept of likelihood based estimation of interaction potentials from a set of displacement vectors fluctuating around a central protein configuration, we generated 10,000

| index | # sns | $\mu$ | $\sigma$ [%] |
|---|---|---|---|
| 1 (38) | 100 | MJ | 10 (100) |
| 2 (39) | 250 | MJ | 10 (100) |
| 3 (40) | 500 | MJ | 10 (100) |
| 4 (41) | 750 | MJ | 10 (100) |
| 5 (42) | 1,000 | MJ | 10 (100) |
| 6 (43) | 2,500 | MJ | 10 (100) |
| 7 (44) | 5,000 | MJ | 10 (100) |
| 8 (45) | 7,500 | MJ | 10 (100) |
| 9 (46) | 10,000 | MJ | 10 (100) |
| 10 (47) | 25,000 | MJ | 10 (100) |
| 11 (48) | 50,000 | MJ | 10 (100) |
| 49 - 56 | 1,000 | MJ | 10 |
| 57 - 64 | 1,000 | MJ | 100 |
| 32 | 10,000 | EQ | 100 |
| 33 | 10,000 | EQ | 150 |
| 34 | 10,000 | EQ | 200 |

| index | # sns | $\mu$ | $\sigma$ [%] |
|---|---|---|---|
| 12 | 10,000 | MJ | 5 |
| 13 | 10,000 | MJ | 10 |
| 14 | 10,000 | MJ | 15 |
| 15 | 10,000 | MJ | 20 |
| 16 | 10,000 | MJ | 30 |
| 17 | 10,000 | MJ | 40 |
| 18 | 10,000 | MJ | 50 |
| 19 | 10,000 | MJ | 65 |
| 20 | 10,000 | MJ | 75 |
| 21 | 10,000 | MJ | 85 |
| 22 | 10,000 | MJ | 100 |
| 23 | 10,000 | MJ | 150 |
| 24 | 10,000 | MJ | 200 |

**Table 1.7.:** Settings for the sensitivity analysis which was performed for MAPE-based fitting of interaction potentials. Values in parentheses give alternative settings. Fitting runs belonging to indices not mentioned here are omitted from further analysis. MJ indicates individual prior means for all interaction types, whereas EQ discriminates only covalent and non-covalent bonds, for the latter the mean MJ value is used to center the normally distributed prior.

snapshots according to a covariance matrix $C$ which was derived from an ANM parametrized with MJ interaction potentials (see section 1.1). The snapshots served as input for the fitting of amino acid specific potentials by an MLE or MAPE approach. MAPE (Eq. 1.36) utilizes prior knowledge in contrast to MLE (Eq. 1.34), which was implemented using a normal distribution. For both approaches, the parameters were derived in subsequent iterations with an increasing amino acid alphabet. The resulting interaction potentials of both MLE and MAPE are compared with the input MJ values. Fig. 1.13 illustrates the quality of both employed fitting routines. Clearly, the parameters obtained by MAPE show higher resemblance with the actual MJ parameters, whereas the results of MLE exhibit more scatter. Thus, invoking prior knowledge improves the quality of the fit dramatically. To quantify this observation, we computed the correlation of obtained and initial interaction parameters as well as their deviation (see Fig. 1.13). MLE achieves a correlation of about 0.6 of initial and fitted force constants used to weight non-covalent amino acid connections. In contrast, the correlation obtained for MAPE is nearly one indicating a qualitatively higher fit, a conclusion that is further supported if we look at the absolute deviation $\Delta_\rho$ of those parameters. The differences were summed up for all 150 non-covalent interaction types present in BPTI. In total, the MLE derived interaction potentials differ from the MJ parameters by a value of about 1 RT/Å$^2$ per parameter, which is about one third of the mean MJ, whereas the deviations are negligible for MAPE. Interestingly, the peptide bond parametrization obtained by MLE is closer to the input value of 82 RT/Å$^2$ than the

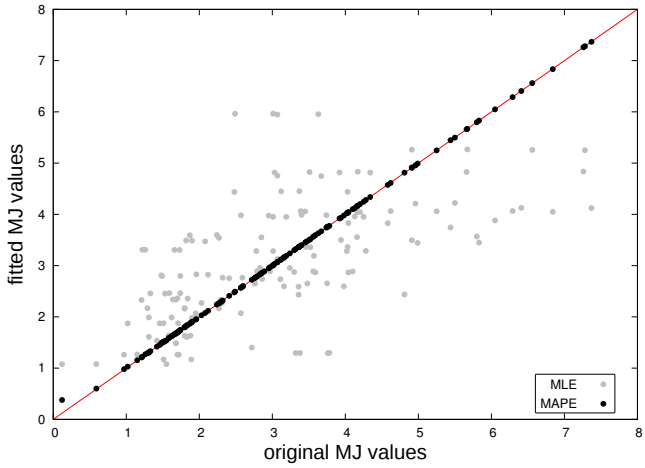



**(a)** Scatterplot **(b)** Correlation

**Figure 1.12.:** For varying numbers of snapshots the correlation of amino acids was determined and compared with the initial correlation that was used for the generation of displacement vectors via Cholesky decomposition. In (a) we illustrate the difference of both correlations in a scatterplot. A quantification using the Pearson correlation coefficient is given in (b).

corresponding MAPE result. A closer inspection of the fitted interaction parameters revealed shortcomings of MAPE to find the correct values for both minimum and maximum parameter values. Deviations were maximal for those interaction potentials. Since we applied a stepwise fitting, we obtained interaction potentials for grouped amino acids as well. Closer inspection of those intermediate results reveals that in case of two parameters, i.e. only a single amino acid type plus peptide bond, the parameter for the covalent bond is recovered correctly for both methods. Fitting without prior knowledge performed better in terms of correlation of contact potentials for a medium number of different interaction types (data not shown).

Since we employed a rather narrow prior for the comparison of MLE and MAPE, we performed a sensitivity analysis for the prior as well as for the number of snapshots for MAPE. We find that the quality of the obtained interaction potentials does not depend on the number of snapshots that were used for fitting. Similarly, we have also shown that the initial correlation among residue fluctuations can be restored for about 500 Cholesky derived snapshots. On the contrary, variation of the standard deviation $\sigma$ of the normally distributed prior around (averaged) MJ parameters has a direct influence on the quality of the fit. Fig. 1.14 depicts the average deviation of a single parameter $\Delta_\rho$ in dependence of the standard deviation $\sigma$ that was applied in the respective fitting, $\sigma$ decribes the width of the normal distribution by percentage of the expectation value $\mu$ which is placed at the MJ value, i.e. for larger parameters we find a wider distribution. The curves for intermediate results are shown as well. In order to validate the results, we used the averaged MJ interaction potentials for each tree level as reference for comparison. Note that we averaged over all possible MJ parameters that were grouped into the respective interaction type, which may be the cause for a reduced resemblance to fitted parameters since we consider only existing amino acid pairings during the fit. However, we find the




| | MLE | MAPE |
|---|---|---|
| $\text{cor}_\rho$ (Pearson) | 0.6177 | 0.9999 |
| $\text{cor}_\rho$ (Spearman) | 0.6725 | 0.9999 |
| $\Delta_\rho$ | 136.11 | 0.695 |
| $\Delta_{\text{pep}}$ | 4.47 | 23.85 |

**Figure 1.13.:** Comparison of the results of MLE and MAPE fitting of interaction potentials. A set of snapshots was created based on the ANM-covariance matrix of BPTI. MJ [Miyazawa & Jernigan, 1996] interaction potentials were used for the computation of the ANM. On the left the parameters derived with both MLE and MAPE fitting are plotted in comparison with the original MJ values. In the table on the right, the correlation between fitted and original potentials is shown ($\text{cor}_\rho$) as well as their absolute deviations ($\Delta_\rho$), and $\Delta_{\text{pep}}$ denotes the absolute difference of the respective peptide bond potentials.

same tendency for all numbers of parameters to deviate more if a larger $\sigma$ is chosen. An analogous effect is not detectable for the deviation of peptide bond potentials, which fluctuates rather randomly for increasing $\sigma$. This is presumably due to the respective prior definition since the root-finding algorithm accepts larger deviations with a smaller penalty. In other words, the interaction potential used for weighting covalent bonds is an order of magnitude larger than that for non-covalent interactions. Therefore, the prior penalizes deviations from the mean value less for peptide bonds. Thus, the algorithm preferably minimizes all non-covalent potentials. Nonetheless, we achieve a correlation larger than 0.9 for Hessian and covariance matrices computed with initial and fitted parameters. Interestingly, the fitted interaction potentials can be improved by averaging the results of independent MAPE applications (data not shown).

In addition, we performed a fitting with a uniform prior (EQ) for all interaction parameters with an expectation value which corresponds to the mean MJ value and relative standard deviations $\sigma = 100, 150, 200\%$. A comparison with fittings that employed individual priors with equivalent $\sigma$ exhibited a reduced capability to determine the true parameters.

In this study, we applied likelihood based estimation approaches to derive interaction potentials that weight amino acid specific contacts. Invoking prior knowledge leads to an improved performance of the applied fitting methods. In the case of BPTI, we are concerned with a total of 150 different non-covalent amino acid pairs plus peptide bonds. Although MLE and MAPE yield per definition the parameters that best describe the data generating process, the performance largely depends both on the number of unknowns and the root-finding algorithm. Since finding the roots of 151 equations at the same time is hardly feasible, we employed a stepwise fitting approach based on reduced amino acid alphabets. Even then, we find approximations of the input parameters whose quality depends massively on the prior definition. Not all fittings converged appropriately, some actually yielded negative parameters. Shortcomings of the routines

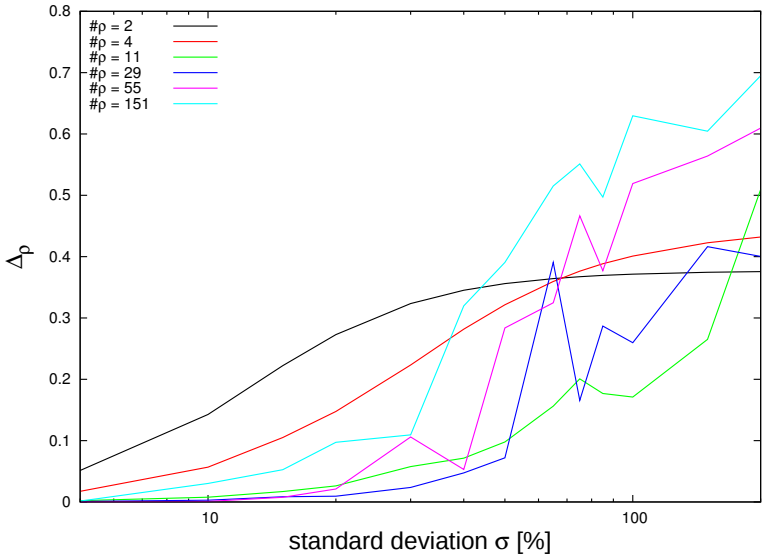



**Figure 1.14.:** Results of the sensitivity analysis. For each tree level (labeled by the respective numbers of parameters $\#\rho$ of each level) we computed the average deviation $\Delta_\rho$ [RT/Å$^2$] of fitted and initial interaction potentials for varying standard deviations $\sigma$.

were also revealed for the results of the extremum potentials, i.e. the most rigid and flexible bond types. As measure for the goodness-of-fit, we correlated both initial and fitted parameters and resulting matrices and computed the respective parameter deviations.

In summary, we have shown the capability to estimate interaction potentials that are implicitly given by a set of displacement vectors and discussed the performance of both MLE and MAPE. The fitting approach depends on the number of different interaction types only rather than on the number of snapshots. Computational time is influenced by the protein size as well since in each iteration eigenvalues of the Hessian matrix need to be derived by SVD (see section 1.1). The algorithm performed best for the smallest number of parameters to be estimated from a fully parametrized set of configurations. Due to its shortcomings, we omit the application of MAPE to experimental data, i.e. snapshots from MD simulation.

### 1.3.3 Semidefinite Programming

In the previous sections 1.3.1 and 1.3.2, we discussed two complementary approaches to determine the “optimal” interaction potentials weighting amino acid contacts from a set of structures fluctuating around an energy minimum state. As optimization criterion for MC-STUN (section 1.3.1), we employed a minimal distance of Hessian or covariance matrix in terms of FN (Eq. 1.26) computed between original matrices and those computed with fitted parameters. Likelihood based parameter estimation approaches aim to maximize either the likelihood (Eq. 1.34) or the posterior function (Eq. 1.36) by invoking prior knowledge. Here, we will describe the last method we examined to retrieve interaction potentials from experimental data

which is formulated as a mathematical optimization problem. In general, an optimization problem is defined as [Boyd & Vandenberghe, 2004]:

$$\begin{aligned} &\text{minimize} && f_0(\vec{x}) \\ &\text{subject to} && f_i(\vec{x}) \leq b_i \qquad \text{for } i = 1, \ldots, m \end{aligned} \tag{1.42}$$

Here, the vector $\vec{x} = (x_1, \ldots, x_N)$ is the $N$-dimensional optimization variable. We call $\vec{x}^*$ optimal or a solution to the specified optimization problem, if the objective function $f_0 : \mathbb{R}^N \rightarrow \mathbb{R}$ is minimal for all vectors that satisfy the constraint functions $f_i : \mathbb{R}^N \rightarrow \mathbb{R}$ for $i = 1, \ldots, m$:

$$f_0(\vec{x}^*) \leq f_0(\vec{x}) \qquad \text{with } f_i(\vec{x}^*) \leq b_i \quad i = 1, \ldots, m \tag{1.43}$$

The limits or boundaries for the (inequality) constraint functions $f_i$ are given by $b_i$. Optimization problems are classified due to particular forms of both objective and constraint functions. An important class of optimization problems for which effective solution methods have been developed is called linear programming. This class comprises problems whose objective and constraint functions are linear, and, hence, satisfy the following condition for constants $\alpha, \beta \in \mathbb{R}$ and variables $\vec{x}, \vec{y} \in \mathbb{R}^N$:

$$f_i(\alpha\vec{x} + \beta\vec{y}) = \alpha f_i(\vec{x}) + \beta f_i(\vec{y}) \tag{1.44}$$

A more general class that also covers linear programming comprises convex optimization problems. Here, a convex form of objective and contraint functions is given, i.e. the following condition is satisfied for $\alpha, \beta \in \mathbb{R}^+, \alpha + \beta = 1$:

$$f_i(\alpha\vec{x} + \beta\vec{y}) \leq \alpha f_i(\vec{x}) + \beta f_i(\vec{y}) \tag{1.45}$$

In contrast to linear programming the condition on $f$ is weakened, the rather strict equality (Eq. 1.44) is replaced by an inequality (Eq. 1.45) that has to be fulfilled only for distinct values of $\alpha$ and $\beta$.

Semidefinite programming (SDP) belongs to the class of convex optimization problems. Semidefinite programs are a generalization to linear programs, but similarly hard to solve. Applications of SDP are mainly found in research to improve optimization algorithms [Alizadeh, 1995; Ben-Tal & Nemirovski, 2002; Scherer & Hol, 2006; Vandenberghe & Boyd, 1996], but also in the field of pattern recognition [Biswas *et al.*, 2006; Weinberger & Saul, 2006], network theory [Bertsimas & Sim, 2003; Karger *et al.*, 1998; Srivastav & Wolf, 1998] and life sciences [Mazziotti, 2004]. Here, we are concerned with the minimization of a linear function which is subject to the constraint that an affine combination of symmetric matrices $F(x)$ is positive semidefinite termed as linear matrix inequality $F(x) \succeq 0$. Hence, SDP is formulated in general as [Vandenberghe & Boyd, 1996]:

$$\begin{aligned} &\text{minimize} && \vec{c}^T\vec{x} \\ &\text{subject to} && F(\vec{x}) \succeq 0 \\ &\text{where} && F(\vec{x}) \triangleq F_0 + \sum_{i=1}^{m} x_i F_i \end{aligned} \tag{1.46}$$

Both the vector $\vec{c} \in \mathbb{R}^m$ and the $m+1$ symmetric matrices $F_0, \ldots, F_m \in \mathbb{R}^{n \times n}$ are problem data.

### Fitting with SDP

Determining the underlying interaction potentials of a Hessian matrix, that weight the strength of amino acid contacts in a protein, can be considered as a matrix norm minimization problem as suggested by Vandenberghe & Boyd [1996]. Therefore, the general form of SDP (Eq. 1.46) is reformulated and includes a slack variable $t$:

$$\begin{aligned} &\text{minimize} && t \\ &\text{subject to} && \begin{bmatrix} tI & A(\vec{\rho}) \\ A(\vec{\rho})^T & tI \end{bmatrix} \geq 0 \end{aligned} \tag{1.47}$$

SDP is employed to minimize the matrix norm $||A(\vec{\rho})||$ of a matrix $A(\vec{\rho}) = A_0 + \rho_1 A_1 + \cdots + \rho_k A_k$, $A(\vec{\rho}) \in \mathbb{R}^{p \times q}$ with the optimization variable $\vec{\rho} \in \mathbb{R}^k$, i.e. amino acid specific interaction potentials, and an additional slack variable $t \in \mathbb{R}$. In this special case, the matrices $A_i$ need not to be symmetric. $I$ represents the identity matrix. The SDP has dimensions $m = k + 1$ and $n = p + q$. When applying SDP to the Hessian-fitting problem, we are concerned with the yet unknown interaction potentials $\rho_i$ and their respective "filtered" partial Hessians $A_i$. Such filtered matrices $A_i$ are computed as a Hessian matrix according to Eq. 1.5 but restricted on entries of amino acid contact type $i$. Since we intend to minimize the distance between the externally provided Hessian matrix $A_0$ and the Hessian based on estimated parameters, the matrix norm subject to minimization is the following:

$$A(\rho) = A_0 - \sum_{i=1}^{k} \rho_i A_i \tag{1.48}$$

The slack variable $t$ which is to be minimized can also be interpreted as the fitting error. The external Hessian matrix denoted with $A_0$ is obtained either as the pseudoinverse of the covariance matrix computed from a set of structures or derived from NMA. For each amino acid contact type $i$ a filtered Hessian matrix $A_i$ which contains only information on spatial distances of the involved residues is computed according to ANM theory (see section 1.1). The corresponding interaction potentials $\rho_i$ are obtained from SDP application. In the following, we perform a proof-of-principle analysis for SDP based parameter fitting and discuss its performance.

#### SDP : Proof of Principle

To test whether SDP is capable to extract interaction potentials of a Hessian matrix, we applied SDP to an artificially constructed ANM-Hessian (Eq. 1.5) of BPTI (PDB code 6PTI [Wlodawer *et al.*, 1987]) as we did for the MC-STUN approach (see section 1.3.1). Analogously, we weighted non-covalent interactions of amino acids within a distance of $r_c = 13$ Å according to MJ contact potentials and peptide bonds by 82 RT/Å$^2$. The filtered Hessians necessary for SDP were constructed using the same structure and contact definition. In a subsequent sensitivity analysis, we investigated the influence of the cutoff distance $r_c$ on the quality of the derived in-

teraction potentials since we intend to employ SDP fitting for data from various processes, such as MD simulations, NMA, Gō-models [Taketomi *et al.*, 1975], etc.. To this end, we varied both the "construction cutoff" $r_c^{(c)}$ which is used for the initial (experimental) Hessian matrix and the "fitting cutoff" $r_c^{(f)}$ which is employed for the computation of the filtered Hessian matrices needed for SDP in the interval $[7.5\ \text{Å}, 20\ \text{Å}]$ with a step size of 0.5 Å. Tackling the overfitting issue, we repeated the analysis for varying sizes of the "construction" and "fitting" alphabet as well. Therefore, reduced alphabets were created according to the alphabet tree that was proposed by Pape *et al.* [2010] (see section 4.2) and employed for the stepwise likelihood based parameter estimation routines (see section 1.3.2). Thus, we can assess the quality of applying SDP based fitting utilizing a more abstract amino acid alphabet. We employed the described variations combined at once.

As SDP implementation we used the stand-alone command-line SDPA software version 7.3.1 [Fujisawa *et al.*, 2008], data preparation was performed in R [R Development Core Team, 2008].

### SDP : Results

Using the same BPTI structure and contact definition, SDP was capable to perfectly retrieve the interaction potentials that were used for construction of the Hessian matrix. No deviation of initial and determined parameters is noticeable, the slack variable $t$, which can be interpreted as the fitting error, having a negligible value of $3.718 \cdot 10^{-5}$.

We furthermore investigated the influence both of differing cutoffs and reduced amino acid alphabets on the outcome of SDP based fitting of contact potentials. Note that for quantifying the impact of alphabet reductions, we may only evaluate the results obtained for fitting Hessians that have been constructed using a larger number of symbols with a reduced alphabet since deriving more detailed parameters by applying fitting routines is impossible.

To estimate the influence of the cutoff distance, we used a construction cutoff $R_c^{(c)}$ to compute the Hessian matrix based on a given amino acid alphabet that is used as input for SDP. To derive the underlying interaction potentials, we used an alternate cutoff definition $r_c^{(f)}$ for determining contacts in the protein. Here, we used the same number of symbols and, thus, interaction types for both construction and fitting. The two distinct cutoff distances were varied and we determined the respective differences of the peptide bond parameter. The results are shown in Fig. 1.15 for the complete amino acid alphabet, where we plotted the deviation of the fitted from the initial contact potential as a function of the cutoff difference $\Delta r_c = r_c^{(c)} - r_c^{(f)}$. In all cases, the fitted peptide bond parameter was larger than the original value of 82 RT/Å$^2$. If the cutoff definitions used for input and fitting are equal, the exact parameter is obtained. Interestingly, deviations that become noticeable for differing cutoffs are not independent of whether $r_c^{(c)} > r_c^{(f)}$ or vice versa. If the construction cutoff distance $r_c^{(c)}$ is smaller than the one used for fitting ($\Delta r_c < 0$), the contact potential of covalent amino acid contacts can be estimated with minor deviations only. In contrast, using a smaller contact defining distance for fitting than for construction, the dynamics of the protein cannot completely be represented by the smaller system, leading to an overestimation of the peptide bond. Hence, the error is getting larger for larger cutoff deviations. In addition, we examined the relationship of the slack variable $t$ that

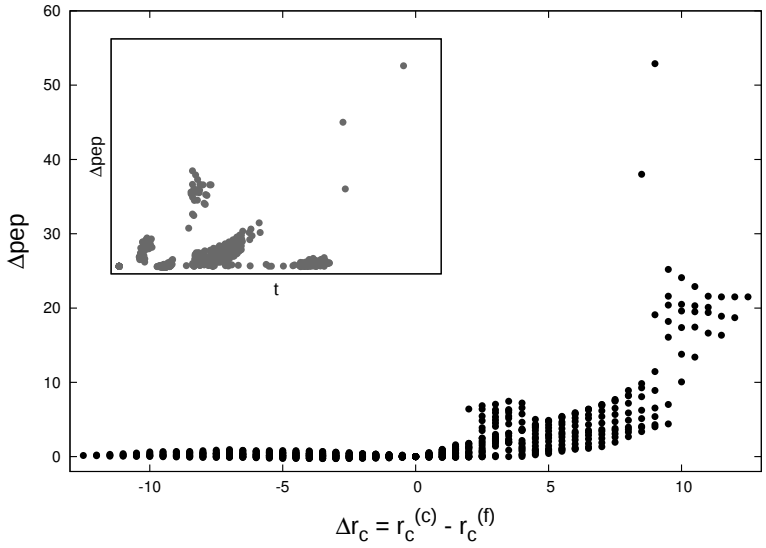
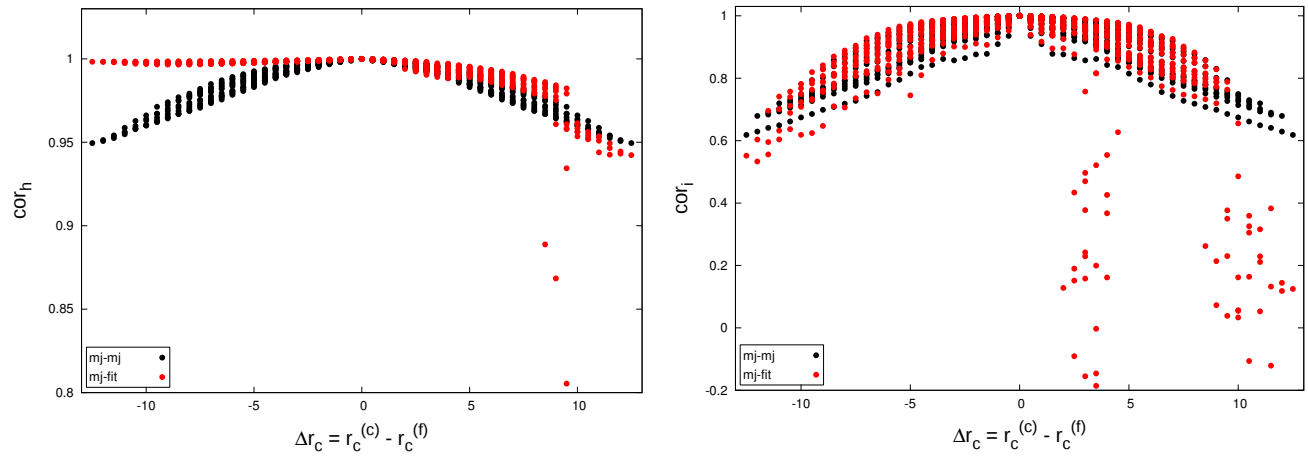



**Figure 1.15.:** Quality of the interaction potential of the peptide bond $\Delta_{\text{pep}}$ in dependence of the utilized cutoff definitions for construction $r_c^{(c)}$ of the intial Hessian matrix and for fitting $r_c^{(f)}$. The deviation of the parameter is shown for $\Delta r_c = r_c^{(c)} - r_c^{(f)}$. The inset depicts a scatterplot of the results of peptide bond deviation and the corresponding slack variable $t$ obtained by SDP. We omitted the absolute scale to illustrate the relationship of the two values only. Results are shown for the complete amino acid alphabet.

is to be minimized by SDP and the corresponding deviation of peptide bond parameters, which is illustrated in the inset of Fig. 1.15. Although smaller values of $t$ imply smaller deviations of peptide bond parametrization, we find no direct correlation of the two variables. Hence, the slack variable cannot necessarily be utilized to judge the quality of fitting. We yield similar results for reduced amino acid alphabets as well (data not shown). Considering the absolute deviations of fitted interaction potentials of non-covalent contacts, an analogous picture reveals itself.

By fitting interaction potentials of a Hessian matrix, we obtain a description of protein dynamics which is represented by the pseudoinverse of the Hessian, the covariance matrix (see Eq. 1.6), which correlates the fluctuations of residues of all spatial directions. Hence, evaluating the quality of the fit implies evaluating how well the dynamics of the protein is approximated. To this end, we computed the correlation coefficient between original and fitted Hessian and covariance matrices, respectively. Fig. 1.16 shows the Pearson correlation coefficient computed for the complete amino acid alphabet using differing cutoff distances for construction and fitting. Again, similar results are obtained for reduced alphabets as well. Additionally, we computed the correlation for both Hessian and covariance matrices that were derived by varied cutoff distances and the same MJ parametrization scheme to work out the effect of the contact definition on an "ideal" fit. Applying MJ potentials for initial and "fitted" matrices yields a symmetric decrease of correlation depending on the cutoff difference. In general, we notice that the covariance matrix is more sensitive to changes in the parametrization and, thus, to the underlying contact definitions, which has already been noted in the context of MC-STUN (section 1.3.1). The correlation computed for the Hessian matrix exhibits a similar behavior as was previously detected for the peptide bond deviation. Using a larger cutoff distance for the fit results in a




**(a)** Hessian **(b)** Inverse

**Figure 1.16.:** Pearson correlation coefficient is plotted for differing cutoff distances used for the constructed and the fitted Hessian matrix. The correlation is plotted for (a) the Hessian and (b) for the covariance matrices in dependence of $\Delta r_c$. The black dots show the correlation of matrices computed with the same parametrization by differing contact definitions, whereas the red dots describe the correlation of the constructed matrix and the fitted matrix which is computed with fitted interaction potentials.

Hessian matrix with nearly perfect correlation to the original Hessian matrix. We notice a higher correlation if we construct the "fitted" Hessian with interaction potentials obtained from SDP instead of MJ parameters. In contrast, if the original Hessian matrix is computed with a larger cutoff distance, hence, providing more contacting residue pairs, the correlation with the matrix fitted with a smaller cutoff decreases with increasing $\Delta r_c$, albeit the minimum correlation that is observed is larger than 0.8. Regarding the correlations computed for the covariance matrices, we notice a significant decline the more the cutoff definitions deviate from each other. Whereas we observe a steady decline of linear correlation down to about 0.6 for $\Delta r_c < 0$, we even find negative correlations for larger $\Delta r_c$ values. The same holds for Spearman correlation and further alphabet definitions as well. Thus, the cutoff chosen to extract interaction potentials from a given Hessian matrix should preferably be equal to or larger than the contact definition that has been used to generate this Hessian matrix.

Examining the influence of reduced amino acid alphabets, we used the same cutoff distance for constructing the initial Hessian and fitting but varied the number of amino acid types according to the reduction scheme proposed by Pape *et al.* [2010]. The impact on the quality of the fitted results is shown in Fig. 1.17 for the Pearson correlation of the Hessian matrices for varying cutoff distances. Again, we computed the correlation of the original Hessian to the Hessian computed with averaged MJ or fitted parameters. In the first scenario, we find a slight decrease for larger cutoff distances for more reduced alphabets presumably due to the fact that the averaged parameters include interactions that are not found in BPTI. However, the impact is negligible since we are concerned with a correlation of more than 0.99. The same holds for the correlation of covariance matrices as well (data not shown). Although the observed correlation for fitted interaction potentials is larger than 0.97 as well, we notice a dependence of the cutoff distance which may arise from the number of amino acid types present for the respective

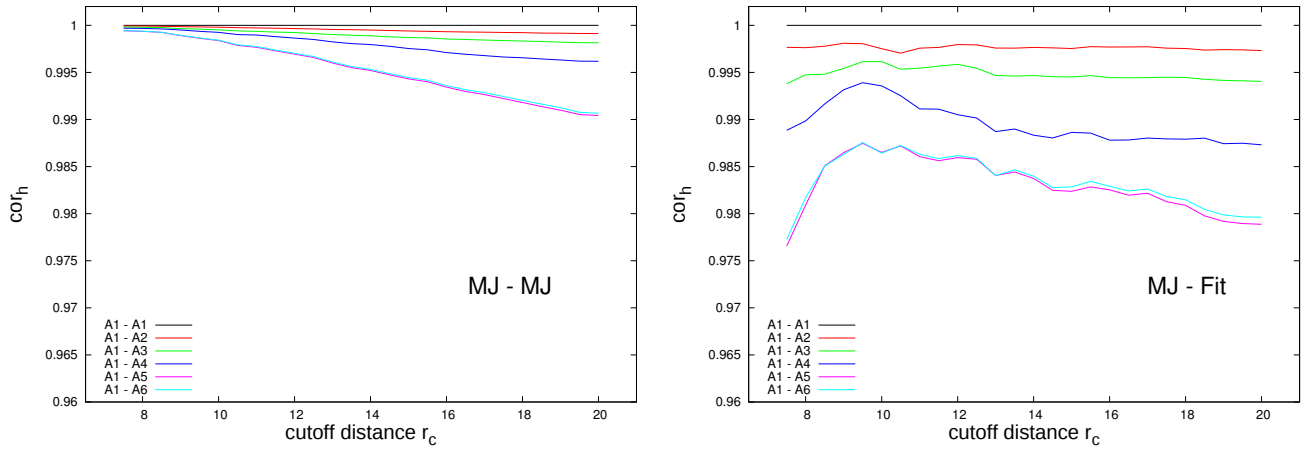



**Figure 1.17.:** Correlation of initial and "fitted" Hessian matrix is shown in dependence on the applied distance cutoff $r_c$ [Å]. On the left, the interaction potentials used to construct the "fitted" Hessian matrix are simply averaged MJ potentials, whereas on the right we used the interaction potentials obtained by SDP. The complete amino acid alphabet (A1) comprising 20 symbols was subsequently reduced to an alphabet containing one symbol only (A6) using the reduction scheme proposed by Pape *et al.* [2010].

contact definitions. Clearly, we find that reducing the number of amino acid symbols from 20 to 1 for fitting yields a slightly reduced correlation of the Hessian matrix with respect to the initial Hessian. Other than for the cutoff definition, the correlation of the covariance matrix as well as the deviation of peptide bond parameters differ to a minor extent only indicating the fitting routine to be less sensitive towards alphabet reductions than to differing cutoff definitions.

Here, we demonstrated the capability to determine the exact interaction potentials from a constructed Hessian matrix using the mathematical concept of SDP. Due to an efficient implementation and a small number of iterations the parameters were obtained expeditiously for the employed BPTI. The runtime depends on both the number of interaction parameters and the system size. In addition, the influence of differing cutoff definitions and reduced alphabets was investigated. Reduced numbers of residue types showed only a minor contribution to the resemblance of initial protein dynamics whereas the cutoff definition used for determining interaction potentials needs to be chosen carefully since fitting cutoffs smaller than the initial cutoff distance may lead to worse fitting results.

### 1.3.4 Summary

In this study, we presented three approaches based on a coarse-grained protein network model to extract interaction potentials that weight the strength of amino acid contacts. Although we may encounter anharmonicity in realistic applications, interactions of amino acids are modeled as harmonic springs connecting residues within a defined cutoff distance $r_c$. Following ANM theory [Atilgan *et al.*, 2001], amino acids are represented as beads placed on their respective $C_\alpha$ atom neglecting side chains. MC-STUN (section 1.3.1) is a stochastic algorithm that searches the best parameters by minimizing the FN computed for the original matrix and a matrix computed with newly derived parameters. The STUN extension provides a better convergence since un-

interesting regions of the parameter space are tunneled. MLE/MAPE (section 1.3.2) is derived from statistical learning theory and accurate to determine the parameters that best describe the data generated from a given process based on maximizing the likelihood function. MAPE invokes prior knowledge in contrast to MLE. Finally, SDP (section 1.3.3) is a class of mathematical optimization problems and is applied to minimize the matrix norm of a given and fitted positive semidefinite matrix. Notably, MC-STUN is the only method to determine amino acid specific interaction potentials directly from protein dynamics as given by the covariance matrix $C$.

For all approaches, we have proven the capability to reconstruct the force constants weighting residue interactions using an artificially generated protein system by help of ANM theory (see section 1.1). Although MC-STUN and the likelihood based approaches were able to estimate the contact potentials in a good approximation, exact reconstruction of the input parameters was achieved by SDP alone. Shortcomings were detected for MLE since invoking prior information improved the results drastically. The sensitivity analysis applied for diverse prior definitions revealed a strong dependence of the quality of estimated parameters on a well-designed prior. Application of both MLE and MAPE for a reduced amino acid alphabet comprising a single amino acid type and, thus, two interaction types (covalent, non-covalent) was able to approximate the peptide bond potentials close to the input, whereas for the complete alphabet of 20 symbols the multidimensional root-finding algorithm exhibited worse convergence behavior. Likelihood based approaches are preferred for extracting a reduced number of interaction potentials only. MC-STUN was able to approximate the contact potentials with a quality similar to MAPE for fitting based on both Hessian and covariance matrix. Parameter estimation based on MC-STUN depends on the internal algorithm parameters that determine the convergence of the algorithm. Since the choice of these parameters is not obvious, various settings need to be applied before the actual fitting can be initiated. Even sophisticated presampling of internal parameters does not necessarily prevent the algorithm from getting trapped in a local minimum. Covariance based fitting is limited by the protein size as each iteration requires computation of the pseudoinverse of the Hessian matrix which is the most time-consuming step of the algorithm. Preferentially, the number of iterations of stochastic algorithms is not bounded which increases the runtime dramatically. Due to those shortcomings the approaches of estimating interaction potentials using MC-STUN and MLE/MAPE are not further pursued. In contrast to those conventional methods, the mathematical optimization problem SDP has shown to efficiently derive the exact interaction potentials of an artificial protein system without the need of additional parameters. Therefore, we restrict the estimation of amino acid specific potentials from data stemming from MD simulations and NMA on SDP based fitting.

## 1.4 Parameter Fitting – Application

In the previous section 1.3 we investigated the capability of three complementary approaches to retrieve interaction potentials from an artificially constructed ANM system. Since the optimization problem SDP (section 1.3.3) yielded encouraging results, we will apply SDP to data sampled from MD and NMA. To this end, we performed MD simulations for BPTI and PA as well as NMA for BPTI. Both proteins are special cases, since BPTI is a rigid protein exhibiting high stability, whereas the short polypeptide PA is extremely flexible. In the following, we discuss the data preparation of the proteins and the application of SDP to it.

### 1.4.1 Data Generation for BPTI

Due to its small size and its stability, BPTI was used as a model protein for numerous experimental and theoretical approaches. Amongst others, structure determination utilizing X-ray diffraction and neutron scattering methods was explored by Wlodawer *et al.* [1984], solution nuclear magnetic resonance (NMR) techniques were employed by Wüthrich *et al.* [1982]. In addition, folding pathways have first been described first for BPTI [Goldenberg & Creighton, 1984; Makhatadze *et al.*, 1993]. Pioneering theoretical studies were published by McCammon *et al.* [1977], who were able to perform MD simulations for BPTI to investigate protein dynamics *in silico*. BPTI was also chosen as model protein of early applications of NMA to describe internal protein motions [Brooks & Karplus, 1983]. Since the protein is extremely stable against denaturation [Moses & Hinz, 1983], we employ the small BPTI as model system for developing and testing of routines that extract amino acid specific interaction potentials from a set of conformations in the extreme case of high stability.

#### Setup of MD Simulations

To sample conformations of BPTI by MD simulations, we used the crystal structure that is deposited in the PDB [Berman *et al.*, 2000] with entry code 6PTI [Wlodawer *et al.*, 1987]. We deleted all hetero atoms and alternative atom positions of residues. The protein structure was preprocessed by the `psfgen` plugin of VMD [Humphrey *et al.*, 1996], missing atoms were added. A total of 3,259 water molecules as well as ions in physiological concentration (5 $Na^+$ and 5 $Cl^-$) were added into a simulation box with box vectors of about 4.5 nm. Coulomb and Lennard-Jones interactions were defined up to a distance of 1.2 nm, but their strengths were decreased starting from 1.0 nm. After setting up the system, we ran a short minimization of 20 ps to ensure correct structure, i.e. no clashes due to misplaced atoms or improper bonds. During the equilibration run (50 ps), constraints for the amino acid positions were added to freeze the protein while water molecules adapt to the simulation temperature. From the equilibration step, we obtained two structures that are used for further simulations: the final structure (MD1) and an intermediate structure (MD2) that has been subject to a short minimization (about 100 fs). After both structures have been simulated for 600 ps, we started a total of 29 independent production runs of 1.8 ns each (14 based on MD1, 15 based on MD2) to generate a set of BPTI conformations. Due to water relaxation insufficiencies, not all simulations finished correctly resulting in shorter trajectories (see Tab. 1.8). MD simulations, minimization and equilibration runs were performed using NAMD [Phillips *et al.*, 2005] and an all-atom additive CHARMM force field [MacKerell *et al.*, 1998, 2004]. We simulated the protein at a temperature of 310 K, temperature and pressure were rescaled during the simulation.

