Complexes Theory

Abstract

This document is showing the theory behind complexes. It contains a summary of the literature we cite in the paper. All major equations and derivations ,if appropriate, are given.

Contents

The complexes model is a hierarchical coarse-grained (CG) protein model that has been implemented with a Monte-Carlo engine to generate structures of protein complexes [1]. The model is commonly refereed to as the Kim-Hummer (KH) model in the literature. Since the original publication, the forcefield has been used in a large variety of applications. Those include the study the binding kinetics of the HIV-1 capsid proteins [2, 3], the kinetic behavior of proteins in crowded environments [4–7], folding of knotted protein [8–13], enhancing the structural resolution of experiments [14–18], protein-protein interactions [19–21], protein design [22], docking [23], multi-enzyme complexes [24–27]. The model has also been extended to simulate intrinsically disordered proteins and protein phase separation [7].

The complexes model is implemented in a C++14 program Complexes++ and a Python tool pycomplexes . Complexes++ implements the Monte-Carlo engine and pycomplexes is a helper library and command line interface (CLI) tool to setup simulations and visualize them, see Figure 1. The Monte-Carlo integrator accepts input files in the CPLX format and configuration files for simulations. The split enforces that a well defined file format exists that uniquely defines a simulation. We have decided to use the YAML standard [28] for the CPLX files. A library to write YAML files exists for many programming languages allowing easier integration into existing workflow without forcing a specific programming language.

The complexes model describes proteins and large macromolecular structures at three levels of coarse-graining. The first level is a bead. Beads are interaction sites that are used to evaluate potentials and represent a single amino acid, centered on the C_α atoms. The second level is a domain. Domains are collections of beads that define how the bead positions are propagated in a simulation. The complexes model has rigid and flexible