#### Setup of NMA

As discussed in section 1.1, NMA is a well-established tool that has often been employed to explore internal protein motions [Brooks & Karplus, 1983, 1985; Brooks *et al.*, 1995; Cui & Bahar, 2006; Janezic & Brooks, 1995; Janezic *et al.*, 1995; Kidera & Gō, 1992; Levitt *et al.*, 1985;

| no. | # frames | no. | # frames | no. | # frames | no. | # frames | no. | # frames |
|---|---|---|---|---|---|---|---|---|---|
| MD1 1 | 6,626 | MD1 7 | 18,000 | MD1 13 | 17,316 | MD2 5 | 9,822 | MD2 11 | 5,733 |
| MD1 2 | 16,260 | MD1 8 | 18,000 | MD1 14 | 18,000 | MD2 6 | 18,000 | MD2 12 | 18,000 |
| MD1 3 | 9,708 | MD1 9 | 13,924 | MD2 1 | 5,446 | MD2 7 | 18,000 | MD2 13 | 18,000 |
| MD1 4 | 6,445 | MD1 10 | 18,000 | MD2 2 | 13,234 | MD2 8 | 14,719 | MD2 14 | 9,277 |
| MD1 5 | 16,678 | MD1 11 | 18,000 | MD2 3 | 17,294 | MD2 9 | 4,158 | MD2 15 | 18,000 |
| MD1 6 | 10,992 | MD1 12 | 18,000 | MD2 4 | 4,635 | MD2 10 | 8,313 | | |

**Table 1.8.:** Overview of the MD simulations that were performed for BPTI. For each trajectory, which is enumerated according to its start structure (MD1, MD2), the number of generated conformations is given. In total, we obtain 388,580 BPTI conformations, i.e. 205,949 structures for MD1 and 182,631 for MD2 based simulations, respectively.

| step | integrator | tolerance |
|---|---|---|
| 1 | `steep` | 1000 |
| 2 | `cg` | 0.1 |
| 3 | `l-bfgs` | $10^{-3}$ |
| 4 | `l-bfgs` | $10^{-5}$ |
| 5 | `l-bfgs` | $10^{-7}$ |
| 6 | `l-bfgs` | $10^{-9}$ |

**Table 1.9.:** Protocol for the energy minimization of BPTI plus water and ions (NM2). The tolerance of the maximum force is given in [kJ mol$^{-1}$nm$^{-1}$].

Nojima *et al.*, 2002; Tama & Sanejouand, 2001; van Vlijmen & Karplus, 1999]. In the following, we present two NMA settings: in the first approach (NM1) we determine normal modes for the protein *in vacuo* starting from its crystal structure (6PTI), whereas the second approach (NM2) includes water molecules and ions as well. Prior to any computation of the normal modes, a thorough minimization protocol has to be applied to the protein in order to eliminate any forces acting on it. For the minimization and calculation of the normal modes, routines from the GROMACS software suite [van der Spoel *et al.*, 2005] were used. For minimization, GROMACS provides three minimizers: steepest descent (`steep`), conjugate gradient (`cg`) and low-memory Broyden-Fletcher-Goldfarb-Shannon (`l-bfgs`). With `steep`, improper bonds and clashes leading to high forces that act on single atoms are eliminated efficiently. In contrast, the `cg` minimizer performs best whenever the biomolecular system is close to an energy minimum. By successively creating better approximations and, thus, moving the system towards the closest minimum, the `l-bfgs` minimizer which is based on inverse Hessian approximations converges faster than `cg`. Hence, its use is required in the last steps of minimization prior to NMA. We define a tolerance of $10^{-9}$ kJ mol$^{-1}$nm$^{-1}$ for the maximum force to act on a single atom.

For NM1, we reduced the maximum force to the defined tolerance within a single run using the `l-bfgs` integrator. Due to the number of water molecules, it was hardly feasible to reduce the maximum force for NM2 to the required value. To circumvent this problem, we increased the cutoff distance for van der Waals and Coulomb interactions. By employing the developed minimization protocol (see Tab. 1.9), the system energy was properly minimized.

From NMA we obtained a mass-weighted Hessian matrix of dimension $3M \times 3M$ with $M$ being the number of atoms of the biomolecular system. To derive interaction potentials by use of ANM based parameter fitting methods, this matrix has to be transformed into a $C_\alpha$-based Hessian.

This can be achieved either by extracting the respective elements from the pseudoinverse or by reducing the Hessian matrix. Initially, all atom weights have to be removed for either method. Extraction of components from the covariance matrix is feasible for small matrices only. Thus, we will discuss the reduction of the all-atom Hessian matrix in the following. In a previous study, Eom *et al.* [2007] described a reduction scheme for Hessian matrices. Thereby, the potential energy $E$ is expressed in terms of:

$$E = \frac{\gamma}{2} u^T H u = \frac{\gamma}{2} \begin{bmatrix} u_1^T & u_2^T \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{1.49}$$

We discriminate between the so-called "master"- and "slave"-residues. It is assumed that fluctuations of slave residues $u_2$ are not significant, thus, these residues are in an equilibrium state of effectively vanishing force. The total fluctuation is represented by the fluctuation of master residues $u_1$ only. Thus, the Hessian matrix is divided into four parts. $H_{11}$ and $H_{22}$ represent pairwise interactions within the individual groups, whereas $H_{12}$ and $H_{21}$ contain values for interactions between master and slave residues. A minimization of the potential energy in terms of slave residues $u_2$ leads to:

$$\frac{\partial}{\partial u_2} E \stackrel{!}{=} 0 \tag{1.50}$$

After algebraic transformations, we obtain the following formula, which can be applied to convert the all-atom Hessian $H$ to a $C_\alpha$-only matrix $\tilde{H}$ that implicitly describes the mechanics of the full protein in case of NM1 and, furthermore, of water and ions in case of NM2.

$$\tilde{H} = H_{11} - H_{12} H_{22}^{-1} H_{21} \tag{1.51}$$

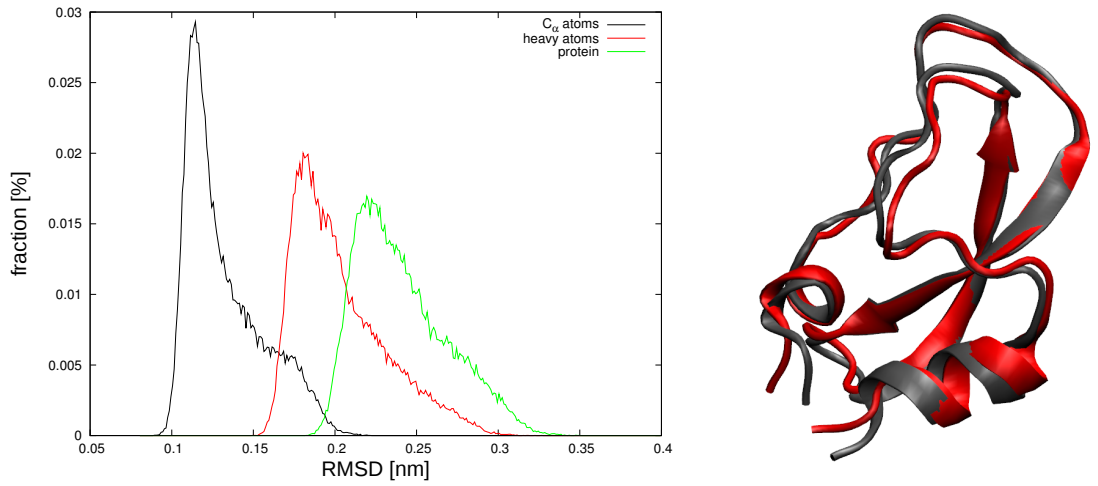
Due to memory restrictions for large matrices, the reduction formula was successively applied to the Hessian after identification of master and slave residues and a corresponding reordering of the matrix. In each iteration step, blocks of size $\leq 501$ of slave residues were eliminated from the full matrix $\Gamma$. Since the effect of the removal order is not obvious, we compared three different scenarios to reduce the all-atom to a $C_\alpha$-only Hessian matrix:

R1. Water is eliminated from the end of the matrix, block by block, and, as a last step, we remove the ions.

R2. We remove ion entries first. Afterwards, the water is eliminated blockwise starting from the end of the matrix.

R3. We first eliminate the water per components (z, y, x in this order), the ions are removed afterwards.

To proof the correctness of the suggested procedures, we first applied the whole protocol to a smaller system. We used a small peptide comprising 7 residues that was extracted from the PDB with code 3HYD [Ivanova *et al.*, 2009]. We added 248 water molecules and a single $Na^+$ ion. The system was energy minimized in three steps by subsequently employing the integrators `steep`, `cg` and `l-bfgs`. Afterwards, NMA was performed and the all-atom Hessian matrix was computed. For the computation of the $C_\alpha$-Hessian, we considered the reduction scenarios




**(a)** RMSD **(b)**

**Figure 1.18.:** Root mean square deviation (RMSD) is computed for all structures of the MD simulations with respect to the crystal structure after a superpositioning of both structures. The resulting RMSD histograms are shown for $C_\alpha$ atoms, heavy atoms and the whole protein (a). Furthermore, two structures obtained from the left and the right end of the histogram are depicted to reveal structural differences (b). Graphics of BPTI structures were generated using VMD [Humphrey *et al.*, 1996].

R1 and R3 to evaluate the effect of different water removal schemes only. We compared the resulting $C_\alpha$-Hessians as well as the corresponding covariance matrices.

## Results

Each conformation of BPTI in the MD trajectories was extracted from the respective trajectory and minimized. Hereafter, we computed the root mean square deviation (RMSD) of each resulting structure with respect to the crystal structure. RMSD measures the average deviation of two structures after a least-squares fit, which is performed to minimize the differences between both structures by superimposing them. We determined the RMSD for $C_\alpha$ and heavy atoms as well as for the whole protein, superpositioning was performed for the respective group. In the resulting histograms (see Fig. 1.18) we observe a single peak for each considered group indicating that many structures share a similar deviation to the reference structure. Thus, it seems that all sampled configurations of BPTI fluctuate around a common energy minimum state. However, we also notice a slight buckling on the right end of the histograms indicating that the sampled structures may belong to two equilibrium states. To ensure that all structures are sampled for a single state only, we picked two distinct snapshots with a small and a high RMSD relative to the reference structure. From the superpostion that is shown in Fig. 1.18, we notice structural deviations particularly for loop regions, whereas the secondary structure elements are nearly identical. Since we detect most of the deviations in the sidechains of the amino acids, we conclude, that all sampled structures of BPTI describe a single ground state, which is in perfect agreement with previous studies that proposed a high stability for BPTI [Moses & Hinz, 1983]. For further analysis, we omitted the first 400 ps of each trajectory to exclude structures of the initial relaxation phase (see Fig. 1.19).

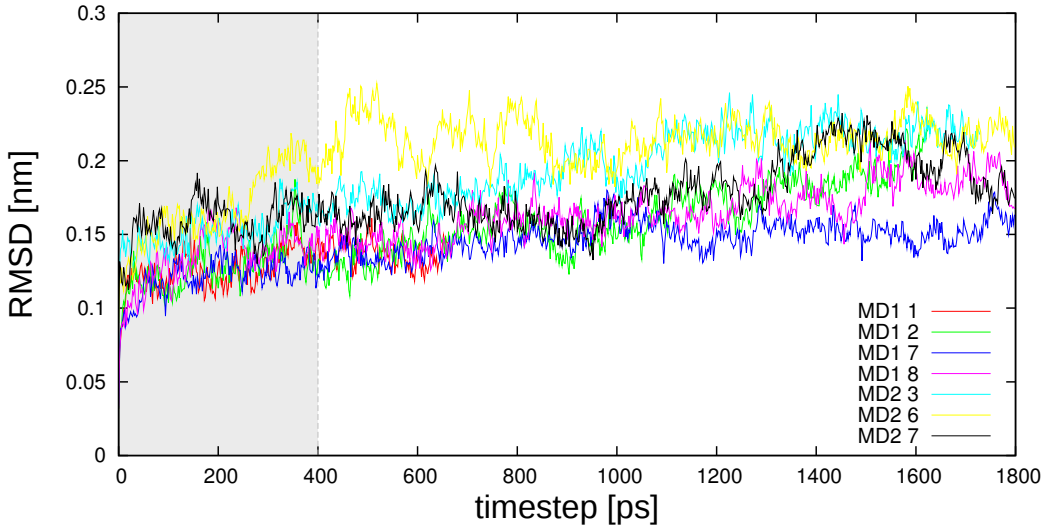



**Figure 1.19.:** RMSD over time plotted for selected trajectories based on MD1 and MD2. Structures of the initial simulation phase (<400 ps) omitted from analysis are marked by the gray box.

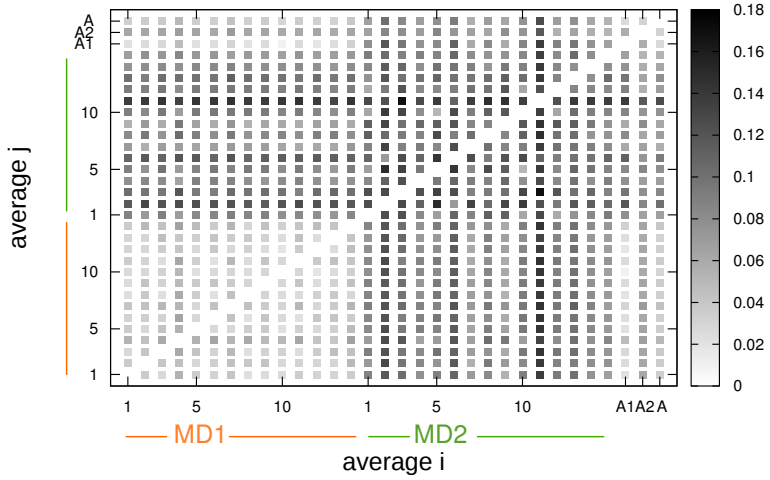



**Figure 1.20.:** Pairwise RMSDs [nm] of average structures from the truncated trajectories. We highlighted simulations based on MD1 (orange) and MD2 (green). Furthermore, we show results for combined trajectories: A1 contains all truncated trajectories based on MD1, A2 those based on MD2, and A contains both A1 and A2. Black points mark larger RMSD values indicating less similarity of the respective structures than for smaller RMSDs.

### Comparing Covariance Matrices from MD simulations

Since the quality of experimental data has a major impact on the quality of obtained parametrizations, we examined the different trajectories and the derived covariance matrices in more detail. For each simulation, the average structure was computed. We have shown above that all snapshots are fluctuations around this central configuration. Note that an average structure is not necessarily a consistent protein structure. We show the pairwise RMSD of all average structures computed after a least-squares fit in Fig. 1.20. We notice that trajectories based on MD1 are more similar than those based on MD2 that exhibit larger RMSD of the respective average structures. Nevertheless, the differences are negligible by a maximum RMSD of 1.8 Å for any two structures.

Covariance matrices that contain information on correlated motions of residues were computed with the GROMACS simulation package [van der Spoel *et al.*, 2005]. To this end, all snapshots were fitted to a reference structure to derive their respective fluctuations. In the following, we will discuss potential types of reference structures that can be used for the computation of covariance matrices. Since the fluctuations of the protein should be modeled for a configuration that corresponds to at least a local minimum of the PES, the crystal structure as well as a thoroughly minimized structure, e.g., the structure that was used as input for NMA, are alternatives. Both structures represent minimum configurations that need not be representative for the MD simulations, although only a single minimum configuration is assumed based on the results above. In addition, we determined the snapshot with a minimum potential energy from all simulations as a possible reference structure. Again, this configuration may not be representative for all simulations, since it was sampled from a single trajectory. We computed the covariance matrix for each trajectory based on this minimum structure as well as on the respective average structure. From the covariance matrices, we derived B-factors and correlated them to experimental B-factors. Atomic fluctuations based on the respective average structure showed a higher resemblance to experimental data than those obtained for the sampled minimum configuration (data not shown). In Fig. 1.21 we show a comparison of experimental and theoretical B-factors for selected trajectories. B-factors computed for MD trajectories exhibit a similar picture, i.e. the same residues are labeled flexible but on a differing scale. If we compare those theoretical results with the experimentally derived B-factors we notice larger deviations. Notably, a general dynamical behavior is observed. For the computation of a covariance matrix for a trajectory, we used the respective average structure. Similar to the pairwise RMSD computation for the average structures obtained from the MD runs, we compute the pairwise correlations of the resulting covariance matrices (see Fig. 1.22). Again, the ensemble generated on basis of MD1 shows a higher intrinsic similarity than the covariance matrices obtained for simulations based on MD2. Those effects are visible for the concatenated trajectories as well. Hence, the dynamics of the protein is not fixated throughout diverse sets of MD runs but rather subject to alterations.

### Comparison of NMA Setups and Reductions

To reduce all-atom Hessian matrices obtained by NMA to $C_\alpha$-only matrices, we proposed three reduction schemes. To avoid numerical issues, we performed the reduction on a small system first. The resulting $C_\alpha$ Hessians of the small peptide 3HYD for reduction schemes R1 and R3 are indistinguishable, and, consistently, show high correlations, i.e. a Pearson correlation coefficient of approximately one. In contrast to ANM constructed Hessian matrices, we do not find six eigenvalues equal to zero due to rotation and translation presumably due to numerical issues, but their values were considerably smaller than other eigenvalues (data not shown). Since a high correlation in the space of Hessian matrices implies high correlation for the respective covariance matrices, we compute the pseudoinverses by either using all or leaving out the first six eigenvalues and the resulting correlation of the covariance matrices. Note that correlation in the space of covariance matrices is achieved only when leaving out the first six eigenvalues that correspond to the six degrees of freedom although being larger than zero. In summary, the order how elements are removed from the all-atom Hessian has only a minor or no influence on the outcome. Minor deviations of resulting coarse-grained matrices are due to numerical issues. For BPTI (NM1, NM2), the reduction of NMA-Hessian again yields the same $C_\alpha$ matrix for either

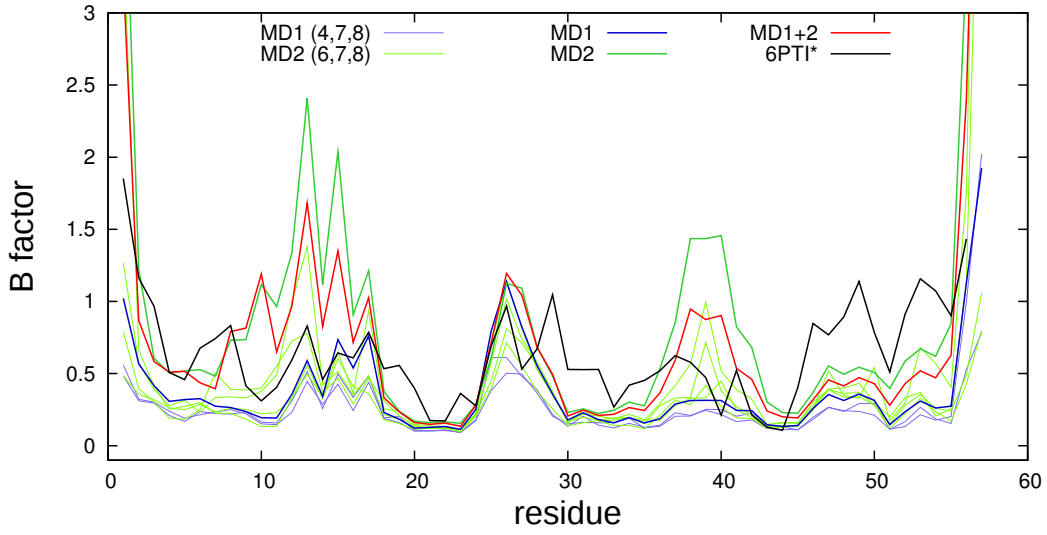



**Figure 1.21.:** Experimental (6PTI) and theoretical B-factors [Å$^2$] that were obtained from covariance matrices. The reference structure used for the computation is the respective average configuration of each trajectory. Results for selected trajectories based on MD1 (runs 4, 7, 8) and MD2 (runs 6, 7, 8) are displayed, as well as the concatenated trajectories MD1 and MD2. B-factors for MD1+2 are derived from a trajectory that contains both MD1 and MD2. Values marked with an asterisk are scaled to allow a better comparison of fluctuation patterns.

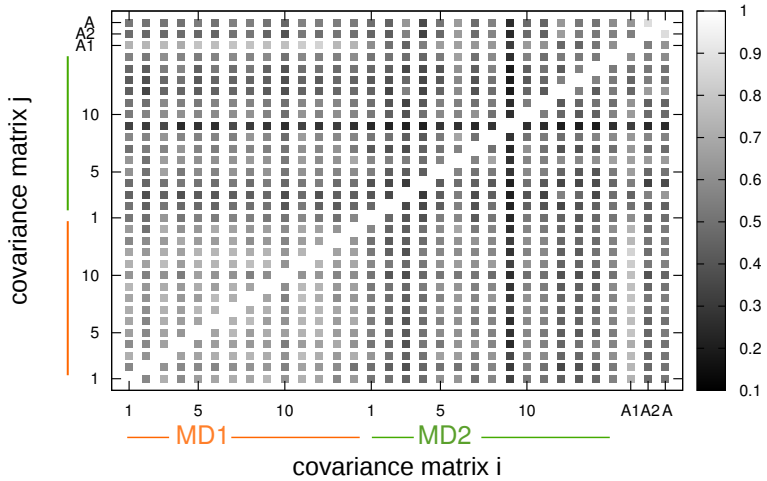
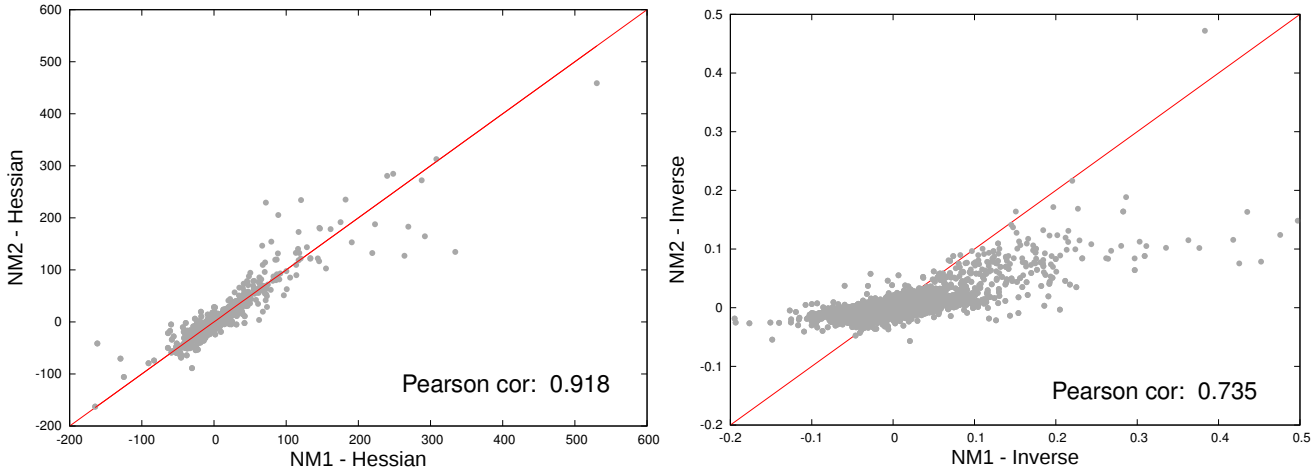



**Figure 1.22.:** Pairwise Spearman correlation coefficients of covariance matrices obtained from the truncated trajectories. The matrices were computed using the average structure of the respective trajectory as reference structure. We highlighted simulations based on MD1 (orange) and MD2 (green). Furthermore, we show results for combined trajectories: A1 contains all truncated trajectories based on MD1, A2 those based on MD2, and A contains both A1 and A2. Dark points mark low correlations indicating a higher diversity of the respective matrices, wheres white points denote more similar dynamics.




**(a)** Correlation for Hessian matrices.

**(b)** Correlation for covariance matrices.

**Figure 1.23.:** Scatterplot of reduced Hessian (a) and covariance (b) matrices obtained after reduction for both settings NM1 and NM2. The correlation indicated by Pearson correlation coefficients is significant with a p-value of $2.2 \cdot 10^{-16}$.

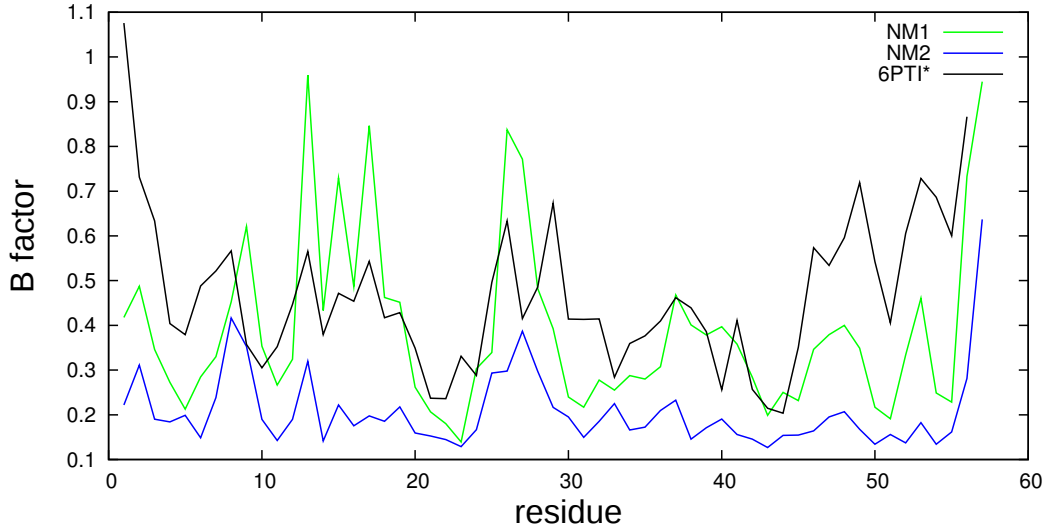



**Figure 1.24.:** Experimental B-factors [Å$^2$] from 6PTI [Wlodawer *et al.*, 1987] are plotted in comparison to B-factors that were obtained from the pseudoinverse of Hessian matrices NM1 and NM2. Values marked with an asterisk are scaled to allow a better comparison of fluctuation patterns.

reduction scenario. The block size used for either reduction scheme has no influence on the result as well (data not shown). Note that extraction of $C_\alpha$ elements of the pseudoinverse leads to a matrix whose eigenvalues are all greater than zero.

We are concerned with two NMA approaches differing in the respective protein environment. By reducing the all-atom Hessian obtained for NM2, the influence of fluctuations of water and ions is implicitly included in the fluctuations of the $C_\alpha$ atoms, whereas for NM1 no dynamical features of the environment are retained. To compare both approaches, we correlated Hessian and covariance matrices. The data shown in Fig. 1.23 are representative for all reduction

schemes. Obviously, we notice a high correlation for the derived Hessian matrices revealing only minor differences. However, for the covariance matrices the correlation is drastically reduced indicating a higher sensitivity of this matrix type to the environment of the protein. Hence, the correlated dynamics of protein residues is altered due to external effects. In addition, we compute B-factors from both covariance matrices and compare those with experimental B-factors as can be found in the PDB with code 6PTI [Wlodawer *et al.*, 1987] (see Fig. 1.24). On an absolute scale, NM2 exhibits a more rigid behavior than NM1 indicated by smaller B-factors. Due to water-ion environment in NM2 the protein fluctuations are limited in contrast to NM1 that was set up *in vacuo*, where the protein is thus allowed to move more freely. Apart from the absolute fluctuations, we notice a similar pattern of dynamics for both NM1 and NM2 resembling experimental B-factors as well but to a minor extent.

### 1.4.2 Data Generation for PA

In proteins, $\alpha$-helices and $\beta$-sheets represent stable secondary structure elements. Since short polypeptides comprising alanines only were shown to form stable $\alpha$-helices [Marqusee *et al.*, 1989], these PAs were subject to experimental and theoretical studies concerning protein stability and folding, particularly focused on helix-coil transitions [Daggett & Levitt, 1992; Soman *et al.*, 1991; Vila *et al.*, 2000; Weber *et al.*, 2000]. According to the theory formulated by Zimm & Bragg [1959], random, unbonded structures are the dominant conformations of short polypeptides. Only for larger chains helical structures dominate the conformation space. Similar observations were made by Levy *et al.* [2001] who observed dramatically different energy landscapes of PA in vacuum and water for MD simulations. In contrast to organic solutions, water destabilizes $\alpha$-helices by interacting with polar groups of the peptide, which may even lead to an unwinding of the helices. Complementary to the highly stable and denaturation-resistent BPTI (see section 1.4.1), we use a small, flexible polypeptide comprising 12 alanine residues to derive interaction potentials of alanine contacts for different conformations that represent energy minima of the PA potential energy surface.

#### Setup for MD Simulations

To sample wide partitions of the PA conformation space, we performed MD simulations starting from two complementary structures (similarly to Levy *et al.* [2001]): an ideal $\alpha$-helix and an ideal $\beta$-sheet. To this end, we extracted the corresponding structure elements from PDB files with PDB codes 3A5A (helix, residues 105–116) [Kuwada *et al.*, 2010] and 2WSJ (sheet, residues 67–78) [Rodríguez *et al.*, 2010] and mutated all amino acids to alanine using the `psfgen` plugin for VMD [Humphrey *et al.*, 1996].

Starting from an $\alpha$-helical and a $\beta$-sheet structure, we set up a cubic simulation box, and filled it with water and ions (see Tab. 1.10), respectively. We applied a two-step minimization subsequently using the minimizers `steep` and `l-bfgs` to reduce the maximum force which acts on a single atom below 10 kJ mol$^{-1}$nm$^{-1}$. Afterwards, two equilibration runs were performed to adjust the biomolecular system to temperature (500 ps) and pressure (100 ps) separately. Both resulting structures served as input for two independent MD simulation runs each. PA was sim-

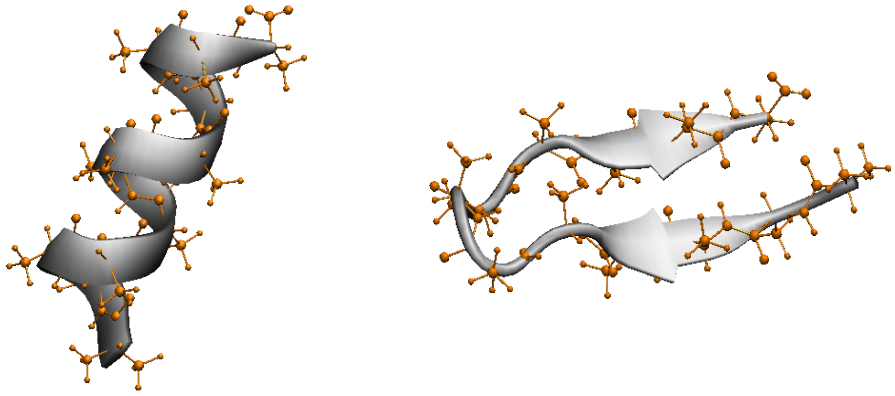


**Figure 1.25.:** Two complementary structures of polyalanine are shown. The ideal $\alpha$-helix (left) was generated on basis of PDB file 3A5A [Kuwada *et al.*, 2010], and the ideal $\beta$-sheet (right) was based on the structural elements of PDB file 2WSJ [Rodríguez *et al.*, 2010]. Graphics were generated using VMD [Humphrey *et al.*, 1996].

| simulation details | PA $\alpha$-helix | PA $\beta$-sheet |
|---|---|---|
| box type | cubic | cubic |
| box vectors [nm] | 6.022 | 6.062 |
| no. solvents | 7,196 | 7,416 |
| ions | 5 $Na^+$, 5 $Cl^-$ | 5 $Na^+$, 5 $Cl^-$ |

**Table 1.10.:** Initial setup of MD simulations of polyalanine. With respect to the different enlargement of $\alpha$-helical and $\beta$-sheet structure, box vectors [nm] and number of water molecules differ in both systems. The box vectors span the cubic simulation box.

ulated for 4 ns at a temperature of 400 K that was chosen to ensure a better sampling of the configuration space of PA, since energy barriers between distinct minima can be overcome more easily for higher temperatures. Thereby, we used an integration step of 2 fs, snapshots were written every 1000 steps. Coulomb interactions were modeled using the particle mesh ewald (PME) method and a cutoff distance of 1.5 nm, van der Waals interactions were switched off between 1.0 nm and 1.4 nm. The resulting trajectories were clustered using the `g_cluster` routine, that is implemented in GROMACS, with the linkage method. Clusters were defined based on pairwise RMSDs of the structures. Hence, a least-squares fit, that eliminates translational and rotational degrees of freedom, was applied prior to the assignment of a structure to one of the clusters.

We yielded 78 different clusters by performing the least-squares fit on the protein backbone and applying an arbitrary cutoff of 0.145 nm to discriminate between structures of neighboring clusters. The central structure of each cluster, i.e. the structure with the smallest RMSD to all other cluster members, served as input for the final production runs. For a better comparison of the final trajectories and resulting PA configurations, we placed all 78 structures in cubic simulation boxes with equal sizes and numbers of water molecules. Since we require a minimum distance of 1.2 nm between protein and box, we examined different settings (see Tab. 1.11). We defined a simulation box spanned by a box vector of length 6.2 nm that was filled with

| box vectors distances [nm] | 1.2 | 1.5 | 1.8 | 2.0 |
|---|---|---|---|---|
| min | 4.101 | 4.701 | 5.301 | 5.701 |
| max | 5.812 | 6.412 | 7.012 | 7.412 |
| avg | 4.928 | 5.528 | 6.128 | 6.528 |
| no. solvents distances [nm] | 1.2 | 1.5 | 1.8 | 2.0 |
| min | 2,238 | 3,399 | 4,909 | 5,784 |
| max | 6,559 | 8,732 | 11,471 | 13,553 |
| avg | 3,995.1 | 5,579.7 | 7,657.1 | 9,297.5 |

**Table 1.11.:** Results for different protein-box distances. We observed various dimensions of PA structures that we would like to place in a box with equal size and number of water molecules. Since we require the distance between the protein and the boundaries of the simulation box to be at least 1.2 nm, we monitored those values for varying protein-box distances.

7,861 water molecules and ions. This specific box size was chosen because each PA structure can be placed in the box with a distance of at least 1.2 nm from the boundaries of the simulation box. The number of solvent molecules was the respective minimum number determined for all simulation boxes. For each structure, a short minimization run was performed followed by two equilibration runs (50 ps). Finally, all PA configurations were simulated independently for 10 ns at a temperature of 300 K.

All MD simulation and analysis steps were performed with the GROMACS software suite [van der Spoel *et al.*, 2005] by employing the implemented OPLS-AA force field [Jorgensen *et al.*, 1996; Kaminski *et al.*, 2001].

In addition, we performed an extensive MD simulation using NAMD [Phillips *et al.*, 2005] with the all-atom additive CHARMM force field [MacKerell *et al.*, 1998, 2004] for about 62.5 ns on the HHLR (Hessischer Hochleistungsrechner). As start configuration we used one of the start structures of the 78 independent MD runs.

Since a clustering of approximately 1.4 million structures is not feasible, we restricted the sampling of minimum configurations by extracting every 1,000th frame from the trajectories. For the resulting 1,403 structures, a short minimization was performed, but did not converge for about 12% of the structures (170 of 1,403). Thereafter, we derived a set $\mathcal{S}_{\text{best}}$ of 20 PA structures as representatives. To ensure high diversity of the chosen structures, they are selected by a geometric approach based on the pairwise RMSDs. Initially, the structures $S_1$ and $S_2$ exhibiting the maximum RMSD are selected: $\mathcal{S}_{\text{best}} = \{S_1, S_2\}$. For any structure $S_i \notin \mathcal{S}_{\text{best}}, i = 1, \cdots, n$, which has not been chosen thus far, we computed the RMSD with respect to any other structure $S_j \notin \mathcal{S}_{\text{best}}, i \neq j$. We extend set $\mathcal{S}_{\text{best}}$ by structure $S_{i*}$ with the maximum geometric mean RMSD:

$$S_{i*} = \arg\max_{S_i} \frac{1}{n-1} \prod_{i \neq j}^{n} \text{RMSD}(S_i, S_j) \tag{1.52}$$

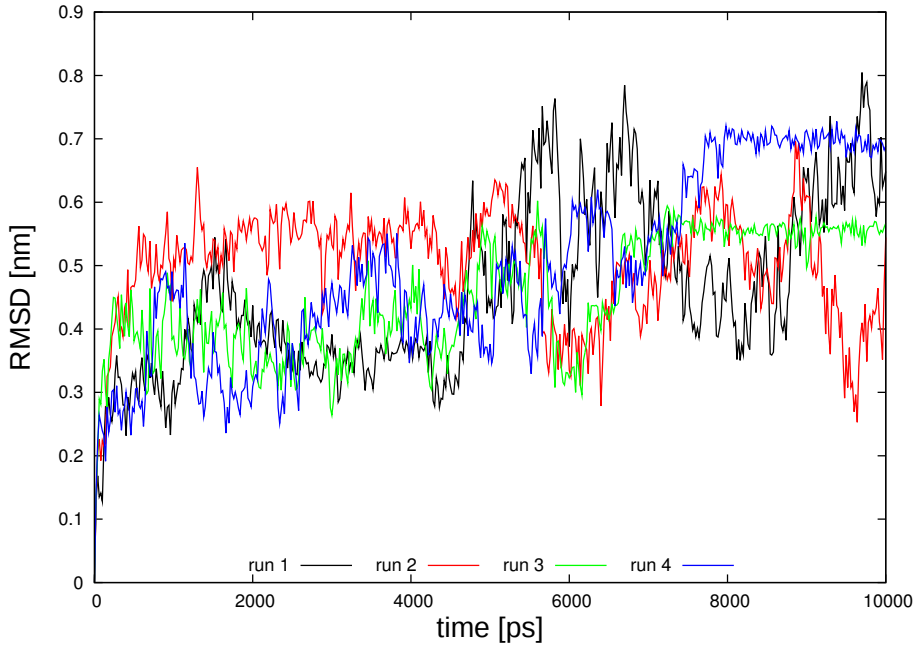



**Figure 1.26.:** For four independent MD simulations, the RMSD over time is computed for the $C_\alpha$ atoms after a least-squares fit with respect to the initial PA structure(s).

where $n$ is the number of configurations that have not been selected as representatives. All remaining structures are assigned to the "closest" (in terms of RMSD) representative structure of set $\mathcal{S}_{\text{best}}$.

### Results

For further analysis, we try to partition the configuration space of PA and find a set of energy minimum structures. At first, we computed the RMSD over time for each simulation with respect to the start configuration. In Fig. 1.26 we show the resulting data for four exemplary trajectories. We observe drastic structural changes indicated by an abrupt increase or decrease of the RMSD value. For two of the depicted trajectories (run 3, run 4), an RMSD plateau is noticeable, i.e. only minor RMSD fluctuations occur, indicating that the respective frames fluctuate around a consensus structure. In addition, we computed pairwise RMSDs for all combinations of minimized configurations as well as the RMSD of each structure before and after minimization. The resulting histograms are shown in Fig. 1.27. We notice a structural difference caused by the minimization algorithm of less than 0.1 nm, whereas on average the RMSD of any two minimized configurations is about 0.4 nm. Hence, minimized structures show a higher diversity than we can explain by simply moving the structures closer to a minimum configuration.

Since we want to extract amino acid specific interaction potentials that describe harmonic fluctuations around a central, energy minimum structure, we defined 20 clusters of PA structures that were identified as the minimized structures with maximum RMSD with respect to the other 1,383 extracted PA configurations. All non-minimized snapshots were assigned to the most "similar" cluster, i.e. the RMSD to the respective cluster structure is smallest amongst all clusters. In Fig. 1.28(a) we show histograms of pairwise RMSDs of cluster structures, regarding original

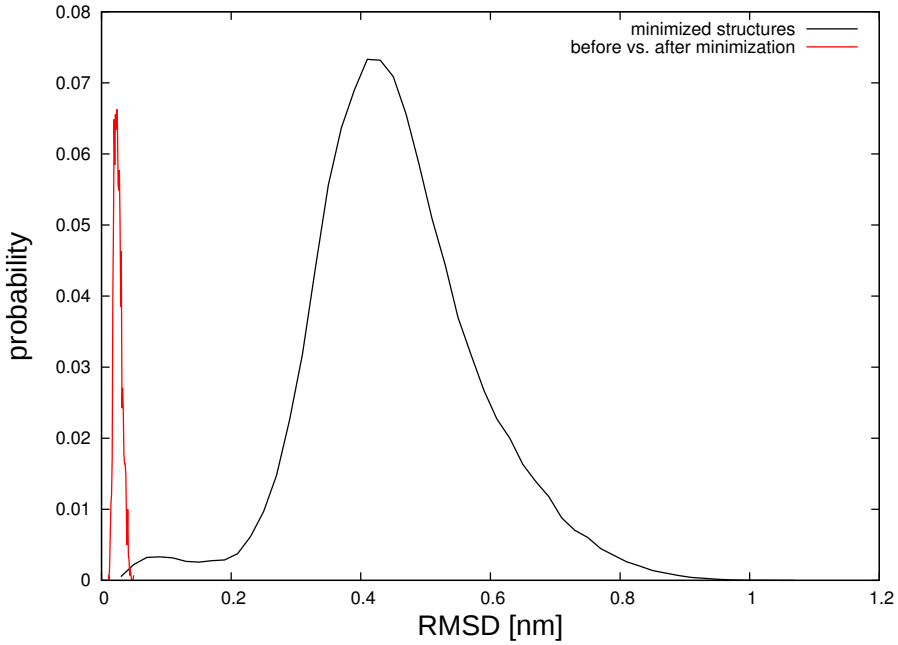
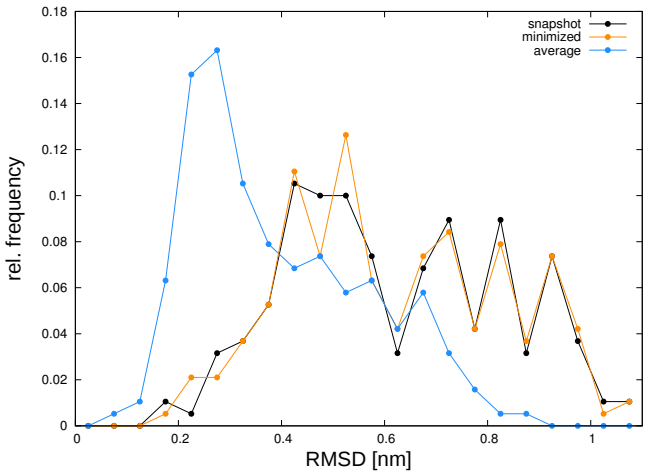
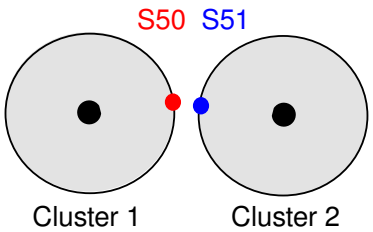



**Figure 1.27.:** Histograms of RMSD values for PA structures are shown. The black curve describes pairwise RMSDs of any two structures of the first representative configuration set which comprises 1,403 PA structures. The RMSD histogram plotted in red is derived by comparing each structure before and after minimization.




**(a)** Histogram of pairwise RMSDs of the 20 cluster structures that were extracted as snapshots from simulation trajectories, obtained after minimization or as average structure of the respective cluster.




**(b)** Illustration of a problematic case of assigning configurations to a cluster which can be used as explanation for smaller pairwise RMSD values for average structures compared to original and minimized snapshots.

**Figure 1.28.:** Results for PA clustering according to the proposed sampling based on the geometric mean to derive structures that differ the most from all other configurations on average.

| ref | no. | ref | no. | ref | no. | ref | no. |
|---|---|---|---|---|---|---|---|
| 31 | 110,146 | 197 | 104,895 | 415 | 118,925 | 658 | 299,017 |
| 55 | 93,385 | 204 | 77,043 | 521 | 27,570 | 734 | 56,382 |
| 78 | 57,431 | 323 | 15,793 | 524 | 98,721 | 780 | 43,392 |
| 124 | 13,825 | 356 | 36,806 | 564 | 41,777 | 818 | 51,732 |
| 180 | 4,510 | 413 | 8,398 | 609 | 27,396 | 1275 | 78,663 |

**Table 1.12.:** For each minimized reference structure (ref) that is used as central structure of a PA cluster the number of assigned snapshots (no.) based on minimal RMSD values is given.

central and minimized structures as well as the computed average structure for all assigned snapshots. In general, the pairwise RMSD of minimized structures and configurations directly extracted from an MD simulation is similar. Notably, the pairwise RMSD computed for average structures indicates an even higher resemblance which may originate from the fixed number of cluster (see Fig. 1.28(b)). Since we restricted the number of clusters, it occurs that we assign similar structures (in the example denoted as S50 and S51) to different clusters according to their RMSDs with respect to the central cluster configurations. Hence, the average structures of their assigned clusters include a contribution of S50 and S51 respectively and, thus, tend to be more similar than the corresponding minimized structures that were used to divide the configuration space. The number of snapshots that were assigned to each central structure is shown in Tab. 1.12. Interestingly, some PA configurations are populated with a much higher probability than others as the number of structures per cluster ranges from about 4,500 up to 300,000 indicating a varying size of the basins of attraction.

### 1.4.3 Estimation of Interaction Potentials for BPTI and PA

In the preceding parts we provided evidence for the correctness of fitting interaction potentials by formulating an SDP (see section 1.3.3). Here, we apply the optimization problem to "real" data originating from MD simulations and/or NMA as described for BPTI (section 1.4.1) and PA (section 1.4.2). To this end, we gathered snapshots from MD trajectories for both BPTI and PA; NMA was performed for BPTI only. The Hessians required for SDP fitting are obtained from MD covariance matrices via SVD, similarly to Eq. 1.6. All BPTI configurations were assigned to a single central structure sparing a clustering of the structures prior to the application of SDP as the RMSD for the MD trajectories suggests the resilence of just one basin of attraction (see section 1.4.1, Fig. 1.18). In contrast, we divided the conformation space of the highly flexible PA into 20 structure clusters (see Tab. 1.12) according to their pairwise RMSD values. As reference structure for MD snapshots, we used the average structure of each trajectory since we assume that single snapshots represent fluctuations around a central structure. Note that all PA structures that have been assigned to the same cluster were merged into so-called cluster trajectories. In addition to MD data, we determined amino acid specific interaction potentials of NMA Hessian matrices which have been computed for BPTI *in vacuo* as well as with solvent and ions. For these matrices, the energy minimized structure was used as reference. In Tab. 1.13 we provide an overview of BPTI data SDP has been applied to.

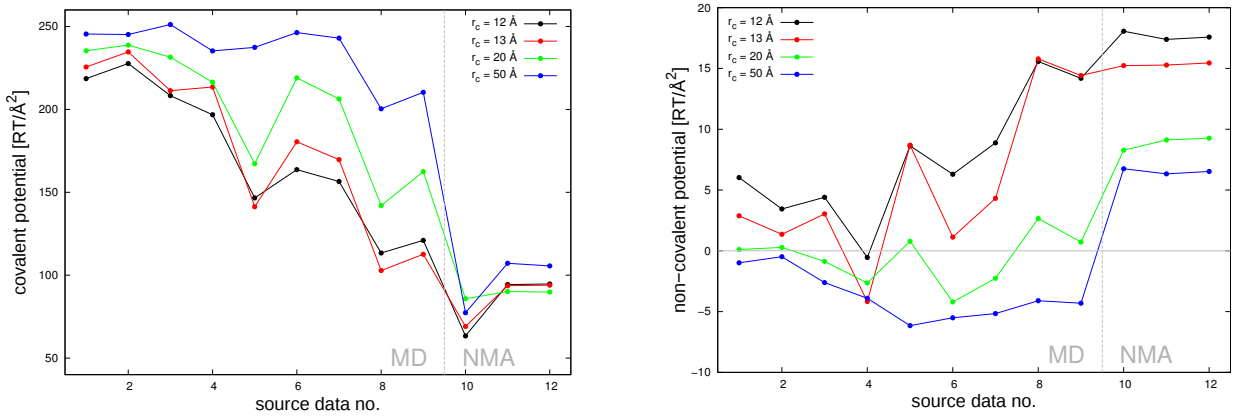
| no. | source | details | no. | source | details |
|---|---|---|---|---|---|
| 1 | MD | MD1 4 | 7 | MD | MD1 |
| 2 | MD | MD1 7 | 8 | MD | MD2 |
| 3 | MD | MD1 8 | 9 | MD | MD1+MD2 |
| 4 | MD | MD2 6 | 10 | NMA | solvent/ions (H) |
| 5 | MD | MD2 7 | 11 | NMA | *in vacuo* (H) |
| 6 | MD | MD2 8 | 12 | NMA | *in vacuo* (I) |

**Table 1.13.:** Overview of BPTI data that served as input for SDP. We considered covariance matrices (or rather the corresponding Hessians) of single (no. 1-6) and combined (no. 7-9) trajectories. The combined trajectories were created by concatenating all simulations based on MD1, MD2 or both. Note that we omitted the first 400 ps of each MD run. NMA Hessians have been computed for the protein *in vacuo* as well as in solution. The all-atom Hessian was reduced to a $C_\alpha$-Hessian matrix by applying the described reduction formula in the Hessian space (H) or by extracting the respective matrix entries of its covariance matrix (I).

The interaction potentials that describe the strength of covalent and non-covalent amino acid interactions were determined with SDP based on an ANM generated for distinct reference structures employing varying cutoff distances $r_c$ for the range of interaction: 12, 13, 20 and 50 Å for BPTI and 9, 10, 11, 12, 13 and 20 Å for PA. Thus, we can estimate the influence of $r_c$ on the SDP-fitted parameters. The cutoff definitions were chosen to account for differing protein sizes. In addition, we determine amino acid specific interaction parameters of BPTI data for reduced amino acid alphabets that were proposed by Pape *et al.* [2010].

### Results

At first, we will discuss results obtained for BPTI data with the simplest amino acid alphabet which contains a single amino acid type only and, thus, just one parameter for non-covalent contacts. Thus, we distinguish covalent and non-covalent force constants of harmonic springs connecting residues within the protein. The results for varying cutoff distances are shown in Fig. 1.29. For the two types of interactions, we notice a general trend: for higher cutoffs the strength of peptide bonds is overestimated in comparison to previously published results of $K = 82$ RT/Å$^2$ [Hamacher & McCammon, 2006] whereas the non-covalent bonds are assigned smaller values. Especially for a cutoff distance $r_c = 50$ Å, the interaction potentials of non-covalent bonds derived for MD data are all negative which is not a reasonable physical value. This behavior may originate from a compensatory effect. Interestingly, the results are in contradiction with the sensitivity analysis we performed where cutoff distances larger than the initial cutoffs yielded more reliable results. Both for single MD1 based simulations (no. 1-3) and for NMA (no. 10-12), the peptide bond potentials are similar for all employed cutoff distances, whereas for MD2 based (no. 4-6) and combined (no. 7-9) trajectories the yielded covalent potentials deviate to a higher extent. In analogy, we already discussed the higher resemblance of MD1 based covariance matrices in comparison to MD2 based matrices in a section 1.4.1 adding more evidence to the consistency of the applied parameter fitting method. Furthermore, we note that the non-covalent interaction potentials we determined for single MD1 simulations resemble the knowledge-based MJ potentials [Miyazawa & Jernigan, 1996] on average ($\approx 3.2$ RT/Å$^2$).




**(a)** covalent **(b)** non-covalent

**Figure 1.29.:** Interaction potentials of covalent and non-covalent amino acid interactions that were determined by applying SDP to BPTI data (see Tab. 1.13) for varying cutoffs. We present the results for the simplest amino acid alphabet containing a single amino acid type only.

Those contact potentials feature a rather stiff peptide bond which is characterized by a force constant of about 250 RT/Å$^2$, in contrast, Hamacher & McCammon [2006] suggested 82 RT/Å$^2$ as parametrization of covalent bonds. Similar values are obtained for combined trajectories ($\approx$ 90 RT/Å$^2$) as well as for NMA Hessians ($\approx$ 70 RT/Å$^2$) with a cutoff distance of 12 or 13 Å in combination with increased strength of non-covalent interactions ($\approx 15 - 20$ RT/Å$^2$). The observed values of peptide bond potentials were also obtained for larger amino acid alphabets, although the non-covalent parameters show only minor correlation (data not shown).

Comparing the interaction parameters we derived for NMA Hessians, we notice a higher resemblance than for individual MD results, particularly for non-covalent interactions. The peptide bonds are parametrized with lower values for the protein solved in water and ions (no. 10) than for the *in vacuo* setup (no. 11-12) as can be seen in Fig. 1.29. Protein backbone flexibility is increased if we include solvent effects. Different reduction schemes that were applied to yield a $C_\alpha$- from an all-atom Hessian matrix exhibit a negligible influence on the interaction potentials. This is also the case for larger amino acid alphabets as well (data not shown). If we increase the number of amino acid symbols, we detect a larger portion of interaction values smaller than zero (even more for increased cutoffs) indicating compensatory contributions due to the matrix structure. To this end, a more homogeneous parametrization is to be preferred to capture the dynamical parameters of a protein based on ANM.

The results for PA based fittings are about an order of magnitude smaller than for BPTI due to an increased flexibility of the PA polypeptide. The cutoff distance $r_c$ that is used to derive the interaction potentials has a similar effect on the interaction strengths as we observed for BPTI. We compare the results for the 20 PA clusters as defined in Tab. 1.12 by correlating the obtained values of covalent and non-covalent interaction weights with the end-to-end distance of the $C_\alpha$ atoms of the cluster average which was used as reference structure for SDP (see Fig. 1.30). Notably, for structures with higher end-to-end distances, i.e. stretched molecules, the harmonic springs are parametrized with smaller force constants. Hence, stretched PAs feature less non-covalent interactions which leads to an increased flexibility. In contrast, compact molecules with a larger number of interactions among residues fluctuate less around a central structure.

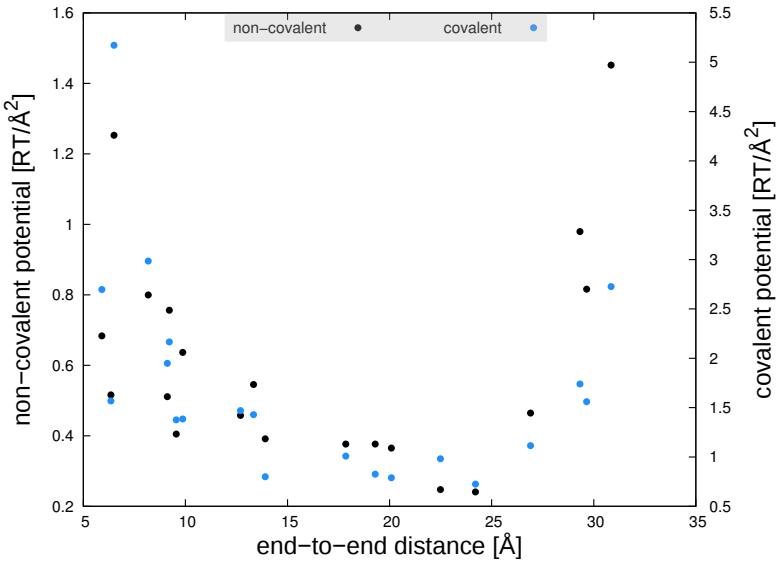



**Figure 1.30.:** Plot of SDP-fitted potentials for parametrizing covalent and non-covalent amino acid interactions of with respect to the end-to-end distance for each PA cluster.

Surprisingly, we find contradictory signals for four clusters with an end-to-end distance larger than 25 Å (see Fig. 1.30) exhibiting a more rigid behavior for extended molecules. According to Tab. 1.12, these clusters are the smallest with less than 20,000 structures each.

## 1.5 Discussion

In this chapter, we examined potential extensions of ANM approaches to enhance the capability of predicting temperature factors. Indeed, introducing an alternative contact definition (Eq. 1.10) lead to a higher correlation with experimental B-factors in comparison with pfENM applying sequence specific interaction potentials. Nevertheless, the correlation achieved by an inhomogeneous parametrization was even higher (see section 1.2.1). However, B-factor prediction without a mechanical model as proposed by Shih *et al.* [2007] could not be improved by adding sequence specific scaling constants (see section 1.2.2). In an additional study, we investigated how to derive interaction potentials from MD or NMA data using ANMs. Remember, those potentials are used to weight interaction of residues within a certain cutoff distance (see section 1.1 for details on ANMs). To this end, we performed a proof-of-principle analysis for MC-STUN (section 1.3.1), MLE/MAPE (section 1.3.2) and SDP (section 1.3.3) using an artificially constructed ANM with known input parameters. Obviously, all methods were able to (at least approximately) retrieve the potentials that have been used for construction. Due to shortcomings (discussed in sections 1.3.1 and 1.3.2), MC-STUN and MLE approaches were not pursued further. SDP has proven itself as an efficient method for this purpose and was applied to data sampled for the proteins BPTI (see section 1.4.1) and PA (see section 1.4.2). As worked out in this context, a thorough data sampling is required to obtain reliable fitting results. The interaction potentials derived for BPTI and PA showed large deviations in the absolute scale both for covalent and non-covalent interactions (considering only one type of amino acids). Notably, the results obtained for a BPTI trajectory that was created as concatenation of smaller trajectories and NMA exhibited a high resemblance under the assumption that all non-covalent interactions

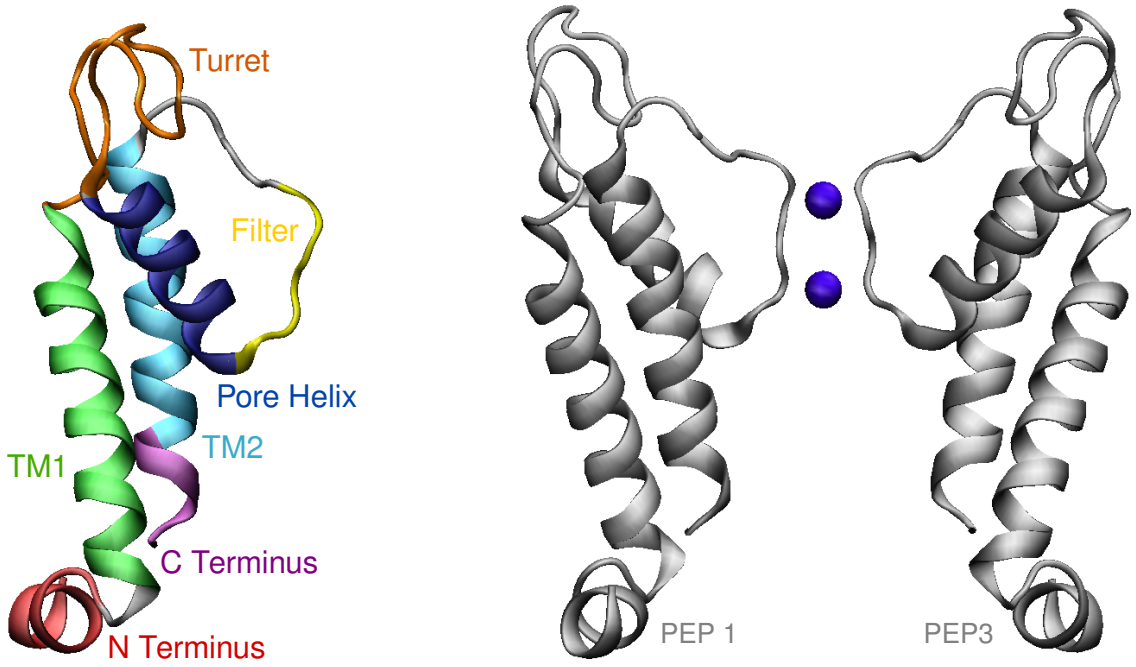
are weighted by the same force constant. Additionally, the obtained peptide bond potential is similar to the potential of 82 RT/Å$^2$ proposed by Hamacher & McCammon [2006] and the absolute value of non-covalent potentials resembles the average of MJ parameters [Miyazawa & Jernigan, 1996]. Correlations of non-covalent interaction potentials fitted for different input data are considerably reduced when considering a more detailed amino acid alphabet. For the highly flexible polypeptide PA, we obtained drastically reduced force constants from cluster structures compared to BPTI that were additionally correlated to the form of the molecules, i.e. compact molecules revealed higher interaction potentials. Summarizing, we have shown SDP to be a highly efficient method to derive interaction parameters of thoroughly prepared MD or NMA data yielding physically meaningful results. Furthermore, it is to be preferred to distinguish covalent and non-covalent interactions only.

# 2 Mechanics of Ion Channels

Ion channels are transmembrane proteins that allow an ion flux across the cell membrane. The resulting bioelectrical signals are, amongst others, the driving force of locomotion, sensory signals and cognition [Hille, 2001]. Ion channels are assemblies of several subunits forming homo or hetero oligomers. Larger channels are classified according to either their respective selectivity to conduct particular ions, such as $K^+$, $Na^+$, and $Ca^{2+}$ channels, or the ligand that induces channel function, e.g glycine for glycine receptors. Usually, such ion channels conduct at least one type of physiologically relevant ions ($K^+$, $Na^+$, $Ca^{2+}$, $Cl^-$, $H^+$) [Fischer & Sansom, 2002].

All known potassium channels are related to a single protein family, and can be found in bacterial, archeal and eukaryotic cells. The diversity of $K^+$ channels is mainly related to the various gating types. Gating or the opening of the water-filled pore is induced either by binding a ligand, such as ions, small organic molecules or even proteins, or by voltage changes within the electrical field of the membrane [MacKinnon, 2003]. Although potassium channels are involved in many physiological reactions, their general structure is conserved throughout prokaryotes and eukaryotes [Thiel *et al.*, 2011]. Usually, they consist of four subunits that are arranged symmetrically to form the water-filled channel pore that is responsible for the ion flux. Each subunit of the assembled tetramer contains between two and eight transmembrane domains (TM). Common to all $K^+$ channels is the so-called pore module, which consists of two TMs connected by a pore helix. The highly conserved signature sequence TxxTxGF/YG [Heginbotham *et al.*, 1994] mediates channel selectivity at the narrow part of the pore, referred to as selectivity filter [Doyle *et al.*, 1998] (see Fig. 2.1). Potassium channels are highly selective for $K^+$ ions, and allow flux of larger alkali metal cations $Rb^+$ and $Cs^+$, as well, but to a smaller extent. Usually, more than one ion can be found in the filter. The intrinsic affinity of a $K^+$ ion for the binding site is overcome by help of repulsatory effects of nearby ions resulting in high conduction rates [MacKinnon, 2003]. Smaller ions, such as $Na^+$ and $Li^+$, are excluded [Doyle *et al.*, 1998; Pagliuca *et al.*, 2007] and may block channel currents similar to toxine peptides [Rodríguez de la Vega & Possani, 2004].

Ion channel activity has also been reported for viruses [Pinto *et al.*, 1992]. Typically, such viral channel proteins are comprised of 50 to 120 amino acids [Wang *et al.*, 2011]. When sequencing the genome of the large, double-stranded *Paramecium bursaria* chlorella virus (PBCV-1) that replicates in unicellular, eukaryotic chlorella-like green algae [van Etten, 2003], a motif resembling the signature sequence of the pore module known for potassium channels was detected [Plugge *et al.*, 2000]. Despite its small size of 94 amino acids per subunit, the miniature, viral potassium channel Kcv contains all structural elements that are shared by pore modules of other $K^+$ channels. Mehmel *et al.* [2003] provided evidence that the miniature potassium channel is required for viral replication of PBCV-1. Host cell depolarization during infection has been suggested as a result of channel activity [Frohns *et al.*, 2006]. Previous studies have shown, that Kcv, which is comprised of four identical subunits, forms a functional channel in heterologous expression systems [Plugge *et al.*, 2000]. Selectivity for $K^+$ ions and sensitivity to channel block-




**(a)** Subunit of Kcv **(b)** Assembly of Kcv subunits

**Figure 2.1.:** Structure of Kcv [Tayefeh *et al.*, 2009]. In (a) structurally and functionally important regions are marked and labeled for a single subunit. The N-terminus is placed between cell membrane and the cytoplasma, the turret region represents an extracellular loop. The Kcv, which resembles the pore module of more complex potassium channels is anchored within the membrane by the outer transmembrane (TM) helix TM1. In (b) two facing subunits of the homotetrameric channels are shown. Blue beads indicate potassium ions that have entered the selectivity filter.

ers ($Na^+$, $Cs^+$, $Ba^{2+}$) resemble those of structurally more complex $K^+$ channels [Pagliuca *et al.*, 2007] making Kcv an ideal model system to study structure-related channel functions. Note that functional properties of Kcv depend on the expression system, which is caused by small changes in its fold due to the different membrane environment [Thiel *et al.*, 2011]. As a consequence, drugs targeting ion channels often have "secondary effects", mediated by modified physical properties of the bilayer membrane [Lundbæk, 2008].

Based on the crystal structure of KirBac1.1 [Kuo *et al.*, 2003], a homology model of Kcv was derived [Tayefeh *et al.*, 2009] by aligning channel parts according to their function. Extensive MD simulations have been performed showing the model's capability of continous ion transport. The structure of Kcv (see Fig. 2.1) features two TM helices, but in contrast to more complex potassium channels the inner TM is too short for the helix-bundle crossing, which has been proposed as common gating motif for the inward rectifying potassium (Kir) channel superfamily [Rapedius *et al.*, 2007]. Furthermore, the C-terminus is virtually at the end of the inner TM, hence, the channel is almost completely embedded in the membrane. Placed at the interface between membrane and cytoplasma, the short N-terminus was shown to be essential for channel function [Moroni *et al.*, 2002]. Hertel *et al.* [2010] provided evidence for a minimal length of the N-terminus by studying N-terminally truncated Kcv mutants. Truncation of up to seven amino acids did not affect the Kcv conductance capabilities, whereas deletion of more than

seven residues rendered the channel inactive, presumably due to missing salt bridges that have shown to be essential for stability and function of the channel. Recent studies [Tan *et al.*, 2010] concerning subunit composition revealed a concerted, "all-or-none" mechanism of the selectivity filter which can be correlated to gating regulation. Hence, introducing mutations in the signature sequence that inactivate a single subunit only lead to a complete loss of channel function. Furthermore, the studies revealed that inter-subunit cooperation is not necessarily required for permeability, as effects of mutated L70 are additive for the tetramer. Therefore, each subunit contributes equally and independently to the modulation of permeability.

Much research has been done in the field of potassium channels and Kcv, in particular. However, there are still unanswered questions concerning structure/function correlates. The miniature, viral potassium channel Kcv which resembles the pore module of more complex potassium channels is used a model system for the studies presented in this thesis.

## 2.1 Analysis of Functional and Stabilizing Modes

Numerous studies employed ENMs to investigate the mechanics of biomolecules, e.g. the HIV-1 protease was thoroughly investigated by Hamacher [2008]. The simplest model to describe thermal fluctuations of a protein is the GNM [Bahar *et al.*, 1997]. We already discussed in section 1.1, that the directionality of fluctuations is not assessed in GNMs other than for the ANM [Atilgan *et al.*, 2001]. In a previous study, Bahar *et al.* [1998] related slow and fast motions to function and stability. Slow motions are often associated with global dynamics of the tertiary structure that involve groups or subunits of biomolecules, since biological function requires concerted, collective motions. Residues that are active in high-frequency modes may also be involved in protein function, e.g. electron transport, but those amino acids are rather assumed to play an important role in maintaining the structure. So-called hot residues are identified by means of GNMs [Demirel *et al.*, 1998]. Previously, Shrivastava & Bahar [2006] described a common mechanism of pore opening for potassium channels that have been crystallized for prokaryotes and eukaryotes. Here, we employ GNMs to analyze the mechanics of the viral potassium channel Kcv.

### 2.1.1 Methods

We used the modeled Kcv structure [Tayefeh *et al.*, 2009] and performed a GNM based analysis. Residue specificity as well as different contact types were not invoked. Dynamical characteristics of a GNM are fully described by the Kirchhoff matrix $\Gamma$ as defined in Eq. 1.2. Two residues $i$ and $j$ are in contact, if their spatial distance $R_{ij}$ in the ground state structure is within a cutoff distance $r_c$. Hence, diagonal entries $\Gamma_{ii}$ are the number of amino acids interacting with residue $i$. We defined two cutoff distances $r_c = 7$ Å and $r_c = 13$ Å, that include contacts within the first and second interaction shell, respectively. Using SVD, we derive eigenvalues and the corresponding eigenvectors in analogy to ANMs (see section 1.1). The protein represented as Kirchhoff matrix has one symmetry. Therefore, we obtain one eigenvalue equal to zero which is omitted from further considerations. The mechanics of Kcv or rather its subunits was assessed by analyzing the motions that are encoded in the computed modes. Structural and functional parts of the

| code | residues | | channel region |
|---|---|---|---|
| N | 1 | 10 | N-terminus |
| TM1 | 13 | 30 | outer TM domain |
| T | 31 | 50 | turret loop |
| P | 51 | 62 | pore helix |
| F | 63 | 68 | selectivity filter |
| TM2 | 75 | 89 | inner TM domain |
| C | 90 | 94 | C-terminus |

**Table 2.1.:** Functionally and structurally important regions of Kcv monomers have been derived from literature [Tayefeh *et al.*, 2009].

| No. | Name | Characteristics |
|---|---|---|
| 1 | Kcv | functional (also referred to as Kcv-HOM-K29$_{\text{deprot}}$) |
| 2 | Kcv-HOM-K29$_{\text{prot}}$ | non-functional |
| 3 | Kcv-NMR-K29$_{\text{deprot}}$ | non-functional |
| 4 | Kcv-NMR-K29$_{\text{prot}}$ | non-functional |
| 5 | KirBac | crystal structure of KirBac1.1 [Kuo *et al.*, 2003] |

**Table 2.2.:** List of structures that have been investigated in this study. Model 1–4 are obtained by homology modeling [Tayefeh *et al.*, 2009]. For the N-terminus two different approaches were pursued: homology modeling (HOM) and modeling based on NMR data (NMR). For residue K29 both protonation states (protonated, deprotonated) were considered.

channel were defined according to Tayefeh *et al.* [2009] (see Tab. 2.1). We used a shorter filter definition TxGFG in the following.

Tayefeh *et al.* [2009] derived a total of four Kcv structures by a template-based modeling approach that included all available experimental data. As template the crystal structure of KirBac1.1 [Kuo *et al.*, 2003] was used. Three of four structural models were labeled inactive channels since only a single structure exhibited features of a functional potassium channel. Note that we used the latter model as reference structure for Kcv. We included both the non-functional variants as well as the template structure into our analysis to compare the dynamics of those channels with the Kcv reference model (see Tab. 2.2).

### 2.1.2 Results

To determine the effect of the distance cutoff $r_c$ for both the slowest and the fastest mode, that are described by the eigenvector belonging to the smallest and largest eigenvalue, respectively, we performed the analysis for $r_c = 7$ Å and $r_c = 13$ Å. Both distances have already been discussed for ENMs [Atilgan *et al.*, 2001; Bahar *et al.*, 1997] to include contacts of the first and second interaction shell, respectively. Slow modes are characterized by collective motions and, thus, involve groups of amino acids. Considering interactions of nearest neighbors only, i.e. a cutoff distance of $r_c = 7$ Å, captures an incomplete picture of global dynamics (see Fig. 2.2).

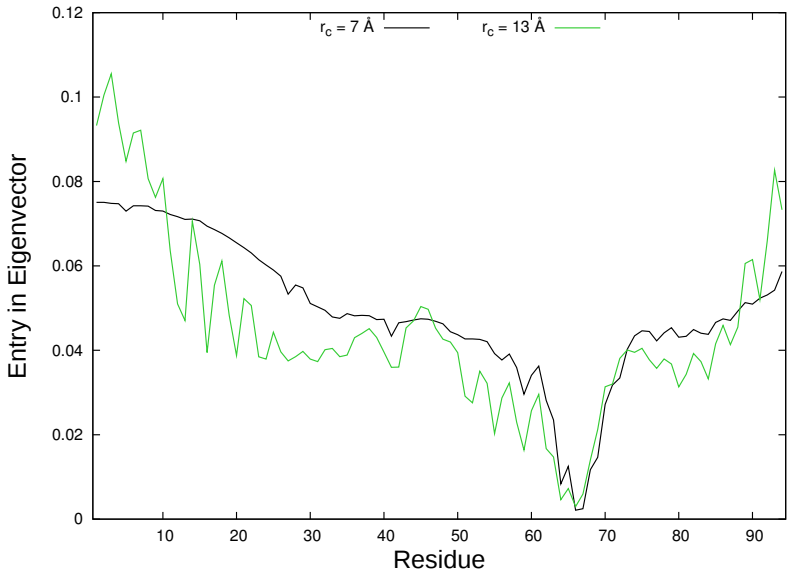
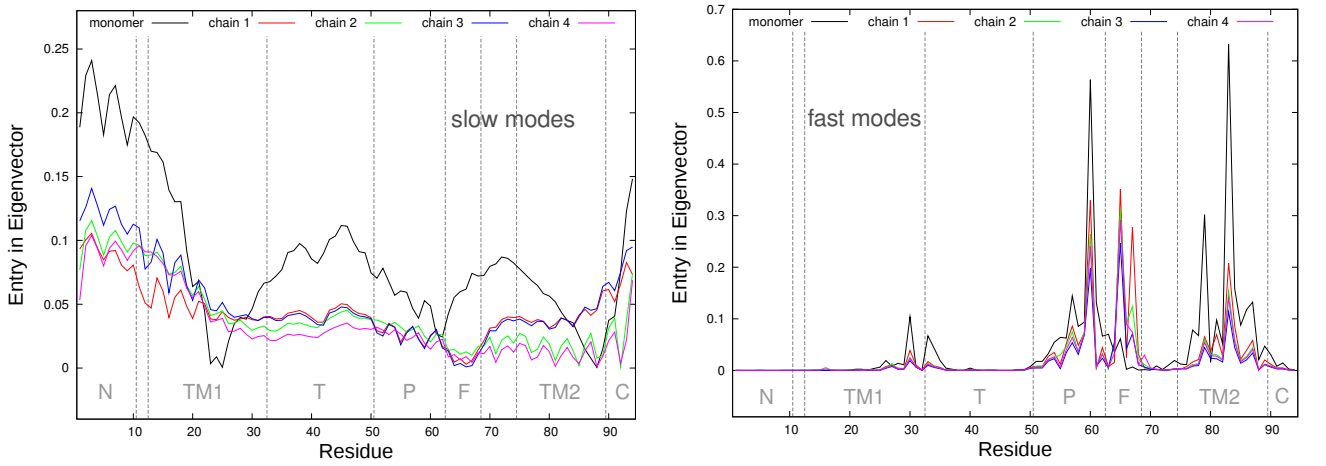



**Figure 2.2.:** Absolute entries of the eigenvector belonging to the smallest eigenvalue, which describes the slowest motions, are plotted for each residue. Results for both cutoff defintions $r_c = 7$ Å and $r_c = 13$ Å are compared for a single chain of the homotetrameric Kcv structure.

Larger distances reveal a more detailed behavior of flexible and rigid subunits as well as of the involved residues. For the fastest modes, that describe the dynamics of individual or small groups of residues, long-range interactions may obscure local dynamics by embedding amino acids in more rigid environment (data not shown). Thus, we propose that smaller cutoff distances should be preferred to describe fast fluctuations of residues labeled as kinetic "hot spots" [Demirel *et al.*, 1998] that are crucial for protein structure and integrity. To identify subunits that are involved in collective motions, which are responsible for protein function, larger cutoffs are recommended.

The selectivity filter, which contains the signature sequence of $K^+$ channels (TxGFGD), mediates the transport of potassium ions through the channel due to electrostatic interactions between carbonyl oxygens of filter residues and the ion(s) [Doyle *et al.*, 1998]. Mode analysis revealed for the filter region localized, high-frequency motions only (see Fig. 2.3). Mutual reactions between filter residues and potassium ions may cause such peak fluctuations. Analysis of the slowest mode allows us to identify the group of amino acids that are part of the signature sequence as the most rigid part of the Kcv. In contrast to more flexible regions, the selectivity filter is not incorporated in collective reorderings, presumably due to gating. In a previous study, Shrivastava & Bahar [2006] detected a similar behavior for eukaryotic and prokaryotic potassium channels as well. From the fastest modes, we detected residues crucial for structural integrity. Those amino acids are part of TM2, pore helix and the junction of TM1 and turret region (see Fig. 2.3). Interestingly, data for the monomer resemble the results for tetrameric chains. Hence, local dynamics is independent of tetramer assembly of the chains. On the contrary, for collective motions we observe dramatic differences between the monomer and the tetramer subunits. For the monomer, we find the termini as well as the turret loop to move concertedly. But in contrast to the tetramer, TM1 is rather rigid, whereas the filter shows more flexibility. Hence, an effect of tetramerization is the burial of the selectivity filter at the narrowest, inner part of the channel to provide rigidity.




**(a)** Slow modes **(b)** Fast modes

**Figure 2.3.:** Absolute entries of the eigenvector belonging to the smallest (a) and the largest (b) eigenvalue are shown for each residue. The smallest eigenvalue represents the slowest mode, whereas a large eigenvalue describes high-frequency motions. The data for the Kcv monomer (black) are compared to the four chains of the homotetramer. Important regions were labeled according to Tab. 2.1.

Remarkably, we observe a dimer-of-dimers behavior for collective motions: TM2 of opposite subunits reveal a similar pattern of dynamics that differs from that of neighboring subunits. Contact of opposite subunits is mediated by the filter region. Analogous findings were not observed for the investigation with the smaller cutoff $r_c = 7$ Å. This suggests that inter-subunit contacts apart from the filter-filter interactions are responsible for such differences. We repeated the analysis for non-functional structures (see Tab. 2.2) as well, but we did not observe a dimer-of-dimers like dynamics for the inner TM helix for either structure. Hence, we propose that this specific characteristics is a feature of a functional, viral potassium channel. To exclude modeling artifacts, we searched for similar dynamical patterns in the KirBac1.1 structure [Kuo *et al.*, 2003] which was used as template for the Kcv structure. The dimer-of-dimers behavior of the KirBac postassium channel that was also proposed by Kuo *et al.* [2003] could not be confirmed by our results.

In this section, we discussed the influence of the cutoff choice on results of the investigation of local and global dynamics. Taking into account contacts of next-nearest neighbors is recommended to examine collective motions since those interactions provide additional boundary conditions for global dynamics. Additionally, we found the filter region to be completely decoupled from collective motions, that were found to be responsible for protein function. Instead, the filter residues show highly localized dynamics which can be referred to ion conductance as well as to structure maintenance. Functionally relevant motions are performed preferentially by N- and C-terminus, and may be related to gating. We propose a dimer-of-dimers behavior to be an important feature of Kcv to ensure channel conductance. A similar pattern of collective dynamics was not found for non-functional Kcv structures.

## 2.2 Tectonics of a $K^+$ channel

The miniature potassium channel Kcv exhibits many features of more complex channels [Thiel *et al.*, 2011], such as ion selectivity, gating, and sensitivity to channel blockers [Gazzarrini *et al.*, 2003; Syeda *et al.*, 2008]. In recent years, extensive investigations of structure/function correlates revealed that long-range interactions of residues are important for channel function. In particular, Gazzarrini *et al.* [2004] reported a communication between the outer TM domain and the selectivity filter which involves the pore helix, but the mechanism remained unsolved. Further studies concerning the short, cytosolic N-terminus revealed that mutations and/or deletions of this N-terminal helix affects the voltage-dependent gating and may even cause a loss of conductivity [Hertel *et al.*, 2010; Moroni *et al.*, 2002]. The structural Kcv model [Tayefeh *et al.*, 2009], which has been developed since no crystal structure was available for the small potassium channel as it is the case for the majority of membrane proteins, has proven itself adequate to predict structure/function relations [Gebhardt *et al.*, 2011; Hertel *et al.*, 2010]. We employ a biophysical approach based on the Kcv homology model to uncover long-range interactions within the miniature channel to eventually extract a map of relevant mechanical connections within different domains. A full understanding of such interactions within the pore module of $K^+$ channels will provide insights into the mechanism how conformational changes of domains are coupled to regulate channel function.

To this end, we have chosen reduced molecular models, in particular ANMs (see section 1.1) that were proposed by Bahar *et al.* [1997], to assess structural and functional modes of the Kcv channel. The rationale behind this approach is two-fold: a) the mathematical simplicity of such models allows investigation of a large number of thought-experiments and varying setups, thus an orthogonal approach to a detailed account (loss of accuracy is mediated by a gain in an overall picture on scenarios); and b) we want to solely focus on the ground state dynamics of the protein here, thus we want to avoid any confusion with other selective pressures such as folding properties or more involved effects.

### 2.2.1 Methods

To investigate the mechanics of the Kcv potassium channel, we derived a coarse-grained representation of the protein utilizing ANM theory [Bahar *et al.*, 1997]. As we already worked out in section 1.1, the underlying principle is to view a protein as a mechanical network of amino acid. The complexity of proteins is reduced to a graph, whose nodes are the residues of the folded protein, each represented by a bead located at the position of its respective $C_\alpha$ atom. The edges in the network represent physical interactions, which in turn are reduced to harmonic interactions for all residues being closer than a certain cutoff distance $r_c$. As discovered in previous studies [Atilgan *et al.*, 2001; Doruker *et al.*, 2002; Micheletti *et al.*, 2004; Yang *et al.*, 2008], a physically sound choice is $r_c = 13$ Å as distance cutoff. To weight the strength of amino acid interactions, we employed two differing parametrization schemes for the interaction potentials: For the homogeneous parametrization, we employed a universal force constant $\kappa$ that is used to describe the interaction of any contacting residue pair $(i, j)$ as put forward by Tirion [1996],

whereas for the inhomogeneous parametrization we distinguished between bonded ($K$) and non-bonded contacts ($\kappa_{ij}$).

The pseudoinverse of the Hessian (Eq. 1.5) constitutes the mechanical covariance matrix $C$ (Eq. 1.6) and resembles the correlated motions, split into $x$-, $y$-, $z$-directions, of any pair of residues in the thermodynamical ensemble. This matrix is computed via SVD (see section 1.1). Experimental B-factors can be reproduced from $C$, as well [Atilgan *et al.*, 2001; Doruker *et al.*, 2002]. Introducing mutations by setting contacts artificially to zero or by changing interaction potentials leads to altered covariance matrices $C^{\mathrm{mut}}$. To quantify the magnitude of change in the protein dynamics introduced by mutations, we used FN (see Eq. 1.26) between $C$ and a "mutated" covariance matrix $C^{\mathrm{mut}}$:

$$\mathrm{FN} = \sqrt{\sum_{ij} \left( C_{ij} - C_{ij}^{\mathrm{mut}} \right)^2} \tag{2.1}$$

Computing the FN solely for the whole matrix may disguise relevant changes that occur only in functional subregions. To cope with such obscuration effects, we computed the FN for several well defined parts of the protein and therefore its covariance matrix as well (see also section 4.4). We defined structurally and functionally interesting channel regions [Gazzarrini *et al.*, 2004, 2003; Kang *et al.*, 2004; Tayefeh *et al.*, 2009] (see Tab. 2.1). From the mechanical covariances one can additionally compute the correlation matrices by dividing each matrix entry by the square root of its corresponding diagonal elements. To investigate the normalized correlation of motions, we performed this step and again restricted the subsequent FN computations to the functional portions of the channel.

### Switch-Off Procedure

In this part of the study we mimicked the effect of mutations by artificially switching off contacts in an ANM for the Kcv structure giving rise to a mutated mechanical covariance matrix $C^{\mathrm{mut}}$. This is equivalent to setting the interaction strength locally to a vanishing value. Since we are working on a homotetramer, we simultaneously switched off corresponding contacts in all chains. We considered the following scenarios:

1. Complete Switch-Off: for each residue we reduced the interaction strength to all other residues dramatically (peptide bonds were excluded and left unchanged). As we discriminate between the homogeneous case, where all interactions are set to a force constant of 1 RT/Å$^2$, and the inhomogeneous case, where we distinguish covalent bonds (parametrized by 82 RT/Å$^2$) and non-covalent bonds (weighted by 3.166 RT/Å$^2$), we used different reductions here as well. In the homogeneous case, switched-off interaction strengths were set to 0.05 RT/Å$^2$ and in the inhomogeneous case to 1 RT/Å$^2$. We considered these smaller values to be almost vanishing ones. The rationale behind this is two-fold: first, an alanine scan reduces biochemically all local interactions, but does not destroy interactions completely, effectively leaving a basal interaction; second, a value of exactly zero for "switched-off" contact might lead to additional singularities in the Hes-

sian matrix, effectively increasing the size of the null space, which might prohibit direct comparison of the singular vectors for real dynamical features.

2. Single Switch-Off: each non-covalent contact within a chain was switched off separately, thus its force constant was set to zero in all chains of the homotetramer at a time. Note that the problem of additional singularities does not occur due to the high connectivity of each residue. Therefore, the maintenance of at least two contacts (one covalent, one non-covalent) for each residue is guaranteed which is required to avoid additional singularities. The protocol is implemented in the `R` package `BioPhysConnectoR` [Hoffgaard *et al.*, 2010] that has also been successfully applied to acetylcholinesterase (see section 4.4).

#### ΔN-Mutants

Removing corresponding residues in each chain at the N-terminus generates ΔN-mutants. We focused on the Δ7, Δ8 and Δ9 mutants, which lack the first 7, 8 and 9 N-terminal amino acids, respectively. We constructed an ANM model for each mutant and computed the difference matrix of the covariance matrices of Δ7 and Δ9, which was restricted to entries of residues present in both mutants. We then proceeded to find a correlate in the ANM mechanics of the mutants with the loss of conductivity as measured in experiments. Additionally, we gained insight into the changes of the movements themselves, rather than the correlation among them as in the previous subsection. To this end, we decided to analyze the eigenvalues and respective eigenvectors for the wild type and mutant systems as well as for different mutants in more detail. Note that the eigenvectors were restricted to entries which are shared by the ΔN-mutant pair under comparison. A comprehensive analysis of the $2 \times 3N$ eigenvectors constitutes a high-dimensional analysis problem, whose complexity needs to be reduced to make a comprehensive analysis possible. We propose the following protocol: To compare mutant or wild-type system $I$ with a mutant $II$, a distance matrix $A$ was constructed computing the overlap distance of each eigenvector $\vec{u}_I$ and each eigenvector $\vec{u}_{II}$ for all pairings $(i, j)$ of eigenvectors (leaving out those belonging to the six vanishing singular values that occur due to rotational and translational degrees of freedom). This constitutes the overlap distance matrix:

$$A_{ij} = 1 - \frac{\vec{u}_{I,i} \cdot \vec{u}_{II,j}}{|\vec{u}_{I,i}| \cdot |\vec{u}_{II,j}|} = 1 - \cos \alpha_{ij} \tag{2.2}$$

with $\alpha_{ij}$ being the angle spanned by the eigenvectors $i$ and $j$. By this definition, which has also been applied to investigate the mechanical impact of the $\pi$-$\pi$ stacking interaction of F30 and H83 [Gebhardt *et al.*, 2011], we provided for entries of $A$ close to zero to indicate a high similarity of the respective eigenvector pairing $(i, j)$. The matrix $A$ was then used to assign corresponding eigenvectors/eigenmovements between the respective proteins. Although there are several efficient optimization schemes available [Hamacher, 2006, 2007a,b], for practical purposes we restricted this step to a greedy approach of local optimization.

For these assignments, we computed histograms of the distribution of the similarity values contained in the $A$ matrices. These histograms are then the sought-for, reduced representation of

the similarity of the accessible space of eigenmovements of the various mutants in comparison to the functional wild-type structure or among each other.

As a potential falsification experiment, we repeated this for $\Delta 1$, $\Delta 2$ and $\Delta 3$ mutants. This allows the comparisons of the absolute changes in the various $\Delta$N-mutants, but also investigating the difference between two $\Delta$N-mutants, which differ by the same number of deleted residues. For example, we can test by this setup whether there is a fundamental distinction between the differences between the $\Delta 1$-$\Delta 2$ mutants in comparison to the differences between the $\Delta 8$-$\Delta 9$ mutants – therefore, accounting for any effect, which is caused by relative differences in the number of deleted residues. We can also check whether any significant change in the mechanics is due to the absolute number of deleted residues.

Moreover, we quantified the similarity between the histograms by computing the Kullback-Leibler divergences $D_{kl}$ [Kullback & Leibler, 1951] between each histogram pair $h_{\Delta\Delta i}$ and $h_{\Delta\Delta j}$. The $D_{kl}$ was already discussed in the realm of MD and chemoinformatics as a measure to quantify dynamical differences [Hamacher, 2007c]:

$$D_{kl}(h_{\Delta\Delta_i}||h_{\Delta\Delta_j}) = \sum_x h_{\Delta\Delta_i}(x) \log_2 \frac{h_{\Delta\Delta_i}(x)}{h_{\Delta\Delta_j}(x)} \tag{2.3}$$

In addition, we performed a clustering based on $D_{kl}$ as distance measure [Hamacher, 2007c]. This procedure was applied to both the Kcv homology model and the truncated KirBac1.1 model.

Computations were performed using the package `BioPhysConnectoR` [Hoffgaard *et al.*, 2010], an extension of the statistical software `R` [R Development Core Team, 2008] for biophysical and evolutionary biology purposes.

### 2.2.2 Results

We modeled the channel dynamics by ANMs (see section 1.1) for the structure of the miniature potassium channel Kcv. We used the following two different weighting schemes for the parametrization of the interaction potentials, which describe the harmonic spring connections between each of two residues that are in contact in the native Kcv structure:

- homogeneous parametrization, that does not discriminate between covalent and non-covalent bonds; this effectively focuses exclusively on structural aspects and
- inhomogeneous parametrization, ensuring that covalent bonds are more rigid in comparison to non-covalent contacts within the protein structure.

First, we found that both the homogeneous and the inhomogeneous parametrization yield the same qualitative results. This insensitivity towards fine-tuning of parameters prompted us to focus on the results for the homogeneous and thus purely structural case alone.

The full dynamical information, which can be derived using ANMs, is contained within the mechanical covariance matrix $C$ (see Eq. 1.6). It describes if and to what extent the fluctuations of the residues of a protein or a protein complex are related. The covariance matrix in addition also denotes the amplitudes of these motions. In fact, normalization of the covariance matrix results in the (mechanical) correlation matrix. Since we are interested here in both the correlation and the amplitude of the motions themselves, we used both matrix types in our study.

### Results : Complete Switch-off

Conceptually, to mimic the effect of alanine scanning mutagenesis in the Kcv channel [Gebhardt *et al.*, 2011], we developed a thought experiment: we artificially "switched-off" all non-covalent interactions of an amino acid to its interaction partners within the Kcv channel. This was achieved by reducing the respective strengths to a very small, but finite value. Since the functional channel is a homotetramer, the reduction on corresponding amino acids was performed in all four monomers simultaneously. To quantify the effect of such reduced mutual amino acid interactions on the biomechanics of specified protein regions, we computed a matrix norm, namely the FN (see Eq. 1.26), of the covariance and correlation matrices. A large norm indicates a major influence of the "mutated" amino acid on the dynamics of the whole protein or a particular region (see Tab. 2.1). The regions of interest correspond to structurally and/or functionally defined regions of the Kcv channel [Tayefeh *et al.*, 2009]. Worth noting is that the focus on such sub-regions also increases the accuracy as the subtle differences are not masked by global effects of the full protein. Similar to previous studies we defined two regions for the filter using TxxTxGFGD or TxGFGD as signature sequence for ion specificity.

In general, the results for covariance and correlation matrices are qualitatively the same differing only in the signal strength as can be seen in Fig. 2.4 that shows the influence of a residue mutation on the monomer dynamics. The data provide such a high resolution that even the periodicity of $\alpha$-helices is apparent from the oscillations in the signals (insets Fig. 2.5 and Fig. 2.6). These findings on both the equivalence of the two parametrizations, as well as the equivalence of the analysis foci, allowed us to restrict ourselves to only one analysis scenario: homogeneous parametrization and covariance matrices.

In Fig. 2.5 we present in more detail a subset of these data, namely the effect of mutagenesis of the entire channel on the N-terminus. In this example, each residue of the Kcv channel is mutated one at a time: all non-covalent interactions of the amino acid under consideration to other amino acids are modeled by weak harmonic springs with a negligible strength. The effect of each mutation on the dynamics of the N-terminus was quantified by the accompanied FN. Clearly, the largest effects on the N-terminus are generated by mutations in the N-terminus itself. The inset shows the same results, but restricted to mutations outside of the N-terminus. Notably, perturbations on the C-terminal residues convey information on the N-terminus. This suggests that the physicochemical and structural properties of the C-terminus reflect back on the mechanics/dynamics of the N-terminus. To a smaller extent we observe this influence to be reciprocal, i.e. a mechanical communication from the C- to the N-terminus. These results are in perfect agreement with experiments that have shown a functional connection between both termini [Hertel *et al.*, 2010; Tayefeh *et al.*, 2009]. Furthermore the data in Fig. 2.5 also show a

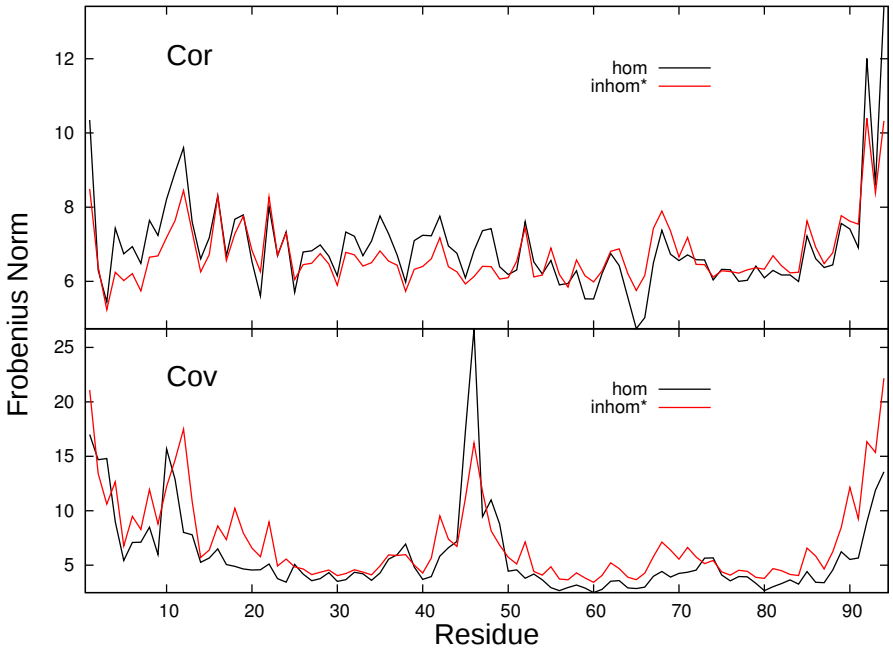



**Figure 2.4.:** Importance of amino acids derived in the complete switch-off scenario mimicking an alanine mutagenesis study: interactions of each residue to all its contacting residues are weakened in all chains individually. Changes in dynamics are quantified by computing the FN of the respective difference matrices for covariance (Cov) and correlation (Cor) matrices. Both interaction strength parametrizations (homogeneous (hom), inhomogeneous (inhom)) have been applied to resemble interaction potentials. The results are compared for each matrix type. The asterisk indicates scaled values.

large effect of mutations in the first part of TM1 on the N-terminal mechanics, whereas residues of the second part of the helix seemed to have no effect. This result corresponds well with the distribution of the thermal B-factors of TM1 obtained from the MD model of Kcv. It occurs that the high FN coincides with the dynamic parts of TM1.

Another key finding of our analysis is that the filter region seems to be kept in a Gimbal-like anchoring in the structure of the Kcv channel. The data show that substantial effects on the filter were generated only by mutated residues within the filter region itself (see Fig. 2.6). Otherwise the filter dynamics is inert against any mutation in the rest of the protein. Our results for the filter region were independent of the particulars of the definition which residues constitute the filter. The finding that the filter region is mechanically so much uncoupled from the rest of the protein is very interesting with respect to the function of a $K^+$ channel. The structural mechanism, which is underlying ion selectivity in the filter, does not tolerate a great deal of flexibility of this part of the protein. Our data show that the protein is indeed constructed in such a way that the filter is more or less mechanically uncoupled from the rest of the channel; this probably guarantees a maximum conservation of the pore in an otherwise dynamic protein. Negligible effects in the FN signal stem from helical residues of TM1 and TM2 as well as from the entrance of the filter region. Worth noting, however, is that mutating residues of the filter region lead to slight changes in the mechanics of TM2, but not vice versa.

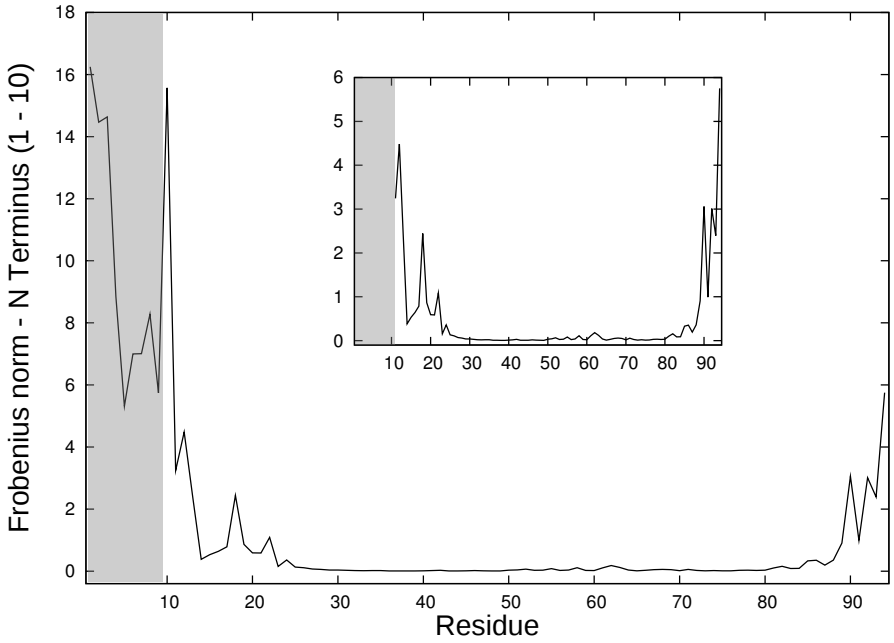



**Figure 2.5.:** The influence on the dynamics of the N-terminus quantified by the Frobenius norm of the respective parts of the covariance matrix. For each residue, whose interactions to all its contacting neighboring amino acids have been reduced – effectively quantifying the impact of those residues on the mechanics of the N-terminus. The results are shown for the homogeneous and thus purely structural case. The inset shows the influence of residues that are not part of the N-terminus itself (indicated as a gray bar).

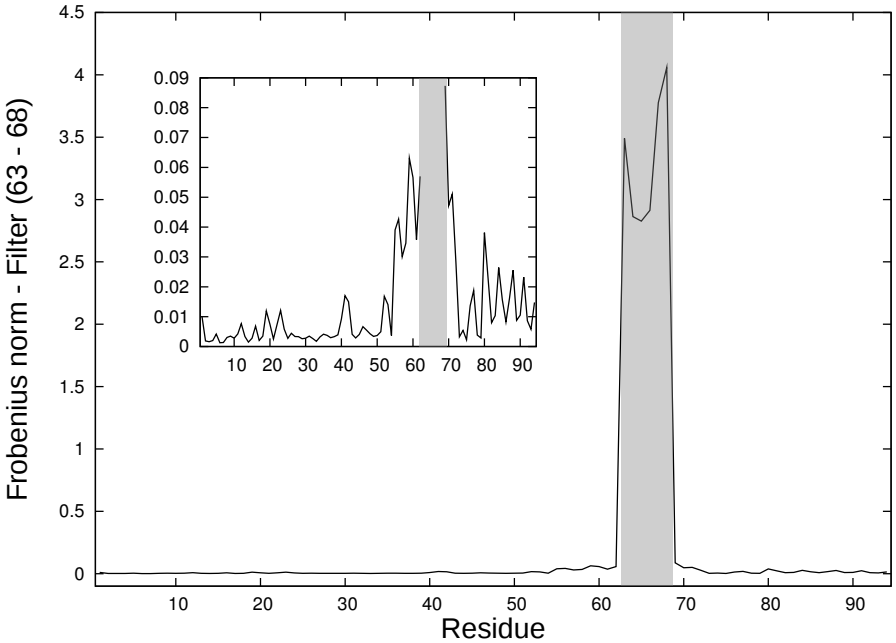



**Figure 2.6.:** Impact of residues on the sensitivity filter. The Frobenius norm for the covariance matrix of the selectivity filter alone sheds a light on how switching off residues affects the dynamics of the filter. The gray bar indicates residues belonging to the sensitivity filter. The inset presents the results for residues outside the filter region. For brevity, we show results for the shorter filter definition (TxGFGD) and homogeneous parametrization, only.

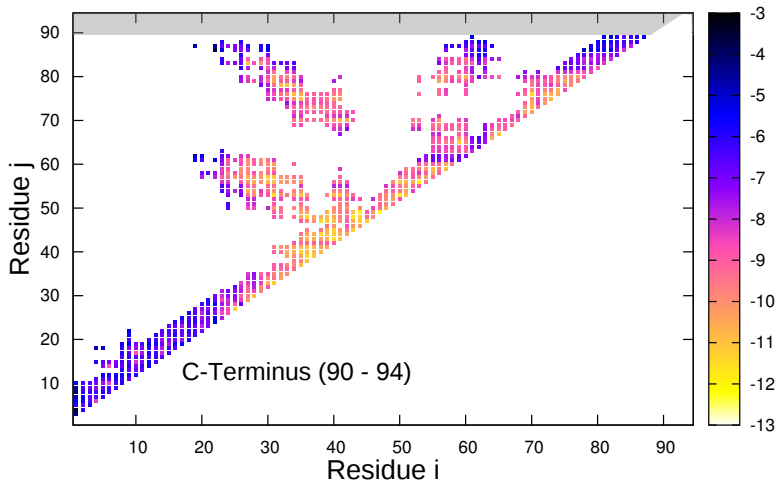



**Figure 2.7.:** Single switch-off results for the C-terminus. Each interaction of contacting residues (i,j) that are not classified as C-terminal residues (indicated by the gray bar) is switched off one at a time in all chains simultaneously. The Frobenius norm is computed for C-terminal parts of the covariance matrices. Blue colors indicate interactions having a high influence on the dynamics of the C-terminus, whereas yellow interactions are negligible.

## Results : Single Switch-off

The aforementioned full switch-off experiment mimics alanine scanning mutagenesis. To gain more detailed insight into the tectonics, e.g. on the functional connectivity, in the miniature potassium channel Kcv, we also applied another protocol to the structure. This protocol was already successfully applied by [Hamacher, 2008] to the HIV protease to identify important contacts (in contrast to whole residues) in the biomechanics of a protein. For this purpose single, corresponding non-covalent contacts were artificially deleted in all monomers simultaneously; all other interactions of the corresponding residues were left untouched. Connections between subunits were retained and the effect was quantified by the FN (see Eq. 1.26) of the covariance matrices $C$ and $C^{\mathrm{mut}}$.

As a case example, we illustrate in Fig. 2.7 the results for the C-terminus. For each existing contact, excluding those with at least one residue belonging to the C-terminal region, the FN is shown. We notice large effects of contacts within the N-terminus as well as between both transmembrane domains TM1 and TM2 on the C-terminus. The most significant residue contacts are between the pairs: F19-V87, A22-T86 and A22-V87 (see Tab. 2.3). Above all, deleting interactions of amino acid M1 leads to major changes in the mechanics of the C-terminus. Those findings underscore again the functional significance for amino acid interactions between N- and C-terminus. Tab. 2.3 shows the top 5 contacts ranked according to their influence on the respective regions measured on the basis of their FNs. The data show that a deletion of contact E12-L94 has drastic effects on the dynamics of most of the functional regions. It occurs that interactions of residues A22 and L94 have large effects on the mechanics of the pore helix. Furthermore contacts T9-A22, F19-L94 and E12-L94 affected the dynamics of TM2 upon deletion.

| N | TM1 | T | P | F | TM2 | C |
|---|---|---|---|---|---|---|
| E12-L94 | E12-L94 | D52-D68 | A22-L94 | F19-S62 | E12-L94 | M1-F4 |
| T11-L94 | L92-L94 | D68-K72 | T9-A22 | F19-H61 | A22-L94 | A22-T86 |
| L18-L94 | T9-T93 | F24-I51 | E12-L94 | M1-I90 | S5-I90 | M1-L8 |
| A22-L94 | T11-L94 | D68-P71 | L18-F24 | I69-I81 | T11-L94 | M1-T9 |
| L18-T20 | S5-L92 | P32-D52 | M1-I90 | M1-V91 | T9-A22 | A22-V87 |

**Table 2.3.:** Contacts with major influence on the dynamics of specified channel regions. Results for single switch-off. For each region (as defined in Tab. 2.1) the contacts with the highest influence (quantified by the Frobenius norm of respective covariance matrix portions) on the dynamics of the respective region are listed.

A key finding of the complete switch-off approach is that none of the amino acids was severely affecting the dynamics of the selectivity filter. To obtain a better insight into this phenomenon we investigated the influence of single contacts on the dynamics of this region. We found individual contacts between F19 in TM1 and the entrance of the filter region (H61, S62) to be distinctly relevant for the mechanics of the filter. The apparent prominent role of F19 and its functional connection to the filter domain favorably confirms and further details previous experimental results. Different natural mutants of Kcv exhibited an exchange of F19V in this position [Kang *et al.*, 2004]. A functional analysis exhibited that the F19V mutation has severe effects on the sensitivity of the channel to $Cs^+$ block, $Rb^+$ permeability and on gating [Gazzarrini *et al.*, 2004]. These altered functions, which are most likely caused by an effect of the filter, were causally related to a long-range interaction of the amino acid in position 19 via amino acids 54 and 66 [Gazzarrini *et al.*, 2004]. The present analysis draws a similar picture in that the amino acid in position 19 is mechanically linked to the filter albeit the more directly connected amino acids might be H61 and S62. Interactions of residue L94 were found to be crucial for the dynamics of several regions. This is in line with earlier findings about the role of L94 as a salt bridge partner for positive residues in the slide helix which turned out to be essential for ion transport [Hertel *et al.*, 2010; Tayefeh *et al.*, 2009].

### Results : ΔN-Mutants

Hertel *et al.* [2010] have shown that deletion of up to 7 N-terminal amino acids (L2 - L8, referred to as Δ7 mutant) maintains activity of Kcv. However, the channel becomes inactive if two more residues, T9 and R10, are removed as well; the intermediate Δ8 mutant (only T9 removed) shows only a reduced conductivity. The reason for the abrupt loss in channel function in response to the truncation of a single amino acid can be causally related to an interruption of the salt bridge pattern between the two TM domains [Hertel *et al.*, 2010]. In order to investigate whether the abrupt loss in function is also influenced by the dynamical modes of such ΔN-mutants, we developed an additional protocol for the truncated structures. The goal was to search for signals that go hand in hand with the complete loss of channel conductance when truncating 9 instead of 7 N-terminal residues, possibly only showing small signals for the Δ8 mutant. Since we did not incorporate any amino acid specificity in our model, we simply removed the first 7, 8 or 9 residues from the N-terminus, respectively. Although M1

is kept in the experimental setups, we deleted it, too. This is justified because M1 is known to be relevant for protein biosynthesis only. It does not influence the protein dynamics itself (see also section 2.2.2 for the N-terminus dynamics in the complete switch-off experiments).

Because the matrices of different mutants have different dimensions we restricted the analysis only to those entries, which correspond to the same residues in the different ΔN-mutants. To quantify which residues are most affected by shortening the N-terminus, we aggregated all squared differences belonging to each amino acid. The largest influence could be observed in the N- and C-terminal regions. This comes as no surprise, because both termini communicate mechanically with each other (see section 2.2.2). Interestingly, half of TM1 was affected as well, whereas the remaining part showed only small deviations from the wild type behavior (data not shown). Furthermore, we also identified motions of the wild-type channel with those in the ΔN-mutants. By this procedure, we obtain an aggregate picture of how the individual motions differ. We extended the set of mutants by Δ1, Δ2, and Δ3 mutants. This allows us to investigate whether substantial mechanical changes occur at an absolute number of truncated residues, or whether the effect is a relative one.

Bahar *et al.* [1998] introduced the notion of functional and stabilizing mechanical motions, which can be classified by their respective frequencies. As aforementioned (see section 2.1), modes with a low frequency represent global motions, which occur mainly due to functional movements e.g. conformational changes. High-frequency modes, on the other hand, are related to localized kinetics crucial for maintenance of structural stability. A pairwise comparison of the modes of both wild-type/mutant and mutant/mutant revealed high similarity mainly for stabilizing modes; the majority of the functional modes, however, differed substantially. To analyze potential mechanical aberrations between different mutants we defined the intra-Δ and the inter-Δ group. The first group contains all comparisons between mutants, whose truncation lengths differ at most by two residues per monomer, e.g. mutants Δ1-Δ2 or Δ7-Δ9. The inter-Δ group is comprised of mutant pairs in which the truncation lengths differ substantially such as Δ3-Δ8 or Δ1-Δ9.

Fig. 2.8 suggests two regimes for the mutant-mutant comparisons: the intra-Δ and inter-Δ group share common, distinct features among each other: mean, shape, location of the maximum in the histograms. The only deviation from this pattern is observed for the comparison of the Δ7-Δ9 mutants. In spite of belonging to the intra-Δ group, the Δ7-Δ9 overlap distances resemble the curvature of inter-Δ pairings. In other words, deleting two N-terminal residues from the Δ7-mutant introduced changes in mechanics as drastic as deleting eight amino acids from the Δ1-mutant. This observation relates very well to the experimental results, which have shown that removing nine instead of seven residues of the N-terminus, rendered the channel inactive.

To analyze the histograms in Fig. 2.8 further we computed a distance measure between the respective histograms, the Kullback-Leibler divergence $D_{kl}$, which is capable to discriminate between the intra-Δ and the inter-Δ group (Fig. 2.9). We found, that the truncation of the channel by a similar number of amino acids leads to similar changes in dynamics in general; this result is almost independent of the location of the truncation. The only exception from this common pattern is observed in comparison of the Δ7-Δ9 mutants. For this particular pair of mutants we see a remarkable difference: the changes induced by going from the Δ7-mutant to

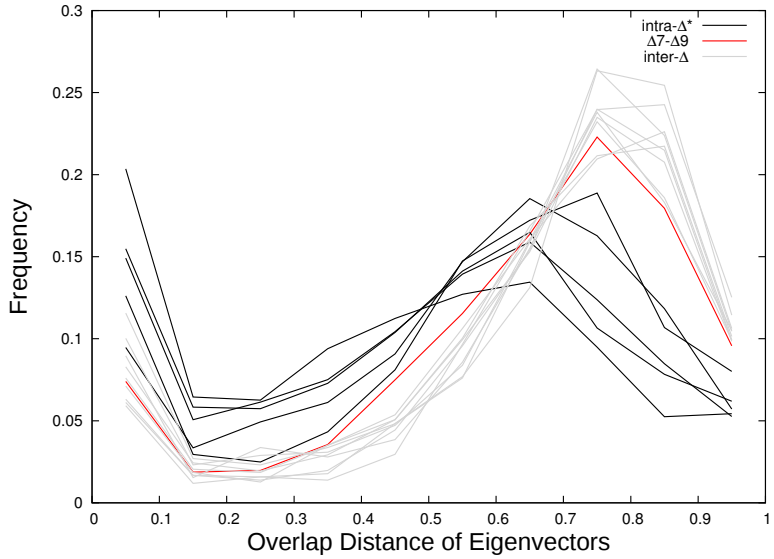



**Figure 2.8.:** Histograms of overlap distance values. For each Δ-Δ mutant pair the similarity is quantified as overlap distance of the eigenvectors/eigenmovements. Small overlap distances indicate a high similarity of the dynamics of the respective mutants. The intra-Δ group includes mutants whose monomer sizes differ by one to three residues only (except Δ7-Δ9) and the inter-Δ group comprises more pronounced mutant pairs, which differ in length by more than three residues. Results are independent of the binning scheme used.

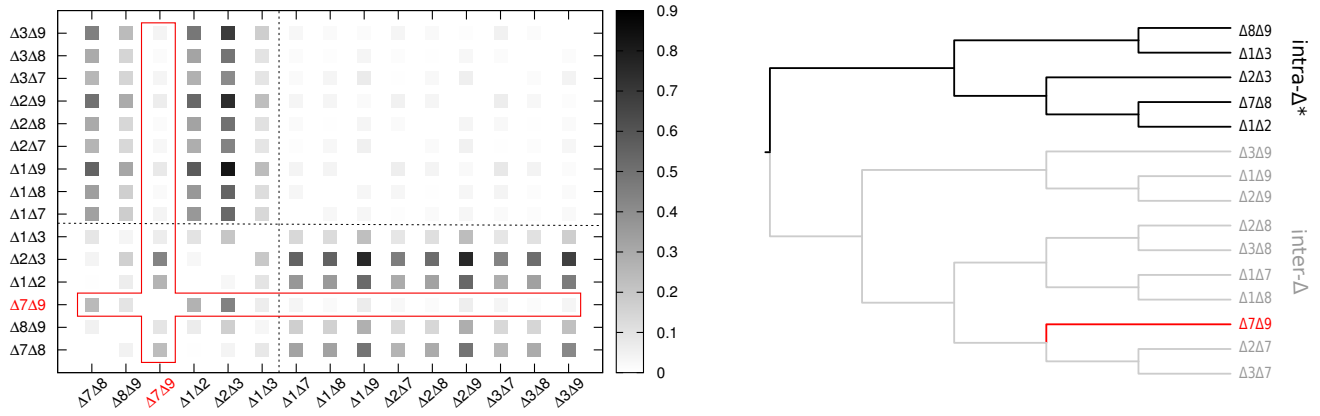
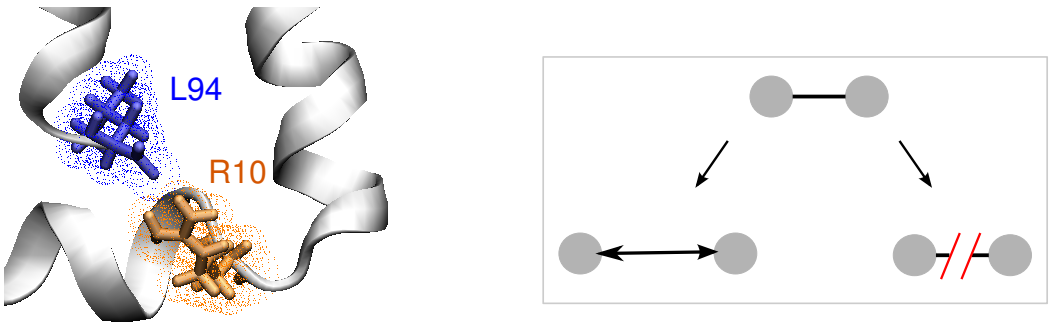



**Figure 2.9.:** Kullback-Leibler divergence of the overlap distance distributions. For each mutant pair, such as Δ7-Δ8 the overlap distance of the corresponding eigenvectors is computed as a measure for mechano-dynamical similarity. We compare those overlap distance distributions amongst all mutant-mutant pairs by the Kullback-Leibler divergence $D_{kl}$ of the respective histograms (left). Small $D_{kl}$ values (white to yellow) indicate a high similarity of the underlying distributions. Note how the Δ7-Δ9 mutant comparison is much closer to the group of mutants, which differ substantially, like Δ1-Δ7 or Δ3-Δ8, eventually showing signatures of the fundamental physiological changes found *in vivo*. Additionally, we applied a clustering to the $D_{kl}$ values; the resulting tree is shown at the right. Results are independent of the binning scheme for the frequencies used in $D_{kl}$.




**(a)** Salt bridge R10 - L94 **(b)** Scheme of sensitivity analysis

**Figure 2.10.:** The salt bridge close to the truncation site where a loss of channel function is observed is shown embedded in the structural context (left). To this particular interaction, we applied a sensitivity analysis to judge its influence. Therefore, we either moved the $C_\alpha$ atoms farther apart along their connection vector or we artificially deleted the respective contact by setting its interaction strength to zero. A sketch to illustrate the described sensitivity analysis is given at the right.

the Δ9-mutant is more severe than the modification of any other Δ$L$-mutant to its respective Δ(L+1) or Δ(L+2)-mutants; eventually the relative change is as severe as going for example from Δ1 to Δ8-mutant. This can also be seen in Fig. 2.9 which presents the clustering of the mutant pairs in accordance with their Kullback-Leibler distance. Again, we observe that only the Δ7-Δ9 pair from the intra-Δ group is assigned to the inter-Δ group.

We further addressed the question whether the assignment of Δ7-Δ9 to the inter-Δ group rather than to the intra-Δ group results from specific, important interactions within the Kcv structure. Hence, we tested the sensitivity of our approach towards local changes. We identified the salt bridge, which is constituted of residues R10 and L94 in Kcv as a crucial interaction that is close to the truncation site (see Fig. 2.10(a)). To test whether this precise mutual interaction between the two amino acids is relevant for the mechanical coupling in the channel we repeated our protocol, but for all mutants this specific interaction was relaxed by a) pushing the residues farther apart or by b) weakening/deleting its interaction (see Fig. 2.10(b)). Notably, with this procedure we obtain the same results as before. This implies that our protocol is insensitive towards local rearrangements and hence mutual interactions between amino acids; the results of these experiments underscore the exceptional character of the Δ7-Δ9 step, which is inherent in the global structure rather than in the local interactions. Altogether, the analysis revealed that truncating more than 7 to 8 residues has a devastating effect on the dynamics of the channel. This altered dynamics in Kcv is correlated with the experimentally observed loss of channel activity; like in the computational analysis we found a sharp transition between active and inactive channels for a truncation of more than 7 to 8 amino acids [Hertel *et al.*, 2010].

To investigate whether these findings are singular for the Kcv alone, we repeated the whole protocol for the KirBac channel, which was truncated to the size of Kcv [Tayefeh *et al.*, 2007]. Analogous results were achieved: again, also in this channel, which shares on the amino acid level in the respective region only moderate homology, the transition from a Δ7 to a Δ9 mutant causes the aforementioned drastic effect. Also in the case of the KirBac channel we detect for

Δ7-Δ8 and Δ8-Δ9 larger deviations to the intra-Δ group than before, which may be due to the absolute number of deleted residues.

Collectively, the results of these experiments imply that the detailed amino acid sequence is not so important for the dynamical behavior of the protein. Apparently, the global structure itself is the main determinant of the dynamics in the pore module of potassium channels.

The present data show that ANMs are capable to generate a global map for the mechanical connections in small $K^+$ channels. This map provides detailed information on long distance interactions in this small channel, which cannot be directly explained by mutual interactions of amino acids. A particular finding is that the selectivity filter is apparently isolated in a mechanical manner from the rest of the channel protein. This architectural arrangement ensures that the delicate geometry of the selectivity filter, which is essential for ion selectivity, is maintained in an otherwise dynamic protein. Some of the main mechanical interactions in the Kcv channel, which were identified from the network model, are in agreement with experimental studies, which have already in the past speculated on long-range interactions between the outer TM domain and the pore in this channel [Gazzarrini *et al.*, 2004]. The ANM also shows that a truncation of the N-terminus of Kcv beyond a critical length has a devastating effect on the channel structure and that this truncation is correlated with a loss of experimentally detectable channel function [Hertel *et al.*, 2010].

In the particular case of the N-terminus, the data demonstrate that, despite substantial local differences, important mechanical properties are governed predominantly by topology. In other words, two sequences such as Kcv and KirBac, which are only moderately similar with a sequence identity of about 23% [Tayefeh *et al.*, 2009], but which generate a $K^+$ channel pore module with an overall similar architecture, generate common functional modes. This is interesting on the background of the finding, that different $K^+$ channels have the same overall topology of the pore module but still exhibit different biophysical features such as different unitary channel conductances, differences in gating and sensitivity to blockers. Hence it is plausible that the functional modes, which are governed by the topology and which may be a genuine feature of all $K^+$ channels, determine generic mechanical interactions in a $K^+$ channel. Any precise and channel specific feature, which is characteristic for a certain type of $K^+$ channel, is then modulated on top of these genuine modes by specific amino acid interactions.

## 2.3 Investigation of a $\pi$-$\pi$ Stacking Interaction in Kcv

Relationships between structure and function have been elucidated for diverse regions of potassium channels and the viral Kcv in particular, e.g. for the N-terminus [Hertel *et al.*, 2010] and the selectivity filter [Chatelain *et al.*, 2009; Cordero-Morales *et al.*, 2006; Gazzarrini *et al.*, 2004; Heginbotham *et al.*, 1994; Tan *et al.*, 2010]. In this study, we investigate the structural significance of the two TM domains for the channel properties. Unbiased information can be obtained by an alanine scanning mutagenesis [Cunningham & Wells, 1989]. To this end, all residues are mutated to alanine, except alanines that are replaced by glycines, one at a time. Hence, all side chain interactions (except for the $C_\beta$ atoms) are eliminated and the contribution of single residues to fold and stability can be determined. Alanine scanning mutagenesis of both TMs

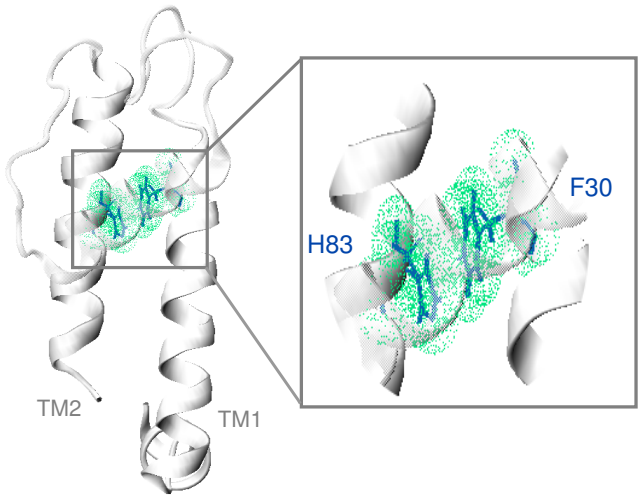



**Figure 2.11.:** $\pi$-stacking is observed between the transmembrane domains TM1 and TM2. A single subunit is shown, the corresponding interaction F30-H83 is highlighted.

revealed that a large number of amino acid positions of the Kcv tolerate the substitution by alanine [Gebhardt *et al.*, 2011]. In general, the inner transmembrane helix TM2 is more tolerant against alanine substitutions than TM1, a total loss of conductance was observed for H83, only. Whereas for TM1, drastic effects are noticed for H17, I20, Y28, and nearly all phenylalanines (F14, F24, F30, F31) leading to a near or complete loss of channel function. Most of those sensitive residues are conserved within the sequences of viral potassium channels (data not shown). Notably, the side chains of the corresponding residues point to the membrane in the structural model of Kcv. Together with the aromatic character of those amino acids these findings suggest a membrane anchoring is provided by these residues. Notably, substituting aromatic residues with other aromatic amino acids is able to rescue the channel function with varying degrees.

As pointed out, the alanine mutagenesis study revealed a complete loss of conductance when replacing residue H83 from TM2. This amino acid is interacting via $\pi$-stack with F30 from TM1 over a distance of about 3.5 Å (see Fig. 2.11); the helical stacking is apparent from the modeled Kcv structure [Tayefeh *et al.*, 2009]. Such $\pi$-$\pi$ stacking interactions occur if two closely located, aromatic molecules are positioned in parallel [McGaughey *et al.*, 1998]. Providing links between helices, they are important for correct fold and stability. $\pi$-$\pi$ interactions have also been discussed for proton gating of acid sensing ion channels [Li *et al.*, 2009]. In the light of the alanine mutagenesis results [Gebhardt *et al.*, 2011], where for F30 mutants appropriate channel function was only rescued by aromatic substitutions, the helical $\pi$-stacking forms important subunit contacts.

We apply an ANM (see section 1.1), a coarse-grained protein representation, to further support the structural importance of this respective interaction. The results of this section have been published in:

Gebhardt, M; Hoffgaard, F; Hamacher,K; Kast, SM; Moroni, A; Thiel, G (2011) Membrane anchoring and interaction between transmembrane domains is crucial for $K^+$ channel function. *J Biol Chem* 286:11299.

### 2.3.1 Methods

To further elucidate experimental findings of the applied alanine mutagenesis study, we examined the homology model of the Kcv [Tayefeh *et al.*, 2009] using ANMs (see section 1.1). In these coarse-grained protein models, residues are represented as beads placed at their $C_\alpha$ atom. Connections between interacting residues are modeled as harmonic springs. From the Hessian matrix (see Eq. 1.5), eigenmodes and eigenfrequencies are derived by SVD. To estimate the importance of the $\pi$-$\pi$ stacking interaction of residues F30 and H83, we artificially delete the respective contact. The impact of the mutation, is quantified by the overlap of corresponding eigenvectors belonging to eigenvalues greater than zero. This measure has been introduced in section 2.2 to compare different $\Delta$N-mutants. In general, let $\alpha_{ij}$ be the angle between any two vectors $\vec{u}_i$ and $\vec{v}_j$, the overlap distance $d_{ij}$ of those vectors is computed as defined in the context of the overlap distance matrix $A$ (see section 2.2, Eq. 2.2):

$$d_{ij} = 1 - \cos \alpha_{ij} \tag{2.4}$$

### 2.3.2 Results

For the application of the ANM to the Kcv model, we used a homogeneous parametrization, i.e. we did not distinguish different types of interactions (covalent vs. non-covalent). Thus, we performed a purely structural analysis focusing on topological issues. To examine the influence of the specific contact F30-H83, we artificially deleted this single interaction from the contact definition. Note that there is no experimental analog to this *in silico* analyis, thus, this analysis can be performed *in silico*, only. The local mechanics of a protein is encoded in the covariance matrix, that can be computed for an ANM (see Eq. 1.6). Wild-type and mutant Kcv are compared with each other by computing the differences of their respective covariance matrices. The impact of the introduced mutation is quantified by FN (see Eq. 1.26). We observed, that both TMs as well as the turret region were affected by deletion of this specific contact. Furthermore, we defined the overlap distance as a measure to compare modes of wild-type and mutant models. Modes, mathematically described as eigenvectors, describe specific fluctuations of the molecule at a specific frequency, given by their respective eigenvalues. Superposition of all but the first six modes, whose eigenvalues are equal to zero due to rotational and translational degrees of freedom, leads to the covariance matrix. As discussed in the context of GNMs (see section 2.1), low frequency modes are often referred to as global modes, that are responsible for collective motions of subunits. Those motions have impact especially on protein function. Fast modes are often localized fluctuations with impact on stability issues [Bahar *et al.*, 1998]. The overlap distance measures the deviation of two corresponding modes $i$ and $j$ of both systems under consideration, $d_{ij} \in [0, 1]$ (see Eq. 2.4). If the modes are identical, the mutation did not change the respective motion, the distance is zero. Deleting the $\pi$-stacking interaction F30-H83 has no effect on the functional modes (see Fig. 2.12). Global, collective motions responsible for gating, etc. are not altered. On the contrary, for stabilizing modes we observe dramatic effects. Hence, this particular amino acid interaction is important to maintain the correct fold and to provide stability.

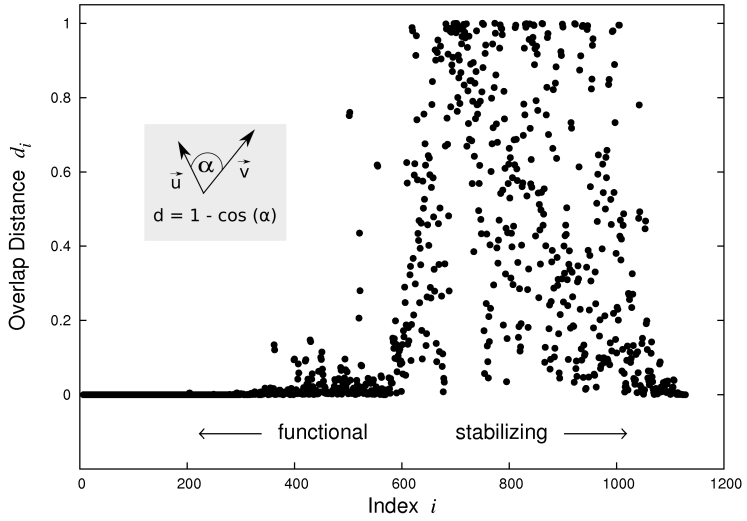



**Figure 2.12:** Overlap distance $d_i$ is plotted for all modes (eigenvector) indexed by $i$. The overlap distance is derived as angle between corresponding wild-type and mutant eigenvectors $\vec{u_i}$ and $\vec{v_i}$. Small indices represent global motions determining protein function, whereas high-frequency modes (high indices) describe local dynamics affecting protein stability. In the mutant channel, the contact F30-H83 is artificially deleted. Picture is adapted from Gebhardt *et al.* [2011].

Our data show that $\pi$-stacking interactions attaching the inner TM2 to the outer TM1, which leads to its immobilization, provide optimal channel function. Experimental data presented by Gebhardt *et al.* [2011] are further supported by results obtained from a theoretical, coarse-grained approach, that revealed modified protein stability for a disruption of the $\pi$-stacking contact F30-H83.

### 2.3.3 Contributions

The theoretical network based analysis (ANM, see section 1.1) of the $\pi$-$\pi$ stacking interaction of F30-H83 was performed in the framework of this PhD thesis.

## 2.4 Discussion

Due to its small size, Kcv has been adapted as model system for potassium channels since its structure corresponds to the conserved pore module shared by $K^+$ channels [Thiel *et al.*, 2011]. Furthermore, Kcv exhibits many features of a functional ion channel [Gazzarrini *et al.*, 2003]. Based on the crystal structure of KirBac1.1 [Kuo *et al.*, 2003], Tayefeh *et al.* [2009] derived a homology model of the Kcv structure. In the presented studies, we utilized this Kcv model to perform theoretical analyses of its structure. We employed network-based approaches (GNM, ANM) that assume a coarse-grained protein representation by placing beads at the $C_\alpha$ positions of amino acids and connecting beads within a given cutoff distance with harmonic springs. GNMs model fluctuations of residues/atoms isotropically, and, thus, provide insights into the general dynamics of a protein. Focusing on the slowest modes, we are able to identify functional regions that undergo concerted collective motion, whereas fast modes reveal hot spots of localized dynamics that are crucial for protein stability [Bahar *et al.*, 1998; Demirel *et al.*, 1998]. In contrast, in ANMs fluctuations are described anisotropically for $x$-, $y$- and $z$-direction. Additionally, the springs that connect contacting residues can be assigned different force constants to invoke amino acid specificity or to stiffen covalent bonds compared to non-bonded interactions.

We applied switch-off scenarios to the Kcv model, i.e. we artificially deleted single interactions or decreased the connection strength of a residue to any other amino acid. We estimated the impact of the "mutations" by the FN of the covariance matrices. Notably, we detected similar signals for both homogeneous and inhomogeneous parameterization mainly differing in the magnitude. From the *in silico* mutation experiments we identfied amino acid contacts that have a major impact on particular functional channel portions upon deletion. In particular, alterations within the C-terminus reflect back on the dynamics of the N-terminus. In contrast, the selectivity filter is rather mechanically uncoupled from the other channel parts; its dynamics is affected by deletion of long-range interactions that involve residue 19 and the entrance of the filter which is in accordance to experimental data [Gazzarrini *et al.*, 2003]. In addition to the identification of mechanically relevant sites, we observed a dimer-of-dimers behavior for Kcv since the inner TM2 of neighboring subunits exhibits differing dynamics which has not been detected for the KirBac1.1 structure and the inactive Kcv structures. Furthermore, we proposed the overlap distance as measure to evaluate the effect of mutations as disturbance of functional or stabilizing modes according to the notion introduced by Bahar *et al.* [1998]. We used the proposed measure to judge on the effect of an artificial deletion of a $\pi$-$\pi$ stacking interaction of residues F30 and H83 yielding drastically modified stabilizing modes whereas the functional modes that mediate collective motion were not affected. In a similar approach, we investigated the effect deletion of N-terminal residues has on the dynamics of the channel and quantified the altered dynamics of different $\Delta$N-mutants by overlap distance histograms. Hertel *et al.* [2010] have reported a complete loss of conductivity of Kcv for deletions comprising more than 8 N-terminal residues. Indeed, $\Delta$7- and $\Delta$9-mutant, which were experimentally shown to be functional and non-functional, respectively exhibit as drastically modified dynamics as the comparison of $\Delta$3- and $\Delta$8-mutants. Altogether, the presented data further support the structural model proposed by Tayefeh *et al.* [2009].

# 3 Coevolution in Hammerhead Ribozymes

Ribonucleic acid (RNA) is a versatile, ubiquitous biomolecule which plays a prominent role in diverse cellular processes. The negatively charged polynucleotide chain is composed of nucleotides consisting of ribose sugar and a phosphate group. The nucleotides are connected by phosphodiester bonds in 5'-3' direction of their sugar and distinguished according to their attached bases. In general, we are concerned with the purines adenine (A) and guanine (G) as well as with the pyrimidines cytosine (C) and uracil (U). RNAs can fold into diverse three-dimensional structures featuring both intramolecular interactions between secondary structure elements, such as kissing loops and pseudoknots [Batey *et al.*, 1999], and intermolecular interactions with ligands including metals [Chow & Bogdan, 1997] and other macromolecules as can be found in large assemblies like the ribosome. Predicting the three-dimensional structure of RNA that mediates its function is difficult. Due to their regulatory or catalytic role in a variety of biological processes such as protein biosynthesis, RNAs are classified in different RNA types, e.g. transfer-RNA, messenger-RNA, ribosomal-RNA, etc.. Other than coding messenger-RNAs that are central to protein synthesis [Crick, 1970], the highly diverse non-coding RNAs seem to be abundant in roles that require high recognition capabilities for specific nucleic acids [Eddy, 2001].

Among non-coding RNAs, we find RNAs with catalytic properties: so-called ribozymes catalyze reactions on themselves or other molecules [Wilson & Lilley, 2009] which was first observed by Cech *et al.* [1981] and Guerrier-Takada *et al.* [1983]. According to their size, ribozymes are classified as large and small ribozymes. The most prominent example of a large catalytic RNA is the ribosome [Cech, 2000], a large complex comprising rRNAs and several proteins that is central to protein biosynthesis and has been found in all domains of life. Further members of this class are group I and group II introns which catalyze their own splicing from primary RNA transcripts by different mechanisms [Saldanha *et al.*, 1993]. Prominent, representative, small ribozymes are the hairpin, hammerhead, and hepatitis delta virus (HDV) ribozymes comprising about 40 to 160 nucleotides [Doudna & Cech, 2002].

Hammerhead ribozymes are the smallest naturally occuring RNA endonucleases that catalyze the site-specific cleavage of their own phosphodiester backbone [Doudna & Cech, 2002; Przybilski & Hammann, 2006]. First discovered as catalytically active element in the rolling-circle replication of certain viroids [Forster & Symons, 1987; Prody *et al.*, 1986] the presence of hammerhead ribozyme motifs was detected amongst others in plant [Gräf *et al.*, 2005; Przybilski *et al.*, 2005], amphibian [Epstein & Gall, 1987] and mammalian genomes [de la Peña & García-Robles, 2010a; Jimenez *et al.*, 2011]. The first three-dimensional structure of a ribozyme was put forward by Pley *et al.* [1994]: a minimal hammerhead ribozyme that consisted of a catalytic core comprising 11 highly conserved nucleotides flanked by three helices arranged in a Y-shaped conformation (see Fig. 3.1). Khvorova *et al.* [2003] and de la Peña *et al.* [2003] made the observation that peripheral elements such as loops and bulges at the end of the flanking helices are required for full activity and may stabilize the ribozyme in a catalytically active form via tertiary

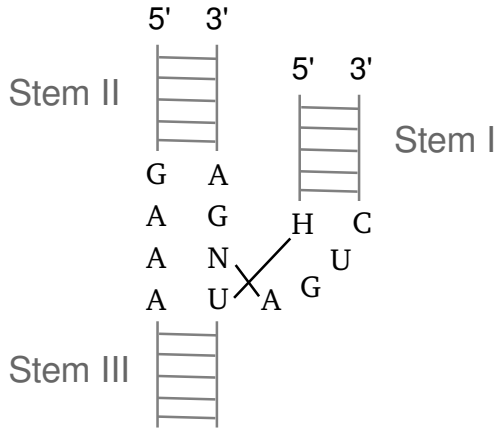



**Figure 3.1:** Secondary structure of a minimalist hammerhead ribozyme. The catalytic core as well as the three flanking helices are shown in a Y-shape conformation mediated by coaxial stacking of stem II and III.

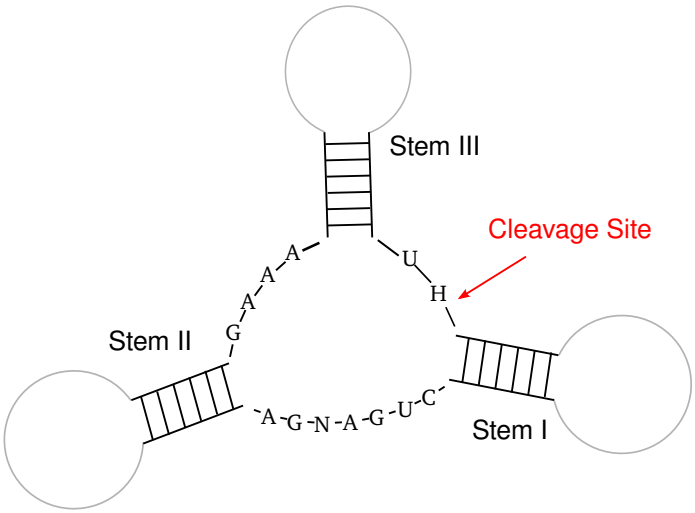



**Figure 3.2:** Secondary structure of a hammerhead ribozyme. Nucleotides of the catalytic core are rather conserved. Hammerhead ribozymes are classified according to the presence or rather absence of loops at the end of the stems flanking the core nucleotides. For example, if the stem helix I is open-ended, we are concerned with type I. Hence, the existence of loops is optional and depends on the hammerhead ribozyme type. In addition, the stems are of variable size. The cleavage site is marked.

interactions. Loops may interact via non-canonical base pairings and stacking of individual, non-adjacent bases [Przybilski & Hammann, 2006]. These findings were further supported by the first crystal structure of a full-length hammerhead ribozyme that was proposed by Martick & Scott [2006]. Hammerhead ribozymes are classified based on the presence or rather absence of loops and/or bulges at the end of the helices. Fig. 3.2 presents a schematic view of a secondary structure of the hammerhead ribozyme to illustrate the classification. Hammerhead ribozymes of type I feature an open end stem helix I, whereas for type II and type III stems II and III are open-ended, respectively. The cleavage site is located between stems I and III behind the motif UH; cleavage depends on a physiological $Mg^{2+}$ concentration [de la Peña *et al.*, 2003; Khvorova *et al.*, 2003; Martick & Scott, 2006]. Although hammerhead ribozymes have been detected among all domains of life [de la Peña & García-Robles, 2010b], the first natural hammerhead ribozyme of type II was not found until recently [Perreault *et al.*, 2011] along with sequence variations of the hammerhead consensus. In this study, we will focus on hammerhead ribozymes of type I and III only. As ribozymes need to maintain their structure, we hypothesize that there needs to exist substantial coevolution within the molecule.

To this end, we employ mutual information (MI) originating from information theory [Shannon, 1948] to detect coevolutionary signals within hammerhead ribozymes. MI quantifies the

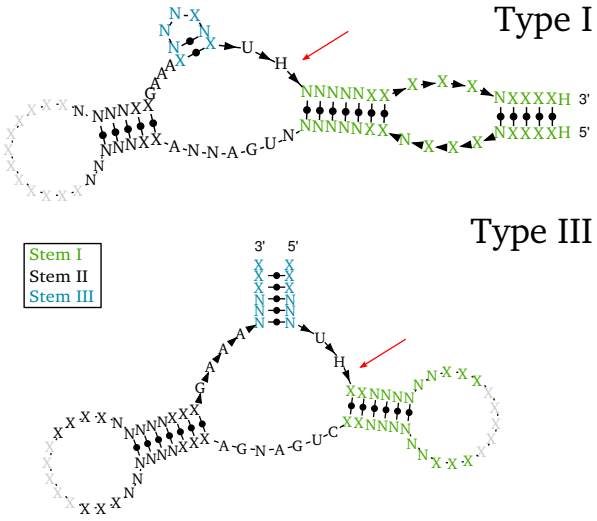



**Figure 3.3:** The structural descriptors of hammerhead motifs type I and III that have been used for identification of potential ribozymes as well as for the truncation of the resulting sequence prior to the coevolutionary analysis are shown. We highlighted the helices flanking the conserved core as well as the cleavage site which is indicated by the red arrow.

amount of information a position $i$ contains about another position $j$. First introduced by Korber *et al.* [1993] to study the highly variable V3 loop of the HIV-1 envelope protein based on sequence information, the non-parametric method has been widely used to examine coevolution among and within proteins [Boba *et al.*, 2010; Caporaso *et al.*, 2008; Fatakia *et al.*, 2009; Gloor *et al.*, 2005; Hamacher, 2008].

## 3.1 Methods

### Data Preparation

Based on a sophisticated genome search encompassing numerous databases, amongst others Ensembl [Flicek *et al.*, 2010], the Subviral RNA Database [Rocheleau & Pelchat, 2006] as well as NCBI sources, Seehafer *et al.* [2011] identified 160 new motifs of hammerhead ribozymes type III originating from various eukaryotes and bacteria as well as two motifs that have already been found in *Arabidopsis thaliana* [Przybilski *et al.*, 2005]. In addition, the pipeline proposed by Seehafer *et al.* [2011] which comprises filtering steps in both sequence space and physical realm requiring the capability to exhibit a hammerhead-like fold proved itself correct since 122 viroid motifs deposited in the Subviral RNA Database [Rocheleau & Pelchat, 2006] have been detected as well. In the following, we will refer to these data sets as A1 (prokaryotes and eukaryotes) and A2 (viroids), respectively. For a coevolutionary meta analysis, we extracted interesting regions of the hammerhead motif applying constraints that have been used for identification of the respective motifs. Besides the core motifs UH, GAAA and CUGANGA with H = {A,C,U} and N = {A,C,G,U}, we applied size restrictions on the flanking helices: stem I consists of 4-6 base pairs, stems II and III comprise 4-7 and 3-6 base pairs, respectively. Note that additionally to standard Watson-Crick base pairs the wobble pair U-G is accepted as well. Since the loop sizes of loop I and II may vary between 4 and 99/100 nucleotides, we expanded stems I and II by up to 5 loop nucleotides each yielding a maximum number of ten loop nucleotides (see Fig. 3.3).

In a complementary approach, a consensus sequence of hammerhead ribozymes type I has been constructed based on all available motifs deposited in various databases. Thus, genomes have

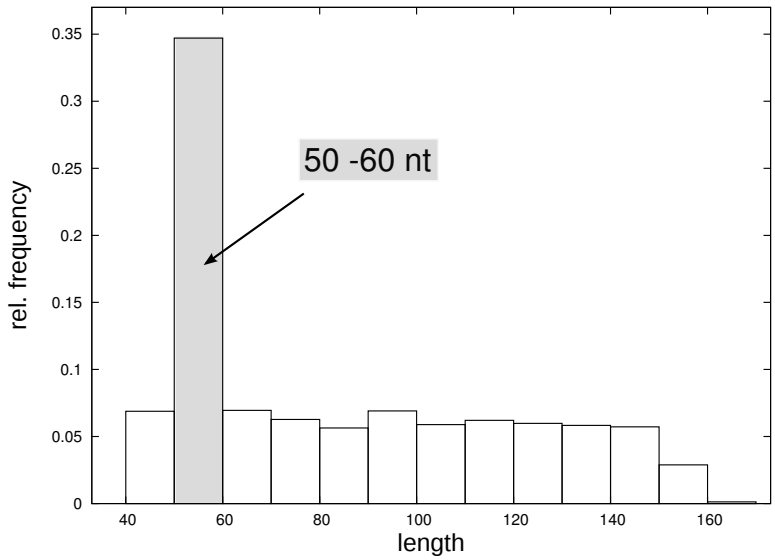



**Figure 3.4.:** Histogram of the lengths of the identified hammerhead motifs type I computed with bins encompassing 10 nucleotides. For further analysis, only sequences with a length between 50 and 60 nucleotides are used (highlighted in gray).

been analyzed using the derived descriptor and a subsequent application of the filter steps described by Seehafer *et al.* [2011] yielding 4,719 hammerhead motifs type I (for descriptor see Fig. 3.3). Note that except for loop II which comprises between 3 and 110 nucleotides, only small variations (up to 4 nucleotides) are allowed for different parts of the motif. A closer inspection of the resulting sequences revealed a distribution of motif lengths biased towards sequences comprising 50 to 60 nucleotides in total (see Fig. 3.4). Based on these results, we focused on the subset of 1,769 sequences of hammerhead ribozymes type I whose lengths are in the respective range assuming the other sequences to be more unlikely variations of the "true" hammerhead motif type I. For this subset of ribozyme sequences, we considered the complete sequences of loop II for further analysis. In the following, we will refer to these sequences as set B. Parts of the motif were defined by help of the structural descriptor that has been used for identification of the ribozyme.

After extracting the described portions of helical and loop sequences for the motif sequences of each set using the `R` [R Development Core Team, 2008] package `Biostrings` from the `BioConductor` software repository [Gentleman *et al.*, 2004], we arranged the substrings in constrained multiple sequence alignments (MSAs) each. Since we aim to maintain a proper ribozyme structure, base pairings and size restrictions need to be retained. Hence, for helical parts, shorter substrings were arranged according to those of maximum length simultaneously for both sequences composing the respective stem helix. By this procedure, we obtained hammerhead ribozymes with maximum lengths of the core helices. Loop regions were aligned using `clustalw2` [Larkin *et al.*, 2007] with standard parameters and subsequent refinement of the yielded alignment by manual optimization utilizing `Jalview` [Waterhouse *et al.*, 2009]. Afterwards, the distinct substrings were recombined for each sequence, i.e. we concatenated the sequence parts while maintaining the correct order of the original sequence. Thus, we yielded a truncated hammerhead motif alignment lacking sequence parts that were not investigated in this study (for sequence lengths see Tab. 3.1).

| Set A1 | Set A2 | Set B |
|---|---|---|
| 78 nt | 55 nt | 77 nt |

**Table 3.1.:** Recombining the constrained multiple sequence alignments that were obtained for each structural part of the hammerhead motifs yielded an alignment for each set of sequences. The obtained sequence lengths are shown in this table. Differences are mainly due to the arrangement of loops.

### Mutual Information

To detect potentially coevolving ribozyme positions, the information theoretical measure MI is utilized. Derived from the Kullback-Leibler divergence [Kullback & Leibler, 1951] MI quantifies the deviation of the joint probability distribution of two sites $i$ and $j$ to the null model of independent evolution of the two sites sites. Hence, MI of two sites, i.e. columns of the MSA, $i$ and $j$ is defined as:

$$\mathrm{MI}_{ij} = \sum_{\sigma_i} \sum_{\sigma_j} P(\sigma_i, \sigma_j) \cdot \log_2 \frac{P(\sigma_i, \sigma_j)}{P(\sigma_i) \cdot P(\sigma_j)} \tag{3.1}$$

where $P(\sigma_i, \sigma_j)$ denotes the joint probability to find the symbols $\sigma_i$ and $\sigma_j$ at alignment positions $i$ and $j$ respectively. $P(\sigma_i)$ and $P(\sigma_j)$ are the corresponding marginal probabilities. We obtain the probabilities as relative frequencies of the symbols $\sigma$ that are drawn from a set comprising the four standard nucleotides as well as a gap character $\mathcal{S} = \{A, C, G, U, -\}$. To account for finite-size effects due to the number of sampled sequences, we employ the shuffle null model that was thoroughly investigated by Weil *et al.* [2009]. Here, in each iteration all alignment columns are shuffled independently and the MI of the obtained alignment is recomputed. The rationale behind this procedure is to maintain the variability of each site by destroying of possible interdependencies between different sites at the same time. We averaged MI values over 10,000 iterations and computed Z-scores from the resulting mean $\overline{\mathrm{MI}}_{ij}$ and standard deviation $S(\overline{\mathrm{MI}}_{ij})$ of each column pair $(i, j)$.

$$Z_{ij} = \frac{MI_{ij} - \overline{\mathrm{MI}}_{ij}}{S(\overline{\mathrm{MI}}_{ij})} \tag{3.2}$$

Z-scores measure the significance of a derived MI signal in terms of standard deviations from an average value. Hence, coevolutionary signals and noise can be separated. To ensure reasonable Z-scores, we set all Z-scores to zero that satisfy one of the following conditions:

- $Z_{ij} = \mathrm{NaN}$, $Z_{ij} = \pm\,\mathrm{Inf}$ (resulting from a division by zero)
- $|Z_{ij}| > 1,000$ (resulting from a division by a value, that is numerically zero)

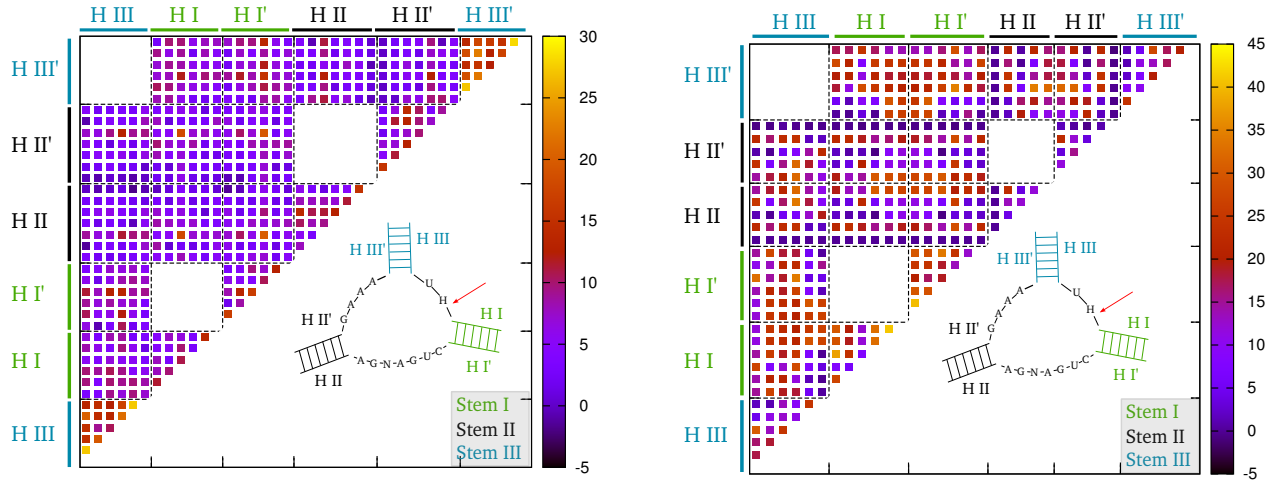
In addition, we computed both gap content and entropy for each alignment position to further evaluate the obtained MI results. All computations were performed using the `R` package `BioPhysConnectoR` [Hoffgaard *et al.*, 2010].

## 3.2 Results

MI and corresponding Z-scores obtained by the shuffle null model are computed for the truncated sequences of sets A1, A2 and B. MI measures the amount of information between any two positions of an MSA, whereas the Z-score quantifies the significance of the derived signals in term of standard deviations from the expectation value of a stochastic null model. Hence, large Z-scores indicate significant results for two coevolving sites. In the following, we will restrict the analysis to Z-scores since the null model eliminates coevolutionary noise. Note that both highly conserved sites and sites with a large number of gaps exhibit only weak MI and, thus, weak Z-score signals. Since we are interested especially in the coevolution of the hammerhead core, we discuss the results for the nucleotides comprising the helices I, II and III that are presumably comparable for all ribozyme sets under consideration.

The core of hammerhead ribozymes type III obtained from eukaryotic and prokaryotic genomes (set A1) is overall rather independently evolved as the Z-scores suggest (see Fig. 3.5(a)). Note that non-normalized MI values exhibit a similar coevolutionary pattern. Clearly, we detect coevolution between nucleotides of any two strands that form a stem helix due to base pairings. We omitted the graphical representation of these apparent coevolutionary effects since they dominate the color scale. In the following, we will discuss signals apart from base pairing only. Notably, we detect significant coevolutionary signals within strands that form stem III. This observation becomes even more evident in comparison with the other stems, where we find coevolution among nucleotides within the same strand as well. However, coevolution is not detectable for all nucleotides and the signal strength is reduced compared to stem III. Besides these signals, we also detect coevolution between sequences that form different helices: both strands of stem I are connected with strands of stem III sharing intermediate coevolution. This observation is rather surprising although it does not imply a direct interaction of stems I and III. Experimental studies have shown that interactions of loops and bulges at the ends of helices I and II are required for function and stability of the ribozyme [de la Peña *et al.*, 2003; Khvorova *et al.*, 2003] which is in perfect agreement to the detected coevolutionary signals. Hence, we may assume to detect coevolutionary signals for the respective stem helix strands as well. However, the "crosstalk" between stems I and II is limited to the coevolution of single nucleotides: nucleotide 3 of strand H I exhibits strong coevolution with nucleotides 3 and 5 of H II and H II' respectively, and consequently nucleotide 4 of H I' is strongly connected to the very nucleotides of stem II.

For set A2 that contains viroid sequences only, the coevolutionary pattern differs from that observed for set A1 (see Fig. 3.5(b)). Similarly as for set A1, the strongest coevolutionary signals were observed between each two strands forming a stem helix. However, significantly more coevolution among sites is detectable for viroid hammerhead ribozymes suggesting a higher preservation of the core motif than for pro- and eukaryotes. As verification for this hypothesis, we computed histograms of pairwise Hamming distances for the sequences within the truncated alignments of sets A1, A2 and B. The Hamming distance of two sequences is defined as number of differing positions; insertions/deletions denoted as gap in one of the sequences are considered as difference as well. The results are shown in Fig. 3.6 and exhibit a higher average sequence identity and, thus, less variation for viroid ribozyme sequences than for sets A1 and B. We find




**(a)** Z-scores of Set A1

**(b)** Z-scores of Set A2

**Figure 3.5.:** The Z-scores have been computed for core nucleotides that form the helices flanking the conserved hammerhead motif as normalized mutual information values with respect to an evolutionary null model. The location (color coded) of the stem helix strands is illustrated by the sketch of the ribozyme structure. The red arrow indicates the cleavage site. Numbering of nucleotides and strands starts at the 5' end. Z-scores between any two strands that form a stem helix (blank squares) are left out since those values dominate the color scale and, thus, mask nucleotides that coevolve not by direct base-pairing selection.

pronounced coevolution within the strands that form stem I, and to a smaller extent also for helices II and III, whereas for set A1 we observed distinct signals of intra-strand coevolution for stem III. Again, we notice a connection between stems I and III, however, more pronounced than for set A1, where only few nucleotides were involved. The data suggest coevolution of the outer parts of stem I with the complete stem III. Interestingly, we also observe strong signals pointing towards a connection between the outer helical parts of stem I and II which is in perfect agreement with experimental results [de la Peña *et al.*, 2003; Khvorova *et al.*, 2003].

The analysis was performed for set B comprising more than 1,500 sequences of hammerhead type I with a sequence length ranging from 50 to 60 nucleotides. The sequences were extracted from various genomes by help of a descriptor based on a structural consensus sequence obtained from all known hammerhead motifs type I. In a subsequent step, a size restriction filter was applied. We again computed both MI and Z-scores for the core residues and additionally for nucleotides of the CUGANGA motif which has been defined less strict as NUGANNA in accordance to Perreault *et al.* [2011] who identified variations of nucleotides that have been considered conserved before. The pairwise Hamming distances computed for the resulting alignments revealed an intermediate level of sequence identity being less conserved than viroid sequences of set A2 but with a more conserved sequence than hammerhead motifs of set A1 (see Fig. 3.6). Apart from high Z-scores due to paired bases within helical regions, we notice a much higher scale for coevolutionary signals than for sets A1 and A2 (see Fig. 3.7) which is presumably an artifact of the sequence numbers included for the analysis: whereas for sets A1 and A2 we are concerned with about 150 sequences, we compute Z-scores of hammerhead ribozymes type I for more than 1,500 sequences. In analogy to the results obtained for set A2, we

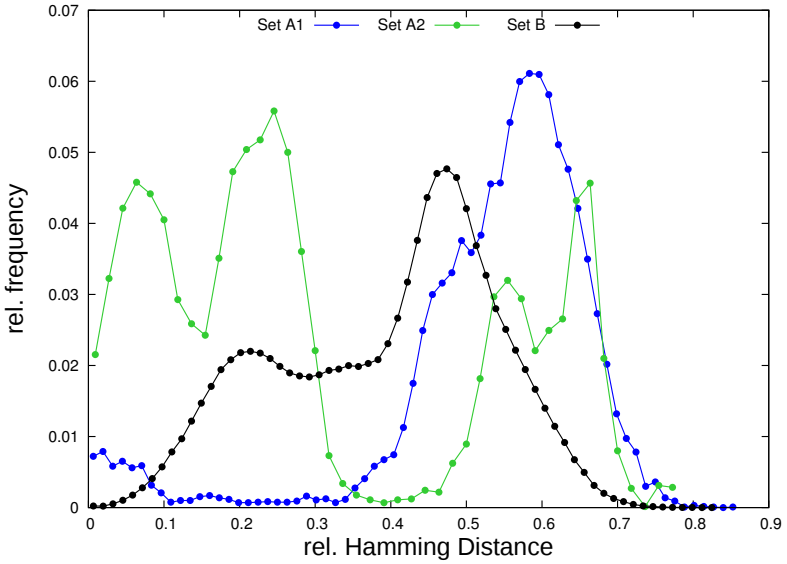



**Figure 3.6.:** Histogram of the pairwise Hamming distance of the truncated alignments that have been reduced to stem helices and loop nucleotides. We obtained the relative distances by normalizing with respect to the sequence length of the alignment. The results are shown for all three sets of ribozymes.

detect coevolution within strands forming stem I. In addition, we notice signals of coevolution among sites connecting inner nucleotides of stems I and II, whereas for set A1 we noticed nearly no coevolution, and for set A2 the signals were detectable especially for the outer parts of those helices. These findings indicate differing selective pressure for different types of hammerhead motifs and organisms. Furthermore, we observe coevolution between the outer part of stem III, which comprises a single nucleotide in the actual consensus structure, and the inner parts of stems I and II. This suggested the identification of coevolutionary signals between strands of helices I and III for each set of ribozyme sequences. An experimental verification of this hypothesis has yet to be confirmed.

For set B, we investigated potential coevolving sites with respect to the NUGANNA motif, which has been examined in its more conserved version CUGANGA for sets A1 and A2 as well. For position 5 of both motifs (emphasized in bold: CUGA**N**GA, NUGA**N**NA) we detect signals of coevolution for sets A1 and B, but not for set A2. Additionally, the pattern of coevolution differs slightly for sets A1 and B since for set B nucleotides of all flanking helices are involved; positions of stem I exhibit negligible signals for set A1 only. We assume the viroid sequences to be more conserved at this position which comes as a natural consequence of both Z-score and Hamming distance results. To verify this assumption, we computed nucleotide probabilities and the resulting entropies for the respective position of the alignments of each set. The results, as shown in Tab. 3.2, actually indicate a higher level of conservation of position 5 of CUGANGA within viroid sequences. Since we still find an intermediate entropy value, the lack of coevolutionary signals cannot completely be explained by computational means alone.

For hammerhead ribozymes type III (sets A1 and A2), we additionally performed the analysis for up to 10 nucleotides of loops I and II. The loop nucleotides were extracted starting simultaneously from the stem nucleotides. The resulting loop sequences were aligned separately

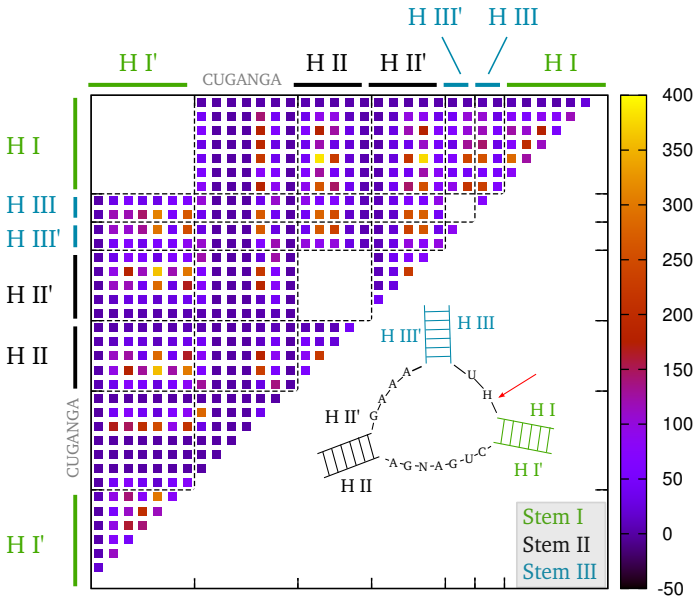



**Figure 3.7.:** Z-scores of core nucleotides that were computed from the alignment of hammerhead ribozymes type I as mutual information normalized with respect to the shuffle null model. Helices and their locations in the motif are highlighted. The red arrow indicates the cleavage site. Z-scores between any two strands that form a stem helix (blank squares) are left out since those values dominate the color scale and, thus, mask nucleotides that coevolve not by direct base-pairing selection.

| | A | C | G | U | Entropy [bit] |
|---|---|---|---|---|---|
| set A1 | 0.247 | 0.130 | 0.117 | 0.506 | 1.740 |
| set A2 | 0.008 | 0.229 | 0.000 | 0.762 | 0.843 |
| set B | 0.126 | 0.114 | 0.474 | 0.287 | 1.760 |

**Table 3.2.:** For each set of sequences (A1, A2, B) the probability of each nucleotide type of position 5 of the conserved CUGA**N**GA, NUGA**N**NA (emphasized in bold) as well as the resulting entropy of the respective position is computed.

afterwards. As loop sizes vary within a range of 4 to 99/100, it is not quite clear how these sequences can properly be arranged in the sequence ensemble. Therefore, we had to exclude these portions from further analysis. For set B, we already applied a size restriction to the sample of sequences, and, hence, restricted the size of the loop located at the end of stem helix II. Before we determine the coevolution between loop II and other sites as well as within itself, we examined the distribution of loop sizes which is shown in Fig. 3.8. Interestingly, small, odd numbers of nucleotides (5, 7, 9) are dominant in the depicted histogram, which are shown to be energetically favored loop sizes where kissing complexes yield the maximum number of base pairs [Tinoco & Bustamante, 1999]. The loop sequences have been aligned, and we examined whether we could find signals of coevolution of loop nucleotides within the loop or with other ribozyme sites. Actually, we observe coevolution of loop positions with nucleotides of the core

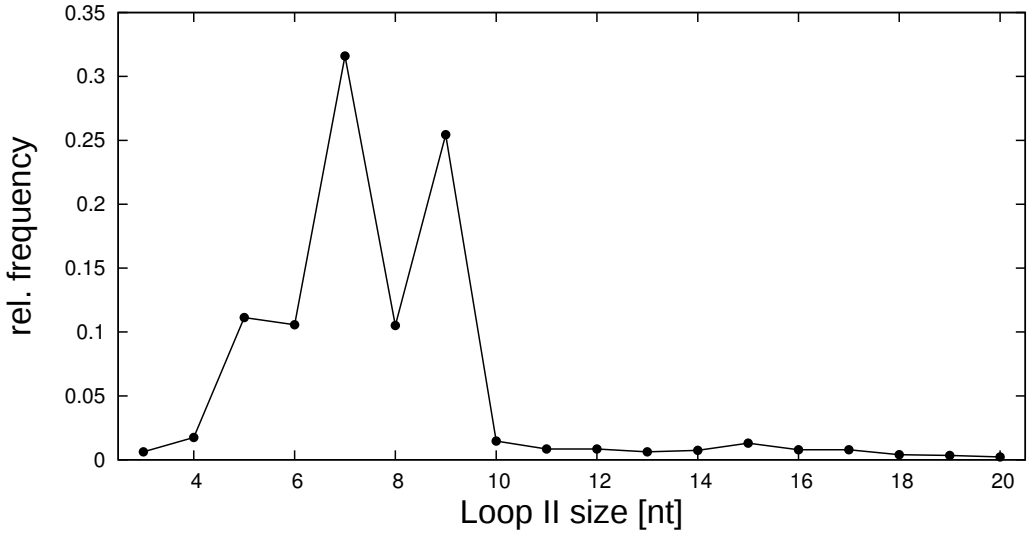



**Figure 3.8.:** Relative frequencies of loop sizes occuring in the set of hammerhead sequences type I with a size ranging from 50 to 60 nucleotide (nt).

helices. Due to the diverse loop lengths, many positions of the loop alignment exhibit no signals, which is an effect originating from the high gap content at those positions (data not shown).

In this study, we examined coevolutionary connection of nucleotides within different sets of ribozyme sequences. We focused on the helices flanking the conserved hammerhead motifs. Structural elements of each sequence have been extracted from the full sequences and recombined into a truncated alignment that contained only parts of interest. Since we are concerned with excerpts of complete hammerhead motifs, duplicates of truncated sequences are not removed from the analysis, since the corresponding complete sequences represent full variability of the motif. The effect reveals itself in the histogram of pairwise Hamming distances (see Fig. 3.6). Here, we find Hamming distances equal zero which corresponds to identical sequences. Nevertheless, removal of duplicate sequences does not change the overall pattern of coevolution, but decreases the signal strength only (data not shown).

Although a direct comparison of explicit helical nucleotide positions is not possible due to differing stem sizes in sets A1, A2 and B, we were able to give evidence for coevolving tendencies that are both different for or rather common to all sequence sets from different kingdoms of life. In general, we notice a higher conservation of viroid hammerhead motifs, which may be responsible for the fact that we find no signal of coevolution of the variable position in the CUGANGA motif which is common to all hammerhead ribozymes. Both for sets A1 and B, we detected signals for this respective position with nucleotides in helices. Additionally, the coevolution within sequences of set A2 is more pronounced than for the other sets. Whereas previous studies revealed tertiary loop interactions to be crucial for fold and function of ribozymes [de la Peña *et al.*, 2003], we detected corresponding coevolution signals for parts of the respective helices for sets A2 and B, and for single nucleotides only in set A1. Interestingly, we identified a connection of helical stems I and III for all sequence sets under consideration which had not been reported yet. To explain this outstanding finding, we examined if those helices are enhanced sites for base pairing with respect to a null model as well. However, this hypothesis could not be verified by computational means alone (data not shown). The results imply an alternate

mechanism of coevolution for these sites, whose verification is called for by experimental means we strongly suggest.

## 3.3 Contributions

The search for hammerhead motifs and the resulting sequence sets have been kindly provided by C. Seehafer and C. Hammann [Seehafer *et al.*, 2011] who discovered the unusual distribution of sequence lengths of hammerhead ribozymes type I as well which is shown in Fig. 3.4. All other analysis and modeling was done in the framework of this PhD thesis.

# 4 Additional Contributions

This chapter gives an overview of contributions to additional projects that have not been discussed thus far, corresponding publications will be referenced. For each section, we will give a short summary of published methods and results, the contributions are declared at the end of each section.

## 4.1 BioPhysConnectoR: Connecting sequence information and biophysical models

We designed an add-on package for the statistical software R [R Development Core Team, 2008], combining both sequence based and biophysical approaches, that have been presented in previous sections of this thesis, to gain insights into a protein's coevolution and dynamics. Further details are presented in the subsequent parts, that have been published in:

Hoffgaard, F; Weil, P; Hamacher, K (2010) BioPhysConnectoR: Connecting sequence information and biophysical models. *BMC Bioinformatics* 11:199.

### 4.1.1 Background

Proteins are ubiquitious in all cells and organisms, and important for essential functions. Yet, a holistic picture annotating functional and evolutionary features is still missing. Huge databases like GenBank [Benson *et al.*, 2011] and PDB [Berman *et al.*, 2000] contain myriads of gene and protein sequences gathered from a wide variety of organisms. Chen *et al.* [2004] collected more than 40,000 sequences of HIV protease and reverse transcriptase, that are crucial proteins within the viral life cycle to reveal both conserved domains and evolutionary hot spots. Furthermore, comparisons of gene or protein sequences among diverse organisms from all branches of the tree of life, may facilitate the understanding about phylogenetic relationships. The sequence space is subject to the mutational operator. Modifications of the underlying sequences can result in differing phenotypes that are subject to selection processes in the biophysical realm. Thus, point mutations may require further compensatory mutations to restore protein structure and/or function. Such coevolutionary relationships within proteins can be identified for example by examination of MSAs using MI [Shannon, 1948] an information-theoretical measure that was discussed in section 3 in the context of ribozymes. Such sequence-based methods allow high-throughput analyses but fail to explain biophysical implications of sequence changes.

Physical properties of biomolecules are investigated by sophisticated methods, such as MD simulations or NMA. By physical methods, we can evaluate implications of mutations in sequence space on protein structure and stability which mediate protein function. In contrast to sequence-based approaches, biophysical methods are computationally expensive and, thus, allow the investigation of a few mutants only. In the past, ENMs [Atilgan *et al.*, 2001; Bahar *et al.*, 1997]

have emerged as coarse-grained protein models that are capable to reflect protein dynamics (see section 1.1). Utilizing these reduced molecular models, large numbers of mutants can be screened for their mechanical aberration caused by sequence changes. Hamacher [2008] developed a protocol to combine both sequence-based and biophysical methods to annotate protein function in terms of structure and coevolution.

### 4.1.2 Methods

As a sequence-based measure to identify coevolving protein positions, we employ MI defined as in section 3.1. To account for finite-size effects, a shuffle null model is included to determine the significance of observed MI results [Weil *et al.*, 2009].

To quantify biophysical implications of introduced sequence changes, we employ ANMs as proposed by Atilgan *et al.* [2001] (see section 1.1). The mechanics of a protein structure can be examined by computing the Hessian matrix $H$ (see Eq. 1.5) and via SVD the corresponding covariance matrix $C$ (see Eq. 1.6). Mutations are introduced by either changing the underlying sequence or by altering amino acid contacts. Sequence changes can be investigated whenever amino acid specific interaction potentials are employed. Structural modifications by altering contacts can be performed by in- or decreasing the rigidity of the respective connection as well as by deleting it. Resulting changes in protein dynamics are directly contained in the resulting covariance matrix $C^{\text{mut}}$ of the mutated system and can be quantified by help of the FN (see Eq. 1.26).

The `R` package `BioPhysConnectoR` includes source code of `bio3d` [Grant *et al.*, 2006] and utilizes routines from the packages `matrixcalc` [Novomestky, 2008] and `snow` [Tierney *et al.*, 2009]. We integrated native `C/C++` code to adress runtime issues for computationally expensive routines. Both low-level functions and protocols that have been included in the software package can be customized by various arguments. Furthermore, we added auxiliary methods to read, write and convert data from PDB and alignment files in fasta format. For ANM computation, both interaction potential matrices MJ and KE are available. `BioPhysConnectoR` can be obtained from CRAN at `http://cran.r-project.org/` or directly from `http://bioserver.bio.tu-darmstadt.de/software/BioPhysConnectoR`.

For further details see: Hoffgaard *et al.* [2010].

### 4.1.3 Results

The `R` package `BioPhysConnectoR` contains utilities to perform an integrated protein analysis combining data from sequence-based (as in section 3) and biophysical approaches (as in section 2). We implemented two mutation scenarios to investigate biophysical implications of sequence changes. Although its application is limited, the routine `sims` computes the changes of mechanical properties of the protein by exchanging the whole sequence information based on an MSA. Limitations of this method arise from the usage of the MSA since gaps have no correspondence in the biophysical realm. Furthermore, effects of altered sequences can only be measured if amino acid specific interaction potentials are employed.

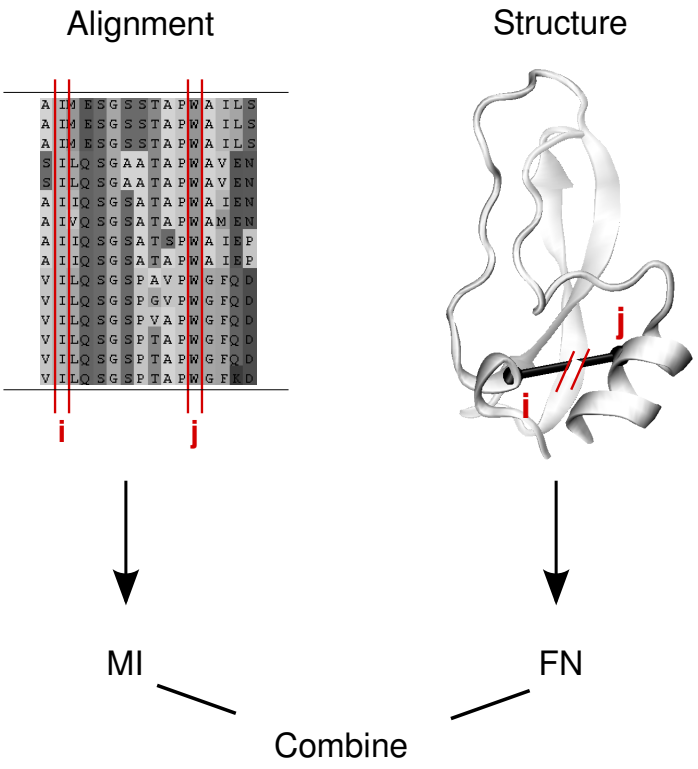



**Figure 4.1:** The flowchart demonstrates the combination of both sequence-based and biophysical approach. From a multiple sequence alignment the MI (Eq. 3.1) is computed for any pair of residues $i$ and $j$. In the biophysical realm, we artificially delete the connection of the same (contacting) residues $i$ and $j$, and compute the FN (Eq. 1.26) that quantifies the changes in protein dynamics for the deleted contact. From the combination of MI and FN, we can identify interactions crucial for stability and function.

An alternative scenario mimics the effect of point mutations: by use of the function `simc` all non-bonded amino acid contacts are deleted ("switched off") one at a time to compute individual FN values. In experiment, by introducing point mutations, e.g., alanine mutagenesis studies, important connections of residues are weakened or destroyed, albeit in experimental studies all interactions of the mutated amino acid are affected. The alteration of protein dynamics is measured by FN of the difference covariance matrices. A similar study was discussed in section 2.2. In the integrated approach which was applied to acetylcholinesterase (see section 4.4), we overlay FN results from the switch-off procedure with corresponding MI obtained for the MSA, a schematic picture is given in Fig. 4.1. As example application, we applied the presented protocol to HIV-1 protease (PDB code 1KZK [Reiling *et al.*, 2002]). Since the data base provided by Chen *et al.* [2004] comprises more than 40,000 sequences, we need not consider finite-size effects for the computation of MI.

The superposition of resulting MI and FN values is illustrated in Fig. 4.2. Note, only MI values are shown that correspond to a deleted contact yielding an FN value. The scatterplot is partitioned into four segments (I, II, III and IV). Residue pairs in quadrant II that feature both low MI and FN values do neither coevolve nor are their contacts crucial for protein dynamics. Additionally, we identify coevolving residues in I whose contact has no implications on structural properties of the protein. We hypothesize that a coevolution of those residues accounts for other than mechanical protein features necessary for functionality, such as electrostatic or size effects. Contacts of segment III feature low MI but we find them being important for the dynamics, i.e. the corresponding amino acids take part in concerted molecular motions to mediate protein function. Residue pairs assigned in IV show high coevolution as well as important connections in the biophysical realm, which is often caused by spatial proximity of the participating amino acids.

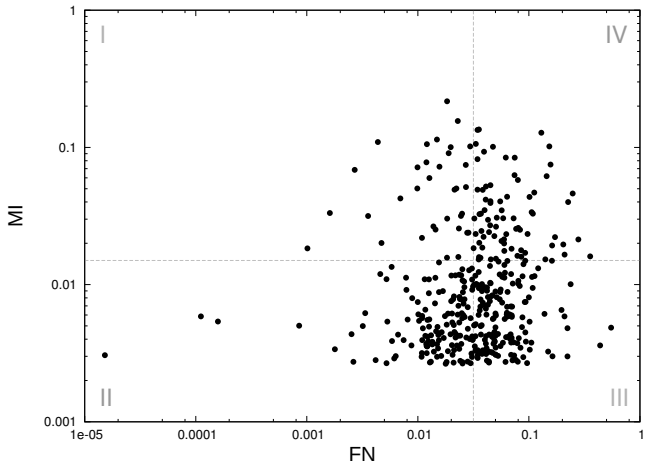



**Figure 4.2:** The switch-off scenario is applied to the HIV-1 protease (PDB code 1KZK). Each amino acid contact is deleted one at a time, the resulting change in protein dynamics is quantified by computing FN (Eq. 1.26) of the difference covariance matrices. The scatterplot illustrates the superposition of MI (Eq. 3.1) and FN values of all contacting residue pairs. The figure is adapted from Hoffgaard *et al.* [2010].

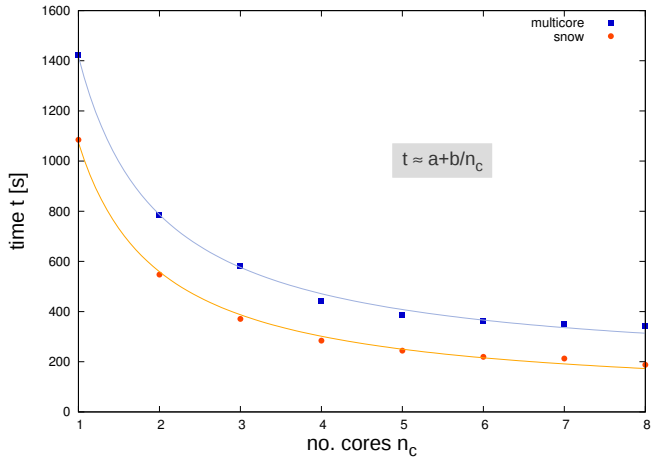



**Figure 4.3:** Performance of the parallelized routine `simc` that simulates the switch-off scenario, i.e. each contact is deleted one at a time and the resulting change in protein dynamics is quantified by help of the Frobenius norm. To this end, we used one up to eight cores and compared the parallelization obtained with the `R` packages `multicore` [Urbanek, 2009] and `snow` [Tierney *et al.*, 2009]. We fitted the elapsed time to a scaling law of the form $t \approx a + b/n_c$ with constants $a$ and $b$. The figure is adapted from Hoffgaard *et al.* [2010].

Depending on the definition of the cutoff distance which specifies contacts between amino acids, we are concerned with large numbers of residue pairs. To this end, we parallelized the switch-off routine to account for runtime issues. We tested the efficiency of parallelized code for two different implementations using the `R` packages `multicore` [Urbanek, 2009] and `snow` [Tierney *et al.*, 2009] for up to eight processors on yielding an efficient parallelization in accordance to Amdahl's law [Amdahl, 1967].

We implemented the package `BioPhysConnectoR` for statistical software environment `R` that is open to a wide community of scientists offering lots of additional package, e.g. from the Bioconductor software project [Gentleman *et al.*, 2004]. `BioPhysConnectoR` contains routines to compare the dynamics of a native protein with systematically mutated proteins with alterations in sequence or single contacts. Mutations are scored by the information-theoretical MI that evaluates the extent of coevolution of residues and FN assessing the importance of amino acid contacts on the dynamics of the protein. Routines of presented package have been utilized for projects of this thesis. Further work is in progress to develop and implement more efficient algorithms for fast network-based protein analysis [Hamacher, 2010].

### 4.1.4 Contributions

KH supplied protocols for connecting sequence information and biophysical properties and for computing the self-consistent pair contact probability, a method proposed by Micheletti *et al.* [2001] to account for non-harmonic effects as well. The R implementation (including the C/C++ code) was carried out by FH and in parts by PW. The parallelization was exclusively done by FH. All authors participated in writing the manuscript.

## 4.2 Distance-dependent classification of amino acids by information theory

We applied an information theoretical approach to derive distance-dependent amino acid alphabets. Reduction schemes resulting from the alphabets were employed in sections 1.3.2 and 1.3.1. Here, we present a summary of the proposed method. The results of this study have been published in:

Pape, S; Hoffgaard, F; Hamacher, K (2010) Distance-dependent classification of amino acids by information theory. *Proteins* 78:2322.

### 4.2.1 Background

A protein sequence, also called primary structure, is represented by the amino acid alphabet, which consists of 20 standard as well as modified or unusual amino acids. An issue of protein folding studies is the reduction of the amino acid alphabet by grouping amino acids by similar chemical or physical properties. Shepherd [1981] was able to detect degenerated DNA patterns on basis of a reduced alphabet, that would hardly be detectable using standard DNA codes. Similar patterns at the protein level may be identified as well based on simplified amino acid sequences. The simplest among the reduction schemes is based on hydrophobicity as only driving force and discriminates hydrophobic and polar amino acids. Wolynes [1997] pointed out that more heterogeneity is necessary to reflect the full complexity of a protein. In experimental studies, it was shown that functional proteins with native fold can be encoded by using a subset of up to five amino acids only [Brown & Sauer, 1999; Munson *et al.*, 1994; Riddle *et al.*, 1997; Rojas *et al.*, 1997]. In addition, the question of minimal amino acid alphabets was tackled by diverse theoretical approaches [Cannata *et al.*, 2002; Cieplak *et al.*, 2001; Cline *et al.*, 2002; Shepherd *et al.*, 2007; Wang & Wang, 1999]. The resulting simplified amino acid alphabets clearly distinguish between hydrophobic and polar amino acids (see Tab. 4.1). Most, if not all, of the studies share a rather strict contact or interaction distance, taking into account only nearest neighbors. Considering long-range interaction of amino acids, we may derive different reduction schemes. In this study, we deduce distance-dependent amino acid alphabets from empirical distance distributions of amino acid pairs using an information theoretical metric.

| | alphabet | hydrophobic | polar |
|---|---|---|---|
| 1 | [Wang & Wang, 1999] | (IVLMWCFY) | (HAT)(GP)(DE)(NQRKS) |
| 2 | [Cieplak *et al.*, 2001] | (FLI)(WMVCY) | (HA)K(NPSTDEGQR) |
| 3a | [Shepherd *et al.*, 2007] | (FLI)(WMVCY) | (HATGP)(DENQRKS) |
| 3b | [Shepherd *et al.*, 2007] | (FL)I(WMV)(CY) | (HAT)(GP)(DES)(NQRK) |

**Table 4.1.:** Reduced amino acid alphabets obtained by various computational studies. Groups of amino acids are indicated by parentheses.

### 4.2.2 Methods

Using the SCOP [Murzin *et al.*, 1995] database, we defined a set of 2,830 proteins[1]. Each protein consists of a single chain and is a representative of a distinct SCOP class. Statistics on contact type distributions are extracted by computing distance-dependent ($r$) and amino acid type-specific ($i, j$) histograms $n_{ij}^{(\mathrm{emp})p}(r)$ for each protein $p$. We observed that distances beyond 50 Å are subject to statistical fluctuations (see Fig. 4.4), and, thus, we restricted the analysis to distances below maximum distance $R_{\max} = 50$ Å. To account for a systematic bias due to different protein sizes, we applied a normalization procedure. For a protein $p$ of size $R_p$ the non-biased number of contacts $n_{ij}(r)$ is related to the empirically measured number of contacts $n_{ij}^{(\mathrm{emp})p}(r)$ at a distance cutoff $r$:

$$n_{ij}^{(\mathrm{emp})p}(r) = n_{ij}(r) \cdot \Theta(R_p - r) \tag{4.1}$$

$\Theta$ denotes the Heaviside step function. The number of measured contacts averaged over $N$ proteins is computed as follows:

$$n_{ij}^{(\mathrm{emp})}(r) = \frac{1}{N} \sum_p n_{ij}^{(\mathrm{emp})p}(r) \tag{4.2}$$

$$= \frac{1}{N} \sum_p n_{ij}(r) \cdot \Theta(R_p - r) \tag{4.3}$$

Thus, the unbiased number of contacts $n_{ij}(r)$ for amino acid types $i$ and $j$ can be derived as:

$$n_{ij}(r) = \frac{n_{ij}^{(\mathrm{emp})}(r)}{h(r)} \qquad \text{with} \quad h(r) = \frac{1}{N} \sum_p \Theta(R_p - r) \tag{4.4}$$

Due to our normalization procedure, only contacts of proteins $p$ with size $R_p \geq r$ are considered in the corrected histograms $n_{ij}(r)$.

[1] For PDB codes see Supporting Information of Pape *et al.* [2010].

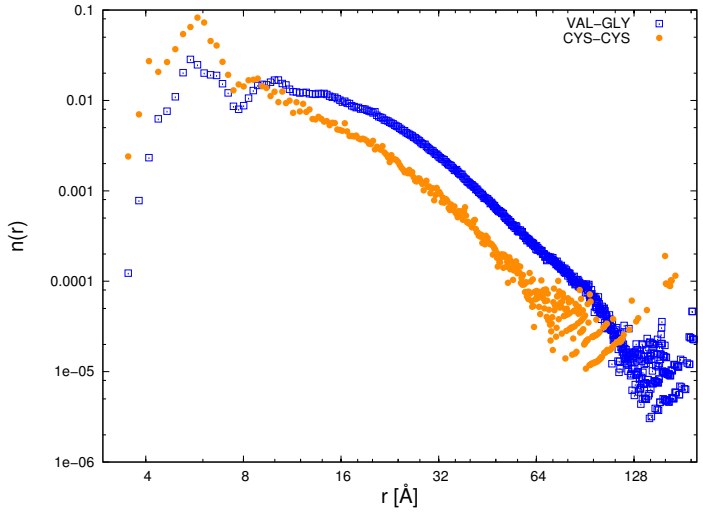



**Figure 4.4.:** The normalized histograms $n_{ij}(r)$ are shown for two types of amino acid pairs $(i, j)$, namely CYS–CYS and VAL–GLY. Normalization was applied due to size effects by $h(r)$ and the spherical volume element $4\pi r^2 \cdot \Delta r$. Picture is adapted from Pape *et al.* [2010].

To determine the information distance of two histograms $n_{ij}(r)$ and $n_{uv}(r)$, we defined a distance matrix $M = \sqrt{H}$, where $H$ denotes the Jensen-Shannon matrix which is computed as follows:

$$H(n_{ij}, n_{uv}) = \frac{1}{2} D_{kl}(n_{ij}||n) + \frac{1}{2} D_{kl}(n_{uv}||n) \qquad \text{with} \quad n = \frac{1}{2}(n_{ij} + n_{uv}) \tag{4.5}$$

The Kullback-Leibler divergence $D_{kl}(P||Q)$ is an information theoretical measure of how much probability distribution $P(x)$ deviates from the distribution $Q(x)$:

$$D_{kl}(P||Q) = \int P(x) \log_2 \frac{P(x)}{Q(x)} dx \tag{4.6}$$

The contribution of any amino acid to the respective amino acid pairing and, hence, to the overall distance matrix $M$ is derived by application of spectral decompositions of $M$ (for more details see Pape *et al.* [2010]). Analyzing the correlation of entries of the resulting eigenvectors quantifies the similarity of the respective amino acids. To this end, we computed the correlation matrix $C$. Since we want to merge similar amino acids into groups of a reduced alphabet, we defined a distance matrix $D$ based on the correlations of the amino acids:

$$D := \frac{\max(C) \cdot I - C}{\max(C) - \min(C)} \tag{4.7}$$

where $I$ is the identity matrix. High correlation coefficients $C_{ij} \approx 1$ that describe a high similarity of amino acids yield small distances $D_{ij} \approx 0$ between the respective amino acids, whereas the distance of anti-correlated amino acids ($C_{ij} \approx -1$) is approximately one. Reduced alphabets, i.e. groupings of similar amino acids, were derived by application of clustering algorithms, e.g. `hclust` [Murtagh, 1985].

| $R = 8$ Å | | $R = 50$ Å | |
|---|---|---|---|
| 20 | ND H ST GP C W FY IV LM RK A QE | 20 | LM W FY C IV NP H ST AG DE Q RK |
| 12 | (ND) H(ST) (GP)C W(FY) (IV)(LM) (RK) A(QE) | 12 | (LM) W(FY) C(IV) (NP) H(ST) (AG) (DE) Q (RK) |
| 7 | (ND)(HST) (GPC) (WFY) (IVLM) (RK)(AQE) | 8 | (LM)(WFY) (CIV) (NP)(HST) (AG) (DE)(QRK) |
| 4 | (NDHST) (GPCWFY) (IVLM)(RKAQE) | 5 | (LMWFY)(CIV) (NPHST)(AG) (DEQRK) |
| 2 | (NDHSTGPCWFY)(IVLMRKAQE) | 3 | (LMWFYCIV) (NPHSTAG)(DEQRK) |
| 1 | (NDHSTGPCWFYIVLMRKAQE) | 2 | (LMWFYCIV)(NPHSTAGDEQRK) |
| | | 1 | (LMWFYCIVNPHSTAGDEQRK) |

**Table 4.2.:** Amino acid grouping schemes for different maximum distances. For each simplified alphabet the number of different symbols is given. Groups are indicated by parentheses. Table was taken from Pape *et al.* [2010].

For further details see: Pape *et al.* [2010].

### 4.2.3 Results

We defined two different maximum interaction ranges $R_{\max}$ =8 Å and $R_{\max}$ =50 Å for the computation of the amino acid pair histograms. The resulting amino acid grouping schemes are presented in Tab. 4.2. Obviously, the first clustering, i.e. going from 20 to 12 amino acid symbols, runs in parallel for both schemes. Further alphabet reduction yields different results. Grouping based on histograms that include only short-range interactions ($R_{\max} = 8$ Å) are primarily due to amino acid sizes. For $R_{\max} = 50$ Å we observe reduction along the lines of charges and aromaticity with a clear separation of hydrophobic and polar amino residues, similar to results from previous studies (see Tab. 4.1). Hence, the assumption to take into account ranges of interactions is supported.

Additionally, we varied $R_{\max}$ to compare all possible ranges of interactions and derived a clustering in form of a tree. Groupings of amino acids are denoted along the branches, the leafs represent amino acid symbols. The nodal distance as measure for dissimilarity of trees is computed using TOPD [Puigbò *et al.*, 2007]. Clearly, we detect potential "structural" breaks, i.e. the classification of amino acids changes dramatically, at $R \approx$17, 27, 41, 46 Å. We also notice recurring patterns of resemblance, for example the clusterings derived at $R = 17$ Å and $R = 30$ Å represent compatible reductions. The underlying biophysical mechanism is not clear, yet. Hence, to derive simplified amino acid alphabets not only amino acid properties need to be considered, but also interaction ranges. The interaction of immediate contacts differs substantially to long-range interactions. The simplified amino acid alphabets were utilized in sections 1.3.2 and 1.3.3 for the fitting of amino acid specific interaction potentials.

### 4.2.4 Contributions

FH and SP prepared the protein data set. SP performed the analysis. KH and FH devised the study. All authors prepared the manuscript.

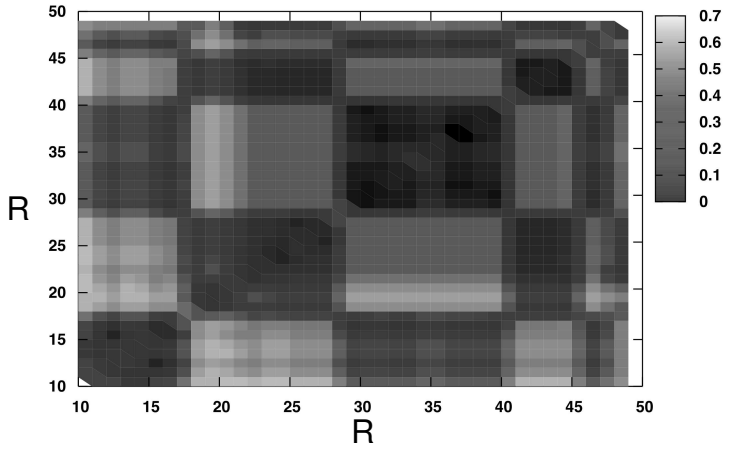



**Figure 4.5.:** Nodal distances between amino acid trees representing simplified amino acid alphabets. The trees were derived as clustering of amino acid correlation distances for varying maximum interaction ranges *R*. The nodal distance, which is a measure of dissimilarity of trees with identical leaf symbols, was computed using the TOPD package [Puigbò *et al.*, 2007]. Picture is adapted from Pape *et al.* [2010].

## 4.3 Structural model of the gas vesicle protein GvpA and analysis of GvpA mutants *in vivo*

A structural model of the gas vesicle envelope protein GvpA has been derived *in silico* by *de novo* modeling. We performed MD simulations to verify the stability of the structure. Results of this study have been published under:

Strunk, T; Hamacher, K; Hoffgaard, F; Engelhardt, H; Zillig, MD; Faist, K; Wenzel, W; Pfeifer, F (2011) Structural model of the gas vesicle protein GvpA and analysis of GvpA mutants *in vivo*. *Mol Microbiol* 81:56.

### 4.3.1 Background

Gas vesicles are protein structures that are found in a wide range of microorganisms in aquatic habitats [Walsby, 1994]. They provide the cells with buoyancy which allows them to adjust their vertical position in aquatic environment in response to light or aeration. In Haloarchaea these structures are spindle- or cylinder-shaped and stretch up to 1 μm in length and 200 nm in diameter. Fourteen genes are involved in the formation of gas vesicles in *Halobacterium salinarum*, but only eight genes have proven to be essential [Offner *et al.*, 2000]. The envelope of the gas vesicle is mainly constituted by the 8-kDa protein GvpA. The wall formed by GvpA has a hydrophobic inner surface to prevent water molecules, that may have entered the structure, from condensation. On the hydrophilic outer surface the protein GvpC is attached, presumably stabilizing the structure. The primary structure of GvpA, which is responsible for many properties of the gas vesicle, is highly conserved. Since GvpA (monomers) can hardly be dissolved without being denaturated, no crystal structure has been determined, yet. In a recent solid-state NMR study, Sivertsen *et al.* [2010] suggested a coil–$\alpha$–$\beta$-$\beta$–$\alpha$–coil fold of GvpA.

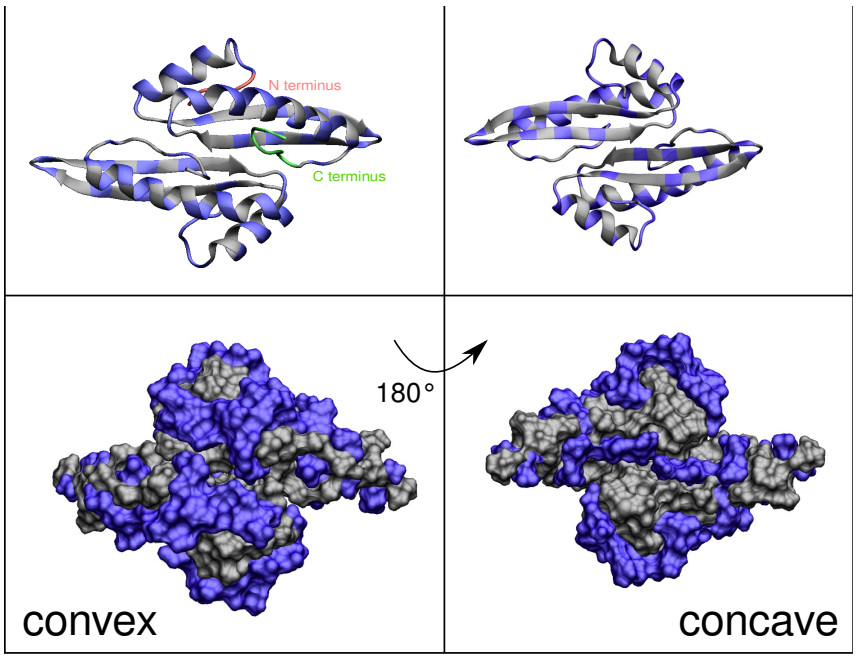



**Figure 4.6.:** Structure of GvpA presented as dimer. We show secondary structure elements and surface representation in the upper and lower graphics, respectively. Hydrophobic residues are colored gray, non-hydrophobic residue are indicated by blue color. The concave surface, i.e. $\beta$-sheets, presumably forms the inner surface of the gas vesicle wall. Pictures were generated using VMD [Humphrey *et al.*, 1996] and in part taken from Strunk *et al.* [2011].

The amino acid sequence of mcGvpA of *Haloferax mediterranei* was used to predict a tertiary structure model *in silico*. Template-based modeling approaches using 3D-Jury [Ginalski *et al.*, 2003], FUGUE [Shi *et al.*, 2001], I-TASSER [Roy *et al.*, 2010; Zhang, 2008] and SAM-T08 [Karplus *et al.*, 2003] failed to detect homology to known proteins, and, thus, yielded no results of promising quality. Therefore, a *de novo* model using ROSETTA [Bonneau *et al.*, 2002] was computed setting the predicted secondary structure elements as constraints. The obtained models were ranked according to scores derived from the all-atom free energy force field PFF02 [Verma & Wenzel, 2009].

In accordance to the solid-state NMR results of Sivertsen *et al.* [2010], the structural model of the major gas vesicle protein GvpA as shown in Fig. 4.6 contains two $\alpha$-helices separated by two antiparallel $\beta$-sheets. Infrared spectroscopy supports the presence of a significant amount of $\alpha$-helical structure other than discussed by Walsby [1994], who assumed GvpA to consist of $\beta$-sheets only. It is supposed that $\beta$-sheets form the surface of the inner gas vesicle wall. Almost all $\beta$-strand residues pointing towards the suggested interior of the gas vesicle are hydrophobic, those with side chains pointing to the gas vesicle wall are rather hydrophilic or charged (see Fig. 4.6). The proposed dimer structure connects two antiparallel GvpA monomers by contacts of their $\beta$-sheet. Hence, the inner, concave surface of the gas vesicle wall is characterized by large hydrophobic patches.

In addition, a dimer structure has been proposed and the importance of single contacts was jugded by application of ANMs [Strunk *et al.*, 2011]. The sensitivity analysis performed for an ANM model of GvpA to judge on the relevance of single contacts uncovered that the most

important interactions are contacts connecting loop regions with stable secondary structure elements. Stabilizing of secondary structures stems from multiple interactions with loops, as the largest changes in dynamics were determined if intrinsically flexible loops were "released" by deleting contacts to $\alpha$-helices or $\beta$-sheets.

Small C-terminal deletions and point mutations were introduced for identified crucial residues *in vivo* to add further evidence on the proposed structure [Strunk *et al.*, 2011]. The *in vivo* studies of $\Delta$-mutants revealed that the last 7 residues at the C-terminus are not required to form the gas vesicle wall, as gas vesicle of size and shape similar to the wild-type could be isolated. Colonies of the $\Delta$11-variant, however, were unable to float in liquid media indicating a lack of gas vesicles. Here, two of the deleted residues belong to the C-terminal helix, which suggests that an intact helix is necessary to form long and stable gas vesicles, whereas the rather unstructured part of the C-terminus is dispensable. Strunk *et al.* [2011] showed in site-directed mutagenesis studies that mutations I34M, E35A, and K60L affect the shape and size of gas vesicles. Interestingly, mutants R15A and R15K lacked any gas vesicles, as residue R15 and other charged amino acids are considered as potential binding site for a second GvpA monomer. In summary, amino acids whose contacts have been marked important for the dynamics of the gas vesicle protein GvpA by the ANM were shown to play an important role for *in vivo* forming and assembling of gas vesicles. Hence, structural features of the GvpA model were further supported.

### 4.3.2 Methods

To judge the stability of the predicted GvpA structure (see Fig. 4.6), we simulated its dynamics for 30 ns using NAMD [Phillips *et al.*, 2005] and an all-atom additive CHARMM force field [MacKerell *et al.*, 1998, 2004]. To this end, we generated two different salt concentrations in reference to habitats of halophilic organisms: 1 M KCl and 1 M NaCl + 5 M KCl.

For details see: Strunk *et al.* [2011].

### 4.3.3 Results

MD simulations of the GvpA monomer in two different salt concentrations were performed to prove stability of the predicted structure. We computed the RMSD, i.e. the average deviation of corresponding atoms of two structures, of the $C_\alpha$ atoms over time. As reference structure, we used the structural model of GvpA that was obtained by *in silico* modeling. As can be seen in Fig. 4.7, for each of the four independent simulations, the RMSD is small (<6 Å), and, thus, simulating the dynamics of the predicted model maintains structural features. Hence, evidence for validity of the modeled GvpA structure is given. Furthermore, only minor differences between both salt concentrations are observed, indicating only a small susceptibility of the structure to ion concentrations. Derived root mean square fluctuations correlate well with results from the ANM (Spearman correlation coefficient 0.68), again revealing only minor differences between both salt concentrations.

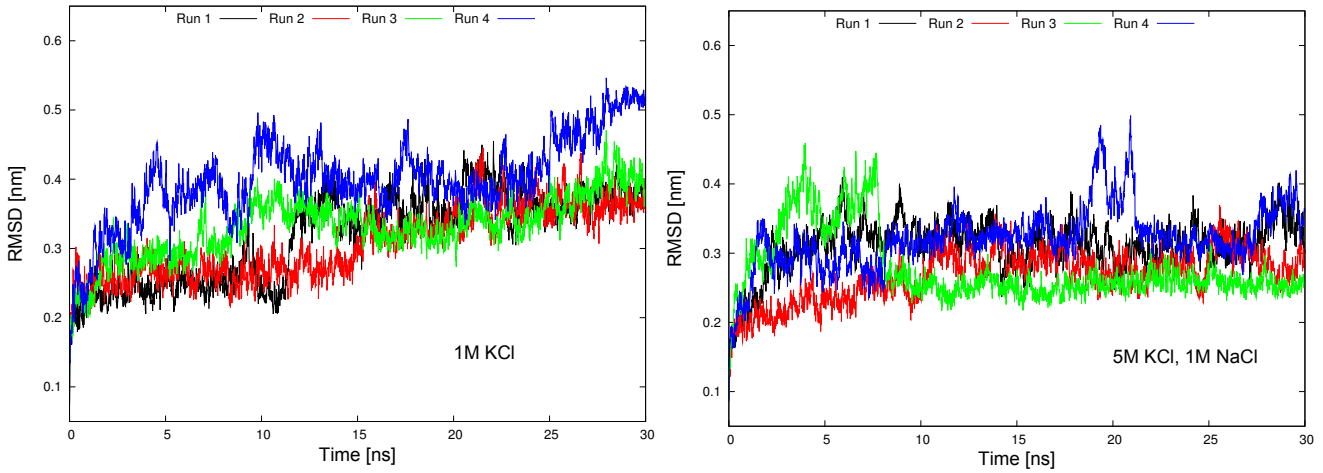



**Figure 4.7.:** Root mean square deviation (RMSD) is plotted versus simulation time. For each salt concentration (1 M KCl and 5 M KCl + 1 M NaCl) the results for four independent simulations are shown. Pictures are adapted from Strunk *et al.* [2011].

### 4.3.4 Contributions

TS, WW predicted the stuctural model of the monomer, KH the structural model of the dimer. FH performed MD simulations and KH the ENM computations. HE realized infrared spectroscopy experiments. FP, KH developed experimental design of GvpA mutants. MDZ, KF, FP carried out mutations in GvpA and microscopic analyses of mutants and isolated gas vesicles. WW, HE, KH, FP wrote the paper.

## 4.4 Structure-based, biophysical annotation of molecular coevolution of acetylcholinesterase

We use an approach that combines results obtained from coevolutionary analysis with mechanical properties to annotate crucial residues in achetylcholinesterase (AChE) that was implemented in the R package `BioPhysConnectoR` [Hoffgaard *et al.*, 2010] (see section 4.1). Coevolutionary analysis as a sequence based method is accomplished using information theory. To investigate the mechanics, we perform *in silico* experiments for a coarse-grained network model of AChE. Methods and results are described in:

Weißgraeber, S; Hoffgaard, F; Hamacher, K : Structure-based, biophysical annotation of molecular coevolution of acetylcholinesterase. *Proteins*, accepted.

### 4.4.1 Background

Acetylcholinesterase (AChE) is a ubiquitous enzyme that has been found for a range of evolutionary diverse vertebrates and invertebrates [Soreq & Seidman, 2001]. Being one of the fastest known enzymes [Lawler, 1961], it hydrolyses and inactivates the neurotransmitter acte-

lycholine. Hence, the concentration of acetylcholine in the synaptic cleft of cholinergic synapses is controlled by AChE. Acetylcholin is fundamental for mediating neurotransmission in the nervous system. Its abrupt blockade is lethal, whereas gradual loss is associated with progressive deterioration of cognitive and neuromuscular function, such as Alzheimer's disease [Wright *et al.*, 1993]. Since neurotransmission depends on the dissociation of acetylcholine from the receptor followed by its diffusion and hydrolysis, AChE has been an attractive target for the rational design of inhibitors, so-called anticholinesterase agents. The core of its active site located at the bottom of a deep, narrow gorge is a catalytic triad composed of serine, glutamate and histidine. The extremely toxic "nerve gas" sarin, for example, phosphorylates the serine residue of the active site and, thus, renders the enzyme inactive [Quinn, 1987; Taylor, 2001]. Additionally, a peripheral binding site of AChE, also referred to as peripheral anionic site (PAS), was identified as target for inhibitors. For example, the snake venom fasciculin reversibly binds to this site and prevents the acetylcholine from entering the channel [Fossier *et al.*, 1986]. Thus, the signal transduction cannot be terminated. Moreover, binding of inhibitors to PAS may induce allosteric effects [Shi *et al.*, 2002]. Owing to its variety of molecular forms (different splicing variants, monomeric vs. multimeric forms) and diverse and unexpected localizations, such as in non-cholinergic neurons, osteogenic, hematopoietic and various neoplastic cells [Soreq & Seidman, 2001], prompted the idea AChE could have non-classical functions [Appleyard, 1992], amongst others an intrinsic proteolytic activity [Small, 1990].

We provide a combined approach based on biophysical and sequence-based information to annotate features of AChE refining the understanding of its enzymatic and non-enzymatic function as well as the modes of actions of anticholinesterase agents and potential resistance mechanisms to such inhibitors. This is implemented in the `R` package `BioPhysConnectoR` [Hoffgaard *et al.*, 2010] (see section 4.1). The rationale behind this study is to relate the sequence space, which is subject to the mutational operator, to the biophysical realm, which is subject to selective pressure acting on phenotypes. Point mutations on single amino acids impose selective pressure on their interaction partners leading to compensatory mutations that may be beneficial to restore previous conditions. Such coevolutionary behavior of amino acids is investigated using the well-established information theoretical MI as discussed in section 3. Biophysical ramifications of mutations accomplished using ANMs [Atilgan *et al.*, 2001] that are used to analyze structural-dynamical aspects of proteins (see section 1.1).

### 4.4.2 Methods

Our study is based on the crystal structure which is deposited in the PDB [Berman *et al.*, 2000] with code 1EA5 [Dvir *et al.*, 2002]. The amino acid sequence was used as query string for BLAST (Basic Local Alignment Search Tool) [Altschul *et al.*, 1990] to search for further AChE sequences. After applying different filter criteria, we computed an MSA using `clustalw2` [Larkin *et al.*, 2007]. Alignment parameters are partly modified to improve the quality of the resulting alignment. For our analysis, we considered alignment positions with less than 40% gaps (arbitrary threshold) only. Hence, we are concerned with more than 200 sequences, which has been shown to be sufficient to yield reasonable results [Weil *et al.*, 2009]. The information-theoretical measure MI is computed as defined in Eq. 3.1. To account for a base level of correlation within the alignment, we applied a null model by reiterated, independent shuffling of each column

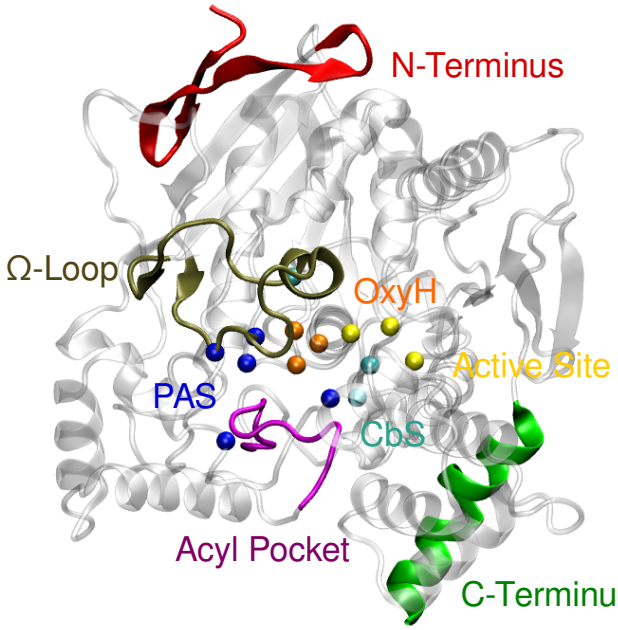



**Figure 4.8.:** Crystal structure of Acetylcholinesterase, deposited in the Protein Data Bank [Berman *et al.*, 2000] with code 3EA5. Parts of the protein that have been identified to be important for structural or functional properties [Shi *et al.*, 2002; Sussman *et al.*, 1991; Zhang *et al.*, 2002] are highlighted. Picture is adapted from Weißgraeber *et al.* [accepted].

[Weil *et al.*, 2009] and recomputating of the respective MI matrix as implemented in the R package `BioPhysConnectoR` [Hoffgaard *et al.*, 2010] (see section 4.1). Based on these values, we determined the Z-score (see Eq. 3.2) that measures the significance of the obtained MI value with respect to the null model. We consider $Z_{ij} > 4$ significant [Gloor *et al.*, 2005].

Furthermore, we used the crystal structure of AChE to derive an ANM (see section 1.1) and computed the respective correlation matrix $C$, which is obtained by normalizing the covariance matrix. Residues whose $C_\alpha$ atoms are closer than $r_c = 13$ Å were defined to be in contact. Each non-covalent contact of the ANM was artificially deleted, one at a time, i.e. the strength of the corresponding interaction is set to zero (see section 4.1). For the mutated protein, we computed the correlation matrix $C'$. The impact of the contact deletion was quantified by determining the FN of $C$ and $C'$ as defined in Eq. 1.26.

As we already discussed in section 2.2, effects on local dynamics may get lost by summing up the complete correlation matrix. To this end, we computed the FN for submatrices that correspond to defined protein regions according to literature [Shi *et al.*, 2002; Sussman *et al.*, 1991; Zhang *et al.*, 2002] (see Fig. 4.8): active site, Ω-loop, peripheral anionic site (PAS), oxyanion hole (OxyH), acyl pocket, choline binding site (CbS) as well as N- and C-terminus. The oxyanion hole, that is located adjacent to the catalytic triad and both acyl and choline binding site, stabilizes the negatively charged carbonly oxygen during catalysis [Warshel *et al.*, 1989]. The Ω-loop at the rim of the channel plays an important role in inhibitor binding [Shi *et al.*, 2002], and PAS establishes contact with the substrate [Bourne *et al.*, 2003].

We combined the results from both approaches to relate potential coevolutionary dependencies to mechanical properties of the AChE.

For further details see: Weißgraeber *et al* [accepted].

### 4.4.3 Results

For each protein region, we determined the top ten contacts whose artificial deletion had major influence on its dynamics, that is yielding the largest FN values. We correlated these contacts with corresponding MI values to assess the amount of coevolution between the respective residues. Annotating residue pairs by both biophysical and sequence-based means results in four distinct classes:

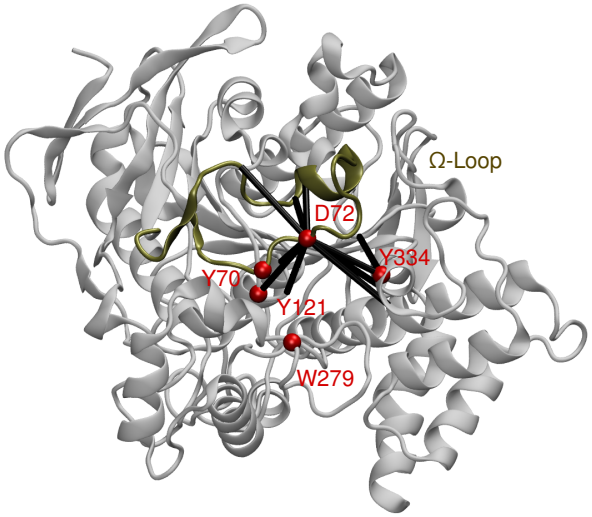
1. Positions with both high MI and FN values are subject to the same coevolutionary pressure and important for structural integrity.

2. Non-interacting residues showing low MI values are regarded independent of each other.

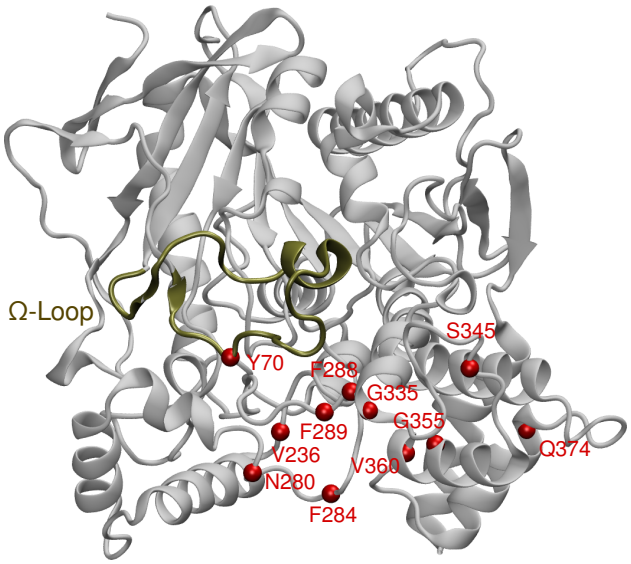
3. Biophysically relevant contact with only little coevolution may be due to a high level of conservation of at least one of the respective residues, that is only little to no variation at that site. Another possible explanation is the independence from side chains, such as electrostatic backbone interactions on the basis of the dipole character of the double bonded resonance form of peptide bonds. Such residue pairs are no subject to coevolution even though they may be crucial for mechanical properties.

4. Amino acid pairs featuring high MI values, but without any mechanical dependence may arise from the underlying physical model, as other than harmonic interactions are not taken into account. The considered interaction may also be crucial for structural elements, which are not subject to our analysis. Furthermore, spatial proximity causes coevolution of two residues, but covalent contacts are not "released" in our analysis. Hence, the impact of artificial deletion of covalent bonds is not correlated with coevolution of the respective residues.

By investigating the top ten FN values for each defined protein region, we were able to identify the loop forming the acyl pocket which holds the acyl group during catalysis as coevolutionary hot spot. Its structural integrity is mainly sustained by contact within the loops itself. All top ten contacts feature considerably high MI values ranging from 0.80 to 1.35 bit, as comparison the average MI is about 0.61 bit and the maximum is found at a value of 2.17 bit. Similarly, coevolution ensures the preservation of the structure of the choline binding pocket, which is necessary to bind the choline group during catalysis. It is stabilized by contacts to the oxyanion hole and surrounding residues. The respective residue pairs show MI values above average. Furthermore, we observe that residue D72 is by far the most important residue maintaining the dynamics of the peripheral anionic site, supported by contacts to the active site, the oxyanion hole and the choline binding site (see Fig. 4.9(a)). The contact between Y334 of PAS and H440 of the catalytic triad yields high FN values which is in agreement with Epstein *et al.* [1979]. According to this study, the catalysis is inhibited by binding of the snake venom fasciculin not only by sterical occlusion of the gorge entrance but also by induction of allosteric changes in the active site conformation. For this contact, we were not able to detect coevolution, as H440 is highly conserved.




**(a)** Using ANMs, we determined the top ten interactions to maintain structure and function of the peripheral anionic site (PAS). Those interactions are indicated by black cylinders in the structure of AChE. The Ω-loop, ranging from F78 to C94 is highlighted as orientation.




**(b)** A group of coevolving residues next to the rim of the gorge that was determined by the sequence-based MI approach is shown. The Ω-loop, ranging from F78 to C94 is highlighted as orientation.

**Figure 4.9.:** Sites identified both from the biophysical (a) and sequence-based (b) approach are labeled in the structure of AChE. Pictures are adapted from Weißgraeber *et al.* [accepted].

Since coevolution of residues that do not share crucial physical contacts is an interesting issue on its own, we identified residue pairs with high MI values. We performed our analysis on a monomer of AChE, but AChE is capable to form homotetramers, as well. Hence, sites with contacts between subunits may not appear in the ANM approach, if the corresponding residues are not in contact within a single subunit. We determined contacting sites using a tetrameric crystal structure (PDB code 1C2O [Bourne *et al.*, 1999]). Indeed, we found inter-subunit contacts, that are not in contact within the subunit, with high MI values, indicating coevolutionary connections due to tetramerization. In a recent study, Gloor *et al.* [2005] made the observation that they could distinguish two classes of coevolving residues: Positions that coevolve with one or two other positions, and, furthermore, clusters of coevolving residues that are presumably part of common functional units. We identified three residues (Y70, Y121, W279) that exhibit MI values above average with Y334, although their spatial distance is beyond the cutoff range. The respective amino acids are a part of PAS, which is crucial for substrate binding. Hence, we suggest the existence of coevolution due to the belonging of the amino acids to the same functional unit. Additionally, we identified a region of AChE that has not been annotated, yet. Five of the top ten highest MI values were detected for pairs including Y70. A detailed analysis of Y70 and coevolving residues revealed a site (see Fig. 4.9(b)) located closely to the opening of the channel, implying potential involvements of this region with binding partners or in non-hydrolytic functions of AChE that were discussed previously [Soreq & Seidman, 2001].

We provided a detailed annotation of structurally and functionally crucial residues of AChE by the application of a combined biophysical and sequence-based protocol. Moreover, we identified functional units within the protein that have not been described, yet. More biophysical charac-

teristics, such as charge, hydrophobicity, need to be included to further enhance the proposed protocol.

### 4.4.4 Contributions

SW performed the analysis. KH, FH and SW devised the study and wrote the manuscript.

# Bibliography


Alizadeh, F (1995) Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM J Optim* 5:13.

Altschul, SF; Gish, W; Miller, W; Myers, EW; Lipman, DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403.

Amdahl, G (1967) Validity of the single processor approach to achieving large-scale computing capabilities. *AFIPS Conference Proceedings* 30:483–485.

Appleyard, ME (1992) Secreted acetylcholinesterase: non-classical aspects of a classical enzyme. *Trends Neurosci* 15:485.

Atilgan, A; Durrell, S; Jernigan, R; Demirel, M; Keskin, O; Bahar, I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* 80:505.

Bahar, I; Atilgan, AR; Demirel, MC; Erman, B (1998) Vibrational dynamics of folded proteins: Significance of slow and fast motion in relation to function and stability. *Phys Rev Lett* 80:2733.

Bahar, I; Atilgan, AR; Erman, B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design* 2:173.

Bahar, I; Erman, B; Jernigan, R; Atilgan, A; Covell, D (1999) Collective motions in HIV-1 reverse transcriptase: Examination of flexibility and enzyme function. *J Mol Biol* 285:1023.

Bahar, I; Jernigan, R (1999) Cooperative fluctuations and subunit communication in tryptophan synthase. *Biochemistry* 38:3478.

Batey; Rambo; Doudna (1999) Tertiary motifs in RNA structure and folding. *Angew Chem Int Ed Engl* 38:2326.

Ben-Tal, A; Nemirovski, A (2002) Robust optimization – methodology and applications. *Math Program, Ser B* 92:453.

Benson, D; Karsch-Mizrachi, I; Lipman, D; Ostell, J; Sayers, E (2011) Genbank. *Nucl Acids Res (Database issue)* 39 (Database issue):D32.

Berman, HM; Westbrook, J; Feng, Z; Gilliland, G; Bhat, TN; Weissig, H; Shindyalov, IN; Bourne, PE (2000) The protein data bank. *Nucleic Acids Res* 28:235.

Bertsimas, D; Sim, M (2003) Robust discrete optimization and network flows. *Math Program, Ser B* 98:49–71.

Biswas, P; Lian, TC; Wang, TC; Ye, Y (2006) Semidefinite programming based algorithms for sensor network localization. *ACM Transactions on Sensor Networks* 2:188.

Boba, P; Weil, P; Hoffgaard, F; Hamacher, K (2010) Co-evolution in HIV enzymes. *BIOINFORMATICS2010*, edited by A Fred; J Filipe; H Gamboa.

Bonneau, R; Strauss, CEM; Rohl, CA; Chivian, D; Bradley, P; Malmström, L; Robertson, T; Baker, D (2002) De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* 322:65.

Bourne, Y; Grassi, J; Bougis, PE; Marchot, P (1999) Conformational flexibility of the acetylcholinesterase tetramer suggested by X-ray crystallography. *J Biol Chem* 274:30370.

Bourne, Y; Taylor, P; Radić, Z; Marchot, P (2003) Structural insights into ligand interactions at the acetylcholinesterase peripheral anionic site. *EMBO J* 22:1.

Boyd, S; Vandenberghe, L (2004) *Convex Optimization* (Cambridge University Press, New York).

Brooks, B; Karplus, M (1983) Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci USA* 80:6571.

Brooks, B; Karplus, M (1985) Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc Natl Acad Sci U S A* 82:4995.

Brooks, BB; Janezic, D; Karplus, M (1995) Harmonic analysis of large systems I. Methodology. *J Comp Chem* 12:1522.

Brown, BM; Sauer, RT (1999) Tolerance of arc repressor to multiple-alanine substitutions. *Proc Natl Acad Sci U S A* 96:1983.

Bryngelson, J; Onuchic, J; Socci, N; Wolynes, P (1995) Funnels, pathways and the energy landscape of protein folding: a synthesis. *Proteins-Struct Func and Genetics* 21:167.

Cannata, N; Toppo, S; Romualdi, C; Valle, G (2002) Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. *Bioinformatics* 18:1102.

Caporaso, JG; Smit, S; Easton, BC; Hunter, L; Huttley, GA; Knight, R (2008) Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. *BMC Evol Biol* 8:327.

Cech, TR (2000) Structural biology. The ribosome is a ribozyme. *Science* 289:878.

Cech, TR; Zaug, AJ; Grabowski, PJ (1981) *In vitro* splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* 27:487.

Černý, V (1985) Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *J of Opt Theory and Applications* 45:41.

Chatelain, FC; Gazzarrini, S; Fujiwara, Y; Arrigoni, C; Domigan, C; Ferrara, G; Pantoja, C; Thiel, G; Moroni, A; Minor, DL (2009) Selection of inhibitor-resistant viral potassium channels identifies a selectivity filter site that affects barium and amantadine block. *PLoS One* 4:e7496.

Chen, L; Perlina, A; Lee, CJ (2004) Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J Virol* 78:3722.

Chow, CS; Bogdan, FM (1997) A structural basis for RNA–ligand interactions. *Chem Rev* 97:1489.

Chu, JW; Voth, GA (2007) Coarse-grained free energy functions for studying protein conformational changes: a double-well network model. *Biophys J* 93:3860.

Cieplak, M; Holter, N; Maritan, A; Banavar, J (2001) Amino acid classes and the protein folding problem. *J Chem Phys* 114:1420.

Cline, MS; Karplus, K; Lathrop, RH; Smith, TF; Rogers, RG; Haussler, D (2002) Information-theoretic dissection of pairwise contact potentials. *Proteins* 49:7.

Cordero-Morales, JF; Cuello, LG; Zhao, Y; Jogini, V; Cortes, DM; Roux, B; Perozo, E (2006) Molecular determinants of gating at the potassium-channel selectivity filter. *Nature Struct & Mol Biol* 13:311.

Crick, F (1970) Central dogma of molecular biology. *Nature* 227:561.

Cui, Q; Bahar, I (eds.) (2006) *Normal Mode Analysis: Theory and Application to Biological and Chemical Systems* (Chapman & Hall / CRC).

Cunningham, BC; Wells, JA (1989) High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* 244:1081.

Daggett, V; Levitt, M (1992) Molecular dynamics simulations of helix denaturation. *J Mol Biol* 223:1121.

de la Peña, M; Gago, S; Flores, R (2003) Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity. *EMBO J* 22:5561.

de la Peña, M; García-Robles, I (2010a) Intronic hammerhead ribozymes are ultraconserved in the human genome. *EMBO Rep* 11:711.

de la Peña, M; García-Robles, I (2010b) Ubiquitous presence of the hammerhead ribozyme motif along the tree of life. *RNA* 16:1943.

Demirel, MC; Atilgan, AR; Jernigan, RL; Erman, B; Bahar, I (1998) Identification of kinetically hot residues in proteins. *Protein Sci* 7:2522.

Dill, K; Chan, H (1997) From Levinthal to pathways to funnels. *Nat Struct Biol* 4:10.

Doruker, P; Jernigan, RL; Bahar, I (2002) Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J Comput Chem* 23:119.

Doudna, JA; Cech, TR (2002) The chemical repertoire of natural ribozymes. *Nature* 418:222.

Doyle, DA; Cabral, JM; Pfuetzner, RA; Kuo, A; Gulbis, JM; Cohen, SL; Chait, BT; MacKinnon, R (1998) The structure of the potassium channel: molecular basis of $K^+$ conduction and selectivity. *Science* 280:69.

Dvir, H; Jiang, HL; Wong, DM; Harel, M; Chetrit, M; He, XC; Jin, GY; Yu, GL; Tang, XC; Silman, I; *et al.* (2002) X-ray structures of Torpedo californica acetylcholinesterase complexed with (+)-huperzine A and (-)-huperzine B: structural evidence for an active site rearrangement. *Biochemistry* 41:10810.

Eddy, SR (2001) Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* 2:919.

Eom, K; Baek, SC; Ahn, JH; Na, S (2007) Coarse-graining of protein structures for the normal mode studies. *J Comput Chem* 28:1400.

Epstein, DJ; Berman, HA; Taylor, P (1979) Ligand-induced conformational changes in acetylcholinesterase investigated with fluorescent phosphonates. *Biochemistry* 18:4749.

Epstein, LM; Gall, JG (1987) Self-cleaving transcripts of satellite DNA from the newt. *Cell* 48:535.

Erman, B (2006) The Gaussian network model: precise prediction of residue fluctuations and application to binding problems. *Biophys J* 91:3589.

Erman, B; Dill, K (2000) Gaussian model of protein folding. *J Chem Phys* 112:1050.

Excoffier, L; Slatkin, M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921.

Fatakia, SN; Constanzi, S; Chow, CC (2009) Computing highly correlated positions using mutual information and graph theory for G protein-coupled receptors. *PLoS ONE* 4:e4681.

Felsenstein, J (1981) Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol* 17:368.

Fischer, WB; Sansom, MSP (2002) Viral ion channels: structure and function. *Biochim Biophys Acta* 1561:27.

Flicek, P; Aken, BL; Ballester, B; Beal, K; Bragin, E; Brent, S; Chen, Y; Clapham, P; Coates, G; Fairley, S; *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res* 38:D557.

Forster, AC; Symons, RH (1987) Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell* 49:211.

Fossier, P; Baux, G; Tauc, L (1986) Fasciculin II, a protein inhibitor of acetylcholinesterase, tested on central synapses of Aplysia. *Cell Mol Neurobiol* 6:221.

Frohns, F; Käsmann, A; Kramer, D; Schäfer, B; Mehmel, M; Kang, M; van Etten, JL; Gazzarrini, S; Moroni, A; Thiel, G (2006) Potassium ion channels of chlorella viruses cause rapid depolarization of host cells during infection. *J Virol* 80:2437.

Fujisawa, K; Fukuda, M; Kobayashi, K; Kojima, M; Nakata, K; Nakata, M; Yamashita, M (2008) *SDPA (SemiDefinite Programming Algorithm) and SDPA-GMP User's Manual – Version 7.1.2.*

Galassi, M (2009) *GNU Scientific Library Reference Manual.* 3rd edn.

Gazzarrini, S; Kang, M; van Etten, JL; Tayefeh, S; Kast, SM; DiFrancesco, D; Thiel, G; Moroni, A (2004) Long distance interactions within the potassium channel pore are revealed by molecular diversity of viral proteins. *J Biol Chem* 279:28443.

Gazzarrini, S; Severino, M; Lombardi, M; Morandi, M; DiFrancesco, D; van Etten, JL; Thiel, G; Moroni, A (2003) The viral potassium channel Kcv: structural and functional features. *FEBS Lett* 552:12.

Gebhardt, M; Hoffgaard, F; Hamacher, K; Kast, SM; Moroni, A; Thiel, G (2011) Membrane anchoring and interaction between transmembrane domains are crucial for $K^+$ channel function. *J Biol Chem* 286:11299.

Gentleman, RC; Carey, VJ; Bates, DM; Bolstad, B; Dettling, M; Dudoit, S; Ellis, B; Gautier, L; Ge, Y; Gentry, J; *et al.* (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* 5:R80.

Ginalski, K; Elofsson, A; Fischer, D; Rychlewski, L (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19:1015.

Gloor, GB; Martin, LC; Wahl, LM; Dunn, SD (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 44:7156.

Goldenberg, DP; Creighton, TE (1984) Folding pathway of a circular form of bovine pancreatic trypsin inhibitor. *J Mol Biol* 179:527.

Gräf, S; Przybilski, R; Steger, G; Hammann, C (2005) A database search for hammerhead ribozyme motifs. *Biochem Soc Trans* 33:477.

Grant, BJ; Rodrigues, APC; ElSawy, KM; McCammon, JA; Caves, LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics* 22:2695.

Guerrier-Takada, C; Gardiner, K; Marsh, T; Pace, N; Altman, S (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* 35:849.

Haliloglu, T; Bahar, I; Erman, B (1997) Gaussian dynamics of folded proteins. *Phys Rev Lett* 79:3090.

Halle, B (2002) Flexibility and packing in proteins. *Proc Natl Acad Sci U S A* 99:1274.

Hamacher, K (2006) Adaptation in stochastic tunneling global optimization of complex potential energy landscapes. *Europhys Lett* 74:944.

Hamacher, K (2007a) Adaptive extremal optimization by detrended fluctuation analysis. *J Comp Phys* 227:1500.

Hamacher, K (2007b) Energy landscape paving as a perfect optimization approach under detrended fluctuation analysis. *Physica A* 378:307.

Hamacher, K (2007c) Information theoretical measures to analyze trajectories in rational molecular design. *J Comp Chem* 28:2576.

Hamacher, K (2008) Relating sequence evolution of HIV1-protease to its underlying molecular mechanics. *Gene* 422:30.

Hamacher, K (2010) Efficient perturbation analysis of elastic network models – application to acetylcholinesterase of *T. californica*. *J Comp Phys* 229:7309.

Hamacher, K; McCammon, JA (2006) Computing the amino acid specificity of fluctuations in biomolecular systems. *J Chem Theory Comput* 2:873.

Hamacher, K; Wenzel, W (1999) Scaling behaviour of stochastic minimization algorithms in a perfect funnel landscape. *Phys Rev E* 59:938.

Heginbotham, L; Lu, Z; Abramson, T; MacKinnon, R (1994) Mutations in the $K^+$ channel signature sequence. *Biophys J* 66:1061.

Hertel, B; Tayefeh, S; Kloss, T; Hewing, J; Gebhardt, M; Baumeister, D; Moroni, A; Thiel, G; Kast, SM (2010) Salt bridges in the miniature viral channel Kcv are important for function. *Eur Biophys J* 39:1057.

Hille, B (2001) *Ionic Channels of Excitable Membranes* (Sinauer Associates).

Hinsen, K (2009) Physical arguments for distance-weighted interactions in elastic network models for proteins. *Proc Natl Acad Sci U S A* 106:E128.

Hinsen, K; Petrescu, AJ; Dellerue, S; Bellissent-Funel, MC; Kneller, GR (2000) Harmonicity in slow protein dynamics. *Chem Phys* 261:25.

Hoffgaard, F; Weil, P; Hamacher, K (2010) BioPhysConnectoR: Connecting sequence information and biophysical models. *BMC Bioinformatics* 11:199.

Humphrey, W; Dalke, A; Schulten, K (1996) VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* 14:33.

Ivanova, MI; Sievers, SA; Sawaya, MR; Wall, JS; Eisenberg, D (2009) Molecular basis for insulin fibril assembly. *Proc Natl Acad Sci U S A* 106:18990.

Janezic, D; Brooks, BB (1995) Harmonic analysis of large systems II. Comparison of different protein models. *J Comp Chem* 16:1543.

Janezic, D; Venable, RM; Brooks, BB (1995) Harmonic analysis of large systems III. Comparison with molecular dynamics. *J Comp Chem* 16:1554.

Jeong, JI; Jang, Y; Kim, MK (2006) A connection rule for $\alpha$-carbon coarse-grained elastic network models using chemical bond information. *J Mol Graph Model* 24:296.

Jimenez, RM; Delwart, E; Lupták, A (2011) Structure-based search reveals hammerhead ribozymes in the human microbiome. *J Biol Chem* 286:7737.

Jorgensen, WL; Maxwell, DS; Tirado-Rives, J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* 118:11225.

Kaminski, GA; Friesner, RA; Tirado-Rives, J; Jorgensen, WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 105:6474.

Kang, M; Graves, M; Mehmel, M; Moroni, A; Gazzarrini, S; Thiel, G; Gurnon, JR; van Etten, JL (2004) Genetic diversity in chlorella viruses flanking Kcv, a gene that encodes a potassium ion channel protein. *Virology* 326:150.

Karger, D; Motwani, R; Sudan, M (1998) Derandomizing semidefinite programming based approximation algorithms. *Journal of the ACM* 45:246.

Karplus, K; Karchin, R; Draper, J; Casper, J; Mandel-Gutfreund, Y; Diekhans, M; Hughey, R (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins* 53 Suppl 6:491.

Keskin, O; Bahar, I; Badretdinov, AY; Ptitsyn, OB; Jernigan, RL (1998) Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. *Protein Sci* 7:2578.

Khvorova, A; Lescoute, A; Westhof, E; Jayasena, SD (2003) Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. *Nat Struct Biol* 10:708.

Kidera, A; Gō, N (1992) Normal mode refinement: Crystallographic refinement of protein dynamic structures. *J Mol Biol* 225:457.

Korber, BTM; Farber, RM; Wolpert, DH; Lapedes, AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *PNAS* 90:7176.

Kullback, S; Leibler, RA (1951) On information and sufficiency. *The Annals of Mathematical Statistics* 22:79.

Kundu, S; Melton, JS; Sorensen, DC; Phillips, GN (2002) Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys J* 83:723.

Kuo, A; Gulbis, JM; Antcliff, JF; Rahman, T; Lowe, ED; Zimmer, J; Cuthbertson, J; Ashcroft, FM; Ezaki, T; Doyle, DA (2003) Crystal structure of the potassium channel kirbac1.1 in the closed state. *Science* 300:1922.

Kuriyan, J; Weis, WI (1991) Rigid protein motion as a model for crystallographic temperature factors. *Proc Natl Acad Sci U S A* 88:2773.

Kuwada, T; Hasegawa, T; Takagi, T; Sato, I; Shishikura, F (2010) pH-dependent structural changes in haemoglobin component V from the midge larva Propsilocerus akamusi (Orthocladiinae, Diptera). *Acta Crystallogr D Biol Crystallogr* 66:258.

Larkin, MA; Blackshields, G; Brown, NP; Chenna, R; McGettigan, PA; McWilliam, H; Valentin, F; Wallace, IM; Wilm, A; Lopez, R; *et al.* (2007) ClustalW and ClustalX version 2. *Bioinformatics* 23:2947.

Lawler, HC (1961) Turnover time of acetylcholinesterase. *J Biol Chem* 236:2296.

Leopold, P; Montal, M; Onuchic, J (1992) Protein folding funnels: a kinetic approach to the sequence–structure relationship. *PNAS* 89:8721.

Levinthal, C (1968) Are there pathways for protein folding? *J Chim Phys* 65:44.

Levitt, M; Sander, C; Stern, PS (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J Mol Biol* 181:423.

Levy, Y; Jortner, J; Becker, OM (2001) Solvent effects on the energy landscapes and folding kinetics of polyalanine. *Proc Natl Acad Sci U S A* 98:2188.

Li, G; Cui, Q (2002) A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to $Ca^{2+}$-ATPase. *Biophys J* 83:2457.

Li, T; Yang, Y; Canessa, CM (2009) Interaction of the aromatics Tyr-72/Trp-288 in the interface of the extracellular and transmembrane domains is essential for proton gating of acid-sensing ion channels. *J Biol Chem* 284:4689.

Li, Z; Scheraga, HA (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci U S A* 84:6611.

Lin, CP; Huang, SW; Lai, YL; Yen, SC; Shih, CH; Lu, CH; Huang, CC; Hwang, JK (2008) Deriving protein dynamical properties from weighted protein contact number. *Proteins* 72:929.

Lindahl, ER (2008) Molecular dynamics simulations. *Methods Mol Biol* 443:3.

Lundbæk, JA (2008) Lipid bilayer-mediated regulation of ion channel function by amphiphilic drugs. *J Gen Physiol* 131:421.

Lyman, E; Pfaendtner, J; Voth, GA (2008) Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophys J* 95:4183.

MacKerell, A, Jr; Bashford, D; Bellott, M; Dunbrack, R, Jr; Evanseck, JD; Field, MJ; Fischer, S; Gao, J; Guo, H; Ha, S; *et al.* (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586.

MacKerell, AD, Jr.; Feig, M; Brooks, CL (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25:1400.

MacKinnon, R (2003) Potassium channels. *FEBS Lett* 555:62.

Makhatadze, GI; Kim, KS; Woodward, C; Privalov, PL (1993) Thermodynamics of BPTI folding. *Protein Sci* 2:2028.

Maragakis, P; Karplus, M (2005) Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J Mol Biol* 352:807.

Marqusee, S; Robbins, VH; Baldwin, RL (1989) Unusually stable helix formation in short alanine-based peptides. *Proc Natl Acad Sci U S A* 86:5286.

Martick, M; Scott, WG (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* 126:309.

Mazziotti, DA (2004) Realization of quantum chemistry without wave functions through first-order semidefinite programming. *Phys Rev Lett* 93:213001.

McCammon, JA; Gelin, BR; Karplus, M (1977) Dynamics of folded proteins. *Nature* 267:585.

McGaughey, GB; Gagné, M; Rappé, AK (1998) $\pi$-stacking interactions. alive and well in proteins. *J Biol Chem* 273:15458.

Mehmel, M; Rothermel, M; Meckel, T; van Etten, JL; Moroni, A; Thiel, G (2003) Possible function for virus encoded $K^+$ channel Kcv in the replication of chlorella virus PBCV-1. *FEBS Lett* 552:7.

Metropolis, N; Rosenbluth, AW; Rosenbluth, MN; Teller, AH (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087.

Metropolis, N; Ulam, S (1949) The Monte Carlo method. *J Am Stat Assoc* 44:335.

Micheletti, C; Banavar, JR; Maritan, A (2001) Conformations of proteins in equilibrium. *Physical Review Letters* 87:088102.

Micheletti, C; Carloni, P; Maritan, A (2004) Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. *Proteins* 55:635.

Ming, D; Wall, ME (2005) Quantifying allosteric effects in proteins. *Proteins* 59:697.

Miyazawa, S; Jernigan, RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256:623.

Moore, EH (1920) On the reciprocal of the general algebraic matrix. *Bull Am Math Soc* 26:394.

Moritsugu, K; Smith, JC (2007) Coarse-grained biomolecular simulation with REACH: realistic extension algorithm via covariance Hessian. *Biophys J* 93:3460.

Moroni, A; Viscomi, C; Sangiorgio, V; Pagliuca, C; Meckel, T; Horvath, F; Gazzarrini, S; Valbuzzi, P; van Etten, JL; DiFrancesco, D; *et al.* (2002) The short N-terminus is required for functional expression of the virus-encoded miniature $K^+$ channel Kcv. *FEBS Lett* 530:65.

Moses, E; Hinz, HJ (1983) Basic pancreatic trypsin inhibitor has unusual thermodynamic stability parameters. *J Mol Biol* 170:765.

Munson, M; O'Brien, R; Sturtevant, JM; Regan, L (1994) Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci* 3:2015.

Murshudov, GN; Vagin, AA; Dodson, EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53:240.

Murtagh, F (1985) *Multidimensional clustering algorithms* (Physica-Verlag).

Murzin, AG; Brenner, SE; Hubbard, T; Chothia, C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536.

Nojima, H; Takeda-Shitaka, M; Kurihara, Y; Adachi, M; Yoneda, S; Kamiya, K; Umeyama, H (2002) Dynamic characteristics of a peptide-binding groove of human HLA-A2 class I MHC molecules: Normal mode analysis of the antigen peptide-class I MHC complex. *Chem Pharm Bull* 50:1209.

Novomestky, F (2008) *matrixcalc.*

Offner, S; Hofacker, A; Wanner, G; Pfeifer, F (2000) Eight of fourteen gvp genes are sufficient for formation of gas vesicles in halophilic archaea. *J Bacteriol* 182:4328.

Pagliuca, C; Goetze, TA; Wagner, R; Thiel, G; Moroni, A; Parcej, D (2007) Molecular properties of Kcv, a virus encoded $K^+$ channel. *Biochemistry* 46:1079.

Pape, S; Hoffgaard, F; Hamacher, K (2010) Distance-dependent classification of amino acids by information theory. *Proteins* 78:2322.

Penrose, R (1955) A generalized inverse for matrices. *Proc Cambr Philos Soc* 51:406.

Perreault, J; Weinberg, Z; Roth, A; Popescu, O; Chartrand, P; Ferbeyre, G; Breaker, RR (2011) Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS Comput Biol* 7:e1002031.

Phillips, JC; Braun, R; Wang, W; Gumbart, J; Tajkhorshid, E; Villa, E; Chipot, C; Skeel, RD; Kalé, L; Schulten, K (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26:1781.

Pinto, LH; Holsinger, LJ; Lamb, RA (1992) Influenza virus M2 protein has ion channel activity. *Cell* 69:517.

Pley, HW; Flaherty, KM; McKay, DB (1994) Three-dimensional structure of a hammerhead ribozyme. *Nature* 372:68.

Plugge, B; Gazzarrini, S; Nelson, M; Cerana, R; Etten, JLV; Derst, C; DiFrancesco, D; Moroni, A; Thiel, G (2000) A potassium channel protein encoded by chlorella virus PBCV-1. *Science* 287:1641.

Press, WH; Vetterling, WT; Teukolsky, SA; Flannery, BP (1992) *Numerical Recipes in C.* (Cambridge University Press, Cambridge).

Prody, GA; Bakos, JT; Buzayan, JM; Schneider, IR; Bruening, G (1986) Autolytic processing of dimeric plant virus satellite RNA. *Science* 231:1577.

Przybilski, R; Gräf, S; Lescoute, A; Nellen, W; Westhof, E; Steger, G; Hammann, C (2005) Functional hammerhead ribozymes naturally encoded in the genome of *Arabidopsis thaliana*. *Plant Cell* 17:1877.

Przybilski, R; Hammann, C (2006) The hammerhead ribozyme structure brought in line. *Chembiochem* 7:1641.

Puigbò, P; Garcia-Vallvé, S; McInerney, JO (2007) TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics* 23:1556.

Quinn, DM (1987) Acetylcholinesterase: Enzyme structure, reaction dynamics, and virtual transition states. *Chem Rev* 87:955.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

Rapedius, M; Fowler, PW; Shang, L; Sansom, MSP; Tucker, SJ; Baukrowitz, T (2007) H bonding at the helix-bundle crossing controls gating in Kir potassium channels. *Neuron* 55:602.

Reichert, J; Jabs, A; Slickers, P; Sühnel, J (2000) The IMB Jena Image Library of biological macromolecules. *Nucleic Acids Res* 28:246.

Reiling, KK; Endres, NF; Dauber, DS; Craik, CS; Stroud, RM (2002) Anisotropic dynamics of the JE-2147-HIV protease complex: drug resistance and thermodynamic binding mode examined in a 1.09 Å structure. *Biochemistry* 41:4582.

Riccardi, D; Cui, Q; Phillips, GN (2009) Application of elastic network models to proteins in the crystalline state. *Biophys J* 96:464.

Riddle, DS; Santiago, JV; Bray-Hall, ST; Doshi, N; Grantcharova, VP; Yi, Q; Baker, D (1997) Functional rapidly folding proteins from simplified amino acid sequences. *Nat Struct Biol* 4:805.

Rocheleau, L; Pelchat, M (2006) The Subviral RNA Database: a toolbox for viroids, the hepatitis delta virus and satellite RNAs research. *BMC Microbiol* 6:24.

Rodríguez, H; Angulo, I; de Las Rivas, B; Campillo, N; Páez, JA; Muñoz, R; Mancheño, JM (2010) p–Coumaric acid decarboxylase from Lactobacillus plantarum: structural insights into the active site and decarboxylation catalytic mechanism. *Proteins* 78:1662.

Rodríguez de la Vega, RC; Possani, LD (2004) Current views on scorpion toxins specific for $K^+$-channels. *Toxicon* 43:865.

Rojas, NR; Kamtekar, S; Simons, CT; McLean, JE; Vogel, KM; Spiro, TG; Farid, RS; Hecht, MH (1997) *De novo* heme proteins from designed combinatorial libraries. *Protein Sci* 6:2512.

Roy, A; Kucukural, A; Zhang, Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5:725.

Saldanha, R; Mohr, G; Belfort, M; Lambowitz, AM (1993) Group I and group II introns. *FASEB J* 7:15.

Scherer, CW; Hol, CWJ (2006) Matrix sum-of-squares relaxations for robust semi-definite programs. *Math Program, Ser B* 107:189.

Seehafer, C; Kalweit, A; Steger, G; Gräf, S; Hammann, C (2011) From alpaca to zebrafish: hammerhead ribozymes wherever you look. *RNA* 17:21.

Shannon, CE (1948) A mathematical theory of communication. *The Bell System Tech J* 27:623.

Shepherd, JC (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc Natl Acad Sci U S A* 78:1596.

Shepherd, SJ; Beggs, CB; Jones, S (2007) Amino acid partitioning using a Fiedler vector model. *Eur Biophys J* 37:105.

Shi, J; Blundell, TL; Mizuguchi, K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 310:243.

Shi, J; Radić, Z; Taylor, P (2002) Inhibitors of different structure induce distinguishing conformations in the omega loop, Cys69-Cys96, of mouse acetylcholinesterase. *J Biol Chem* 277:43301.

Shih, CH; Huang, SW; Yen, SC; Lai, YL; Yu, SH; Hwang, JK (2007) A simple way to compute protein dynamics without a mechanical model. *Proteins* 68:34.

Shrivastava, IH; Bahar, I (2006) Common mechanism of pore opening shared by five different potassium channels. *Biophys J* 90:3929.

Sivertsen, AC; Bayro, MJ; Belenky, M; Griffin, RG; Herzfeld, J (2010) Solid-state NMR characterization of gas vesicle structure. *Biophys J* 99:1932.

Small, DH (1990) Non-cholinergic actions of acetylcholinesterases: proteases regulating cell growth and development? *Trends Biochem Sci* 15:213.

Soheilifard, R; Makarov, DE; Rodin, GJ (2008) Critical evaluation of simple network models of protein dynamics and their comparison with crystallographic B-factors. *Phys Biol* 5:026008.

Soman, KV; Karimi, A; Case, DA (1991) Unfolding of an $\alpha$-helix in water. *Biopolymers* 31:1351.

Soreq, H; Seidman, S (2001) Acetylcholinesterase–new roles for an old actor. *Nat Rev Neurosci* 2:294.

Srivastav, A; Wolf, K (1998) Finding dense subgraphs with semidefinite programming. *Approximation Algorithms for Combinatiorial Optimization, LNCS* 1444:181.

Stillinger, FH; Weber, TA (1984) Packing structures and transitions in liquids and solids. *Science* 225:983.

Strunk, T; Hamacher, K; Hoffgaard, F; Engelhardt, H; Zillig, MD; Faist, K; Wenzel, W; Pfeifer, F (2011) Structural model of the gas vesicle protein GvpA and analysis of GvpA mutants *in vivo*. *Mol Microbiol* 81:56.

Sussman, JL; Harel, M; Frolow, F; Oefner, C; Goldman, A; Toker, L; Silman, I (1991) Atomic structure of acetylcholinesterase from *Torpedo californica*: a prototypic acetylcholine-binding protein. *Science* 253:872.

Syeda, R; Holden, MA; Hwang, WL; Bayley, H (2008) Screening blockers against a potassium channel with a droplet interface bilayer array. *J Am Chem Soc* 130:15543.

Taketomi, H; Ueda, Y; Gō, N (1975) Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int J Pept Protein Res* 7:445.

Tama, F; Sanejouand, YH (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng* 14:1.

Tan, Q; Shim, JW; Gu, LQ (2010) Separation of heteromeric potassium channel Kcv towards probing subunit composition-regulated ion permeation and gating. *FEBS Lett* 584:1602.

Tayefeh, S; Kloss, T; Kreim, M; Gebhardt, M; Baumeister, D; Hertel, B; Richter, C; Schwalbe, H; Moroni, A; Thiel, G; *et al.* (2009) Model development for the viral Kcv potassium channel. *Biophys J* 96:485.

Tayefeh, S; Kloss, T; Thiel, G; Hertel, B; Moroni, A; Kast, SM (2007) Molecular dynamics simulation of the cytosolic mouth in Kcv-type potassium channels. *Biochemistry* 46:4826.

Taylor, P (2001) *The Pharmacological Basis of Therapeutics* chap. Anticholinesterase agents (McGraw-Hill) (239).

Thiel, G; Baumeister, D; Schroeder, I; Kast, SM; van Etten, JL; Moroni, A (2011) Minimal art: or why small viral $K^+$ channels are good tools for understanding basic structure and function relations. *Biochim Biophys Acta* 1808:580.

Tierney, L; Rossini, AJ; Li, N (2009) Snow: A parallel computing framework for the R system. *Int J of Parallel Computing* 37:78.

Tinoco, I, Jr; Bustamante, C (1999) How RNA folds. *J Mol Biol* 293:271.

Tirion, MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* 77:1905.

Trylska, J; Tozzini, V; McCammon, JA (2005) Exploring global motions and correlations in the ribosome. *Biophys J* 89:1455.

Urbanek, S (2009) *multicore.*

van der Spoel, D; Lindahl, E; Hess, B; Groenhof, G; Mark, AE; Berendsen, HJC (2005) GROMACS: fast, flexible, and free. *J Comput Chem* 26:1701.

van Etten, JL (2003) Unusual life style of giant chlorella viruses. *Annu Rev Genet* 37:153.

van Vlijmen, HWT; Karplus, M (1999) Analysis of calculated normal modes of a set of native and partially unfolded proteins. *J Phys Chem B* 103:3009.

Vandenberghe, L; Boyd, S (1996) Semidefinite programming. *SIAM Rev* 38:49.

Verma, A; Wenzel, W (2009) A free-energy approach for all-atom protein simulation. *Biophys J* 96:3483.

Vila, JA; Ripoll, DR; Scheraga, HA (2000) Physical reasons for the unusual $\alpha$–helix stabilization afforded by charged or neutral polar residues in alanine-rich peptides. *Proc Natl Acad Sci* 97:13075.

Walsby, AE (1994) Gas vesicles. *Microbiol Rev* 58:94.

Wang, J; Wang, W (1999) A computational approach to simplifying the protein folding alphabet. *Nat Struct Biol* 6:1033.

Wang, K; Xie, S; Sun, B (2011) Viral proteins function as ion channels. *Biochim Biophys Acta* 1808:510.

Warshel, A; Naray-Szabo, G; Sussman, F; Hwang, JK (1989) How do serine proteases really work? *Biochemistry* 28:3629.

Waterhouse, AM; Procter, JB; Martin, DMA; Clamp, M; Barton, GJ (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189.

Weaver, LH; Grütter, MG; Matthews, BW (1995) The refined structures of goose lysozyme and its complex with a bound trisaccharide show that the "goose-type" lysozymes lack a catalytic aspartate residue. *J Mol Biol* 245:54.

Weber, W; Hünenberger, PH; McCammon, JA (2000) Molecular dynamics simulations of a polyalanine octapeptide under Ewald boundary conditions: Influence of artificial periodicity on peptide conformation. *J Phys Chem* 104:3668.

Weil, P; Hoffgaard, F; Hamacher, K (2009) Estimating sufficient statistics in co-evolutionary analysis by mutual information. *Comp Biol Chem* 33:440.

Weinberger, KQ; Saul, LK (2006) Unsupervised learning of image manifolds by semidefinite programming. *Int J Comp Vision* 70:77.

Wenzel, W; Hamacher, K (1999) Stochastic tunneling approach for global minimization of complex potential energy landscapes. *Phys Rev Letters* 82:3003.

Wilson, TJ; Lilley, DMJ (2009) Biochemistry. The evolution of ribozyme chemistry. *Science* 323:1436.

Wlodawer, A; Nachman, J; Gilliland, GL; Gallagher, W; Woodward, C (1987) Structure of form III crystals of bovine pancreatic trypsin inhibitor. *J Mol Biol* 198:469.

Wlodawer, A; Walter, J; Huber, R; Sjölin, L (1984) Structure of bovine pancreatic trypsin inhibitor. *J Mol Biol* 180:301.

Wolynes, PG (1997) As simple as can be? *Nat Struct Biol* 4:871.

Wright, CI; Geula, C; Mesulam, MM (1993) Neuroglial cholinesterases in the normal brain and in Alzheimer's disease: relationship to plaques, tangles, and patterns of selective vulnerability. *Ann Neurol* 34:373.

Wüthrich, K; Wider, G; Wagner, G; Braun, W (1982) Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J Mol Biol* 155:311.

Yang, L; Song, G; Carriquiry, A; Jernigan, RL (2008) Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure* 16:321.

Yang, L; Song, G; Jernigan, RL (2009) Protein elastic network models and the ranges of cooperativity. *Proc Natl Acad Sci U S A* 106:12347.

Zhang, Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:40.

Zhang, Y; Kua, J; McCammon, JA (2002) Role of the catalytic triad and oxyanion hole in acetylcholinesterase catalysis: an ab initio QM/MM study. *J Am Chem Soc* 124:10572.

Zimm, BH; Bragg, JK (1959) Theory of the phase transition between helix and random coil in polypeptide chains. *J of Chem Phys* 31:526.

# Appendices

# A Data sets

This section lists data sets that have been used in projects of this thesis. For protein sets A and B the PDB codes [Berman *et al.*, 2000] are specified. For PA simulations the structures that were used as references are given.

## A.1 Protein Set A

1A55 1A7G 1ABA 1ABE 1AMP 1AQB 1ARU 1BYB 1C53 1CD2 1CHD 1CNV 1COO 1COT 1CTF 1CTT 1D3Z 1DHR 1ECA 1EDE 1ENH 1ES6 1ESC 1EZM 1FKJ 1FLP 1FXD 1GCA 1GOF 1HFC 1HYP 1IAE 1LID 1MBA 1MJC 1MML 1NPK 1NXB 1OYC 1PAZ 1PEA 1PHP 1PHT 1PIQ 1POA 1PTQ 1R69 1RCB 1RIS 1SBP 1TCA 1TDE 1TEN 1TIG 1UBQ 1XNB 2CMD 2CPL 2CTC 2EBN 2END 2I1B 2LIV 2MCM 2MNR 2RN2 2SAS 2SN3 3DFR 3IL8 3LZM 451C 4FGF 5P21

## A.2 Protein Set B

153L 16VP 1A0I 1A17 1A1I 1A1V 1A1Z 1A26 1A32 1A62 1A6F 1A6Q 1A7J 1A8D 1A8Y 1ABV 1ACC 1AD2 1ADT 1AEP 1AF7 1AFP 1AH7 1AIE 1AIL 1AIN 1AJ2 1AK0 1AKO 1ALG 1ALY 1AMM 1AMX 1AOL 1AOY 1AQT 1ARB 1ASS 1AT0 1ATA 1AUA 1AZO 1B0U 1B0X 1B63 1B74 1B9P 1B9W 1BAM 1BBY 1BCO 1BDO 1BF4 1BF5 1BFG 1BG6 1BGF 1BHU 1BIA 1BIF 1BJT 1BKF 1BL0 1BLE 1BM8 1BOL 1BOO 1BQG 1BRT 1BT3 1BUO 1BUP 1BUU 1BX7 1BY2 1BYR 1C0A 1C1K 1C25 1C3D 1C3G 1C3P 1C4K 1C4X 1C7K 1CA1 1CB8 1CBF 1CBH 1CBY 1CDW 1CDZ 1CEX 1CFE 1CFR 1CHD 1CHU 1CII 1CIS 1CKQ 1CKT 1CQQ 1CSH 1CT5 1CTF 1CTQ 1CUK 1CYO 1CYX 1CZS 1D2S 1D2T 1D5T 1D8B 1D8C 1DAB 1DBH 1DCF 1DCQ 1DD5 1DD9 1DF4 1DFU 1DHN 1DHS 1DI6 1DIV 1DKC 1DL2 1DLC 1DLJ 1DLW 1DMG 1DMU 1DNV 1DOV 1DP7 1DPQ 1DQ3 1DQG 1DS1 1DT4 1DT9 1DU1 1DUR 1DUS 1DVO 1DVP 1DZF 1E0M 1E2S 1E3H 1E3O 1E43 1E4C 1E4M 1E5K 1E7U 1EB7 1ECR 1EDQ 1EFD 1EGS 1EH3 1EI5 1EJ0 1EJE 1EKG 1EKR 1EM2 1EOV 1ES5 1ES9 1ESD 1ETL 1EU1 1EW0 1EWN 1EYB 1EYH 1F00 1F0N 1F2V 1F32 1F3L 1F44 1F53 1F5N 1F7S 1F7U 1FC6 1FCQ 1FEH 1FI4 1FID 1FJJ 1FNC 1FP1 1FRB 1FSF 1FUS 1FVG 1FYE 1FYV 1G12 1G3W 1G5M 1G7S 1G8F 1G8S 1GA8 1GBG 1GK7 1GKU 1GKY 1GM5 1GNT 1GNY 1GOF 1GPR 1GQE 1GRJ 1GS5 1GSA 1GTR 1GU3 1GV9 1GVP 1GYV 1GZ8 1H05 1H2C 1H5P 1H6H 1H6I 1H6L 1H6O 1H70 1H8L 1H99 1HB6 1HCB 1HCR 1HCZ 1HF8 1HFC 1HH2 1HK8 1HLL 1HLV 1HMS 1HN0 1HO8 1HP1 1HP9 1HQ0 1HQ1 1HQV 1HS6 1HUF 1HUS 1HW7 1HXN 1HXX 1HY9 1HYP 1HZT 1I1I 1I27 1I3J 1I52 1I60 1I8A 1I9G 1IAE 1IB8 1IDK 1IFR 1IG8 1IGR 1IIR 1IJ5 1IKO 1IMJ 1IO1 1IOW 1IPA 1IR6 1ISP 1IUQ 1IWL 1IXK 1IZM 1J09 1J1L 1J1T 1J23 1J3A 1J3E 1J5X 1J5Y 1J6O 1J6U 1J77 1J8R 1J98 1J9B 1JB3 1JDW 1JEO 1JFX 1JGS 1JHG 1JHJ 1JHN 1JHS 1JID 1JL1 1JMM 1JNI 1JPC 1JPU 1JQN 1JR7 1JRL 1JSG 1JSX 1JU3 1JU8 1JWQ 1JYH 1K04 1K12 1K24 1K4N 1K4T 1K6K 1K7C 1K8T 1KBL 1KCL 1KFT 1KHC 1KID 1KKH 1KL9 1KLO 1KLX 1KNB 1KON 1KP6 1KR7 1KS9 1KT7 1KWG 1KWI 1KYH 1KZF 1L2H 1L2L 1L2P 1L3K 1L5O 1L6P 1L7Y 1L8Q 1L9V 1LC0 1LCI 1LG7 1LI4 1LJ8 1LL2 1LLA 1LLP 1LMI 1LML 1LN4 1LNS 1LO7 1LOU 1LOX 1LRI 1LRV 1LRZ 1LSL 1LV3 1LVA 1LWB 1M1H 1M2K 1M5I 1M65 1M66 1M6E 1M73 1M8Z 1M9S 1MA4 1MAI 1MG4 1MGP 1MHN 1MHU 1MJC 1MJN 1MK0 1MKY 1ML8 1ML9 1MLA 1MML 1MN3 1MNN 1MRK 1MSC 1MSK 1MSW 1MUG 1MUS 1MVL 1MWP 1MXA 1MZB 1N1T 1N4K

1N5U 1N67 1N81 1N93 1N9P 1N9U 1NC5 1NEP 1NFN 1NFP 1NG6 1NH1 1NH8 1NI3 1NI5 1NI9 1NIF 1NIJ 1NKD 1NKG 1NNX 1NOS 1NOX 1NQK 1NTH 1NW3 1NZA 1NZE 1O1Z 1O22 1O3U 1O4W 1O59 1O88 1O9G 1OBR 1OCY 1ODH 1OFC 1OFL 1OGL 1OGQ 1OH4 1OHL 1OI1 1OKC 1OKG 1OKS 1OPC 1OUO 1OXE 1OXJ 1OY8 1OYG 1OYW 1OYZ 1OZ9 1P1M 1P2Z 1P90 1P97 1P9I 1PB5 1PBE 1PDA 1PDO 1PDR 1PEF 1PEN 1PFO 1PFV 1PG1 1PG6 1PGS 1PHR 1PHZ 1PI1 1PIE 1PIN 1PJ5 1PKM 1PKP 1PMI 1PNE 1POA 1POC 1PP7 1PSF 1PSW 1PTQ 1PUC 1PW4 1PXE 1PZT 1PZW 1Q0H 1Q0R 1Q1H 1Q2B 1Q4R 1Q5Z 1Q7H 1Q8B 1Q8C 1Q8D 1Q92 1QAZ 1QBA 1QCS 1QCZ 1QDD 1QFM 1QG8 1QGO 1QGV 1QHD 1QHX 1QJ4 1QJP 1QLM 1QME 1QOY 1QPG 1QR0 1QRE 1QSA 1QTO 1QTW 1QW2 1QWG 1QWY 1QYI 1QYS 1R0U 1R1H 1R3D 1R4V 1R4X 1R6F 1R75 1R7J 1R89 1R8E 1R8I 1R9F 1RA0 1RA6 1RA9 1RC9 1RE9 1REP 1RH4 1RHS 1RI5 1RI6 1RIE 1RKD 1RL6 1RLH 1RLJ 1RLR 1RMG 1RO2 1ROC 1RP4 1RQB 1RR7 1RRO 1RRQ 1RRZ 1RTQ 1RU4 1RW2 1RW6 1RWR 1RWU 1RYQ 1RZ4 1RZY 1S21 1S2W 1S2X 1S68 1S7C 1S7I 1S7Z 1S9Z 1SAY 1SDO 1SFE 1SFP 1SG7 1SIG 1SJW 1SKN 1SQG 1SQH 1SQW 1SR8 1SRA 1SSK 1SU8 1SUM 1SUR 1SUU 1SVB 1SWX 1SZI 1T1D 1T27 1T2S 1T3J 1T3T 1T5J 1T6A 1T6C 1T95 1TA0 1TCA 1TCH 1TD6 1TDJ 1TF5 1TFF 1TFR 1TG7 1TGJ 1THQ 1TIF 1TIG 1TJ1 1TJN 1TJX 1TKE 1TL2 1TOH 1TOL 1TOP 1TOV 1TP6 1TQH 1TS9 1TT8 1TTU 1TUH 1TUL 1TUW 1TWU 1TXL 1TYX 1U02 1U04 1U14 1U2C 1U4G 1U7G 1U7L 1U84 1U94 1UAE 1UBY 1UDB 1UDS 1UDX 1UEK 1UFA 1UG9 1UJ8 1UKF 1ULY 1UMG 1UMH 1UOY 1URU 1UTG 1UW1 1UWF 1UWV 1UX6 1UXO 1UXX 1UXY 1UYN 1V04 1V0A 1V0D 1V2X 1V2Z 1V4A 1V77 1V9M 1VBV 1VCC 1VFY 1VHE 1VHH 1VHU 1VI7 1VJW 1VK1 1VK5 1VK6 1VK9 1VKB 1VKW 1VLI 1VMG 1VMH 1VNS 1VPQ 1VPR 1VPT 1VRM 1VSR 1W0P 1W1O 1W5D 1W66 1W8M 1WBA 1WC9 1WCD 1WD3 1WER 1WFX 1WHI 1WHO 1WHZ 1WJ9 1WNA 1WPA 1WUB 1WV3 1WV8 1WVK 1WY6 1WZU 1X38 1X82 1X9N 1XAB 1XAK 1XCL 1XD7 1XEO 1XER 1XFI 1XG8 1XHB 1XKS 1XM9 1XMX 1XNB 1XO8 1XOV 1XQ8 1XTO 1XTP 1XW8 1Y08 1Y0N 1Y6I 1Y8C 1Y9Q 1YB3 1YDL 1YDX 1YFQ 1YFU 1YGE 1YGS 1YI9 1YLN 1YQG 1YQY 1YS5 1YT3 1YU0 1YU5 1YUB 1YVR 1YWF 1Z0P 1Z21 1Z67 1ZAR 1ZAT 1ZBP 1ZD0 1ZDY 1ZFO 1ZHV 1ZHX 1ZIN 1ZOD 1ZPW 1ZRN 1ZTN 1ZX3 2A0B 2A4H 2A8E 2AAK 2ABK 2ACT 2ACY 2AE9 2AP3 2APL 2AQA 2ASR 2ATZ 2AXO 2B0J 2B4W 2B5H 2BAI 2BDE 2BDT 2BIB 2BL7 2BNH 2BOP 2BSC 2BV3 2BYO 2BZ1 2C1I 2C5S 2C6J 2C9A 2CC6 2CFQ 2CN1 2CUL 2CW9 2CX1 2CXA 2CXF 2D2S 2D5B 2D5U 2DAP 2DDH 2DP9 2DPK 2DPM 2DRI 2EIF 2END 2ENG 2ERL 2ES9 2ET1 2ETD 2EWH 2F09 2FB7 2FDI 2FFM 2FFT 2FGC 2FGG 2FI0 2FM9 2FPN 2FQ3 2FRN 2FSJ 2FYG 2G7O 2G9D 2GC6 2GHR 2GKE 2GQV 2GS5 2GTI 2GTV 2H85 2HBB 2HBJ 2HGS 2HK6 2HKJ 2HQV 2HUJ 2HVM 2HXM 2IBA 2ICS 2IGD 2ILK 2JAK 2JEK 2JF2 2JQA 2JV3 2JW6 2KFZ 2LIS 2MCM 2NLY 2NML 2NR9 2NXC 2O0Q 2OGQ 2PIA 2PII 2PIL 2PK8 2PTD 2PTH 2PUB 2Q4M 2SAK 2SLI 2TPT 2TS1 3BTA 3CGW 3CLA 3COX 3GRS 3IL8 3VUB 4AIG 4BCL 5CSM 5EAU 6XIA 7ACN 7FD1

# B Abbreviations

## Amino acid codes

| | | | | | |
|---|---|---|---|---|---|
| A | ALA | Alanine | M | MET | Methionine |
| C | CYS | Cysteine | N | ASN | Asparagine |
| D | ASP | Aspartic Acid | P | PRO | Proline |
| E | GLU | Glutamic Acid | Q | GLN | Glutamine |
| F | PHE | Phenylalanine | R | ARG | Arginine |
| G | GLY | Glycine | S | SER | Serine |
| H | HIS | Histidine | T | THR | Threonine |
| I | ILE | Isoleucine | V | VAL | Valine |
| K | LYS | Lysine | W | TRP | Tryptophan |
| L | LEU | Leucine | Y | TYR | Tyrosine |

## Nucleotide codes

| | | | |
|---|---|---|---|
| A | Adenine | T | Thymine (DNA) |
| C | Cytosine | U | Uracil (RNA) |
| G | Guanine | | |

## Further Abbreviations

| | |
|---|---|
| AChE | Acetylcholinesterase |
| ANM | Anisotropic Network Model |
| *asShihB* | amino acid specific *ShihB* |
| BLAST | Basic Local Alignment Search Tool |
| BPTI | Bovine Pancreatic Trypsin Inhibitor |
| CbS | Choline Binding Site |
| DNA | Deoxyribonucleic acid |
| eANM | extended ANM |
| ENM | Elastic Network Model |
| FN | Frobenius Norm |
| GNM | Gaussian Network Model |
| GSL | Gnu Scientific Library |
| HIV | Human Immunodeficiency Virus |
| KE | Keskin *et al.* (interaction parameters) |
| MAPE | Maximum *a posteriori* Estimation |
| MC | Monte Carlo (algorithm) |

## Further Abbreviations ... continued

| | |
|---|---|
| MD | Molecular Dynamics |
| MI | Mutual Information |
| MJ | Miyazawa-Jernigan (interaction parameters) |
| MLE | Maximum Likelihood Estimation |
| NMA | Normal Mode Analysis |
| NMR | Nuclear Magnetic Resonance (spectroscopy) |
| OxyH | Oxyanion Hole |
| PA | Polyalanine |
| PAS | Peripheral Anionic Site |
| PES | Potential Energy Surface |
| PDB | Protein Data Bank |
| pfENM | parameter-free ENM |
| *psShihB* | protein specific *ShihB* |
| RMSD | Root Mean Square Deviation |
| RMSF | Root Mean Square Fluctuation |
| RNA | Ribonucleic acid |
| SCOP | Structural Classification of Proteins |
| SCPCP | Self-Consistent Pair Contact Probability |
| SDP | Semidefinite Programming |
| *ShihB* | B factor model according to Shih *et al.* [2007] |
| STUN | Stochastic Tunneling |
| SVD | Singular Value Decomposition |
| TM | Transmembrane Domain |
| WCN | Weighted Contact Number (model) |

# Curriculum Vitae

## Personal

| | |
|---|---|
| Name | Franziska Martina Hoffgaard |
| Date of Birth | 30.04.1981 |
| Place of Birth | Leipzig |

## Career

| | |
|---|---|
| 09/1987 – 08/1992 | Ernst-Schneller Schule Leipzig |
| 09/1992 – 08/1999 | Wilhelm-Ostwald Gymnasium Leipzig |
| 09/1999 – 02/2002 | (shortened) apprenticeship as land surveyor at Städtisches Vermessungsamt Leipzig |
| 03/2002 – 08/2002 | land surveyor at Städtisches Vermessungsamt Leipzig |
| 10/2002 – 01/2008 | study of bioinformatics at Martin-Luther University Halle-Wittenberg |
| 02/2008 – 08/2011 | Phd at TU Darmstadt |

## Publications


Weil, P; Hoffgaard, F; Hamacher, K (2009) Estimating sufficient statistics in co-evolutionary analysis by mutual information. *Comp Biol Chem.* 33:440.

Boba P; Weil, P; Hoffgaard, F; Hamacher, K (2010) Co-evolution in HIV enzymes. *Proc. of BIOINFORMATICS 2010*, A. Fred, J. Filipe, H. Gamboa (eds.), p. 39.

Pape, S; Hoffgaard, F; Hamacher, K (2010) Distance-dependent classification of amino acids by information theory. *Proteins* 78:2322.

Hoffgaard, F; Weil, P; Hamacher, K (2010) BioPhysConnectoR: Connecting sequence information and biophysical models. *BMC Bioinformatics* 11:199.

Gebhardt, M; Hoffgaard, F; Hamacher, K; Kast, SM; Moroni, A; Thiel, G (2011) Membrane anchoring and interaction between transmembrane domains is crucial for $K^+$ channel function. *J Biol Chem* 286:11299.

Strunk, T; Hamacher, K; Hoffgaard, F; Engelhardt, H; Zillig, MD; Faist, K; Wenzel, W; Pfeifer, F (2011) Structural model of the gas vesicle protein GvpA and analysis of GvpA mutants *in vivo*. *Mol Microbiol* 81:56.

## Publications ... continued

Weißgraeber, S; Hoffgaard, F; Hamacher, K (2011) Structure-based, biophysical annotation of molecular coevolution of acetylcholinesterase. *Proteins*, accepted.

Wächter, M; Hamacher, K; Hoffgaard, F; Widmer, S; Goesele, M (2011) Is Your Permutation Algorithm Unbiased for $n \neq 2^m$? *9th International Conference on Parallel Processing and Applied Mathematics (PPAM 2011)*, accepted.

# Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbst angefertigt habe. Sämtliche aus fremden Quellen direkt oder indirekt übernommene Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und noch nicht veröffentlicht.

Darmstadt, 30.08.2011

Franziska Hoffgaard

# ... und zu guter Letzt

All jenen, die mich stets unterstützt haben, mir Rat gaben und mit mir Sachverhalte diskutierten, egal ob es sich um Wissenschaft oder die korrekte Verwendung von Artikeln handelte, die meine Arbeit stets kritisch aber wohlwollend betrachteten, die Kuchen gebacken und Nudeln gekocht haben, mit denen ich kalte Tage frierend und warme Tage schwitzend verbringen durfte, mit denen ich viele angenehme und auch anstrengende Stunden verlebte, die mir im Leben das Wichtigste sind, die schon immer für mich da waren, die mir Mut machten oder mich auch einfach ins kalte Wasser geschubst haben, die meinen Körper und Geist stets herausforderten, die mir Ruhe in turbulenten Zeiten gaben, die mir Lieder, Sprachen und auch Dialekte näherbrachten und zum Teil Ohrwürmer einpflanzten, die mit mir Rätsel jedweder Art lösten, die mit mir Höhen und Tiefen durchlebten, die verständliche, unverständliche oder auch immer wieder dieselben Witze machten ... all jenen möchte ich sagen:

Danke!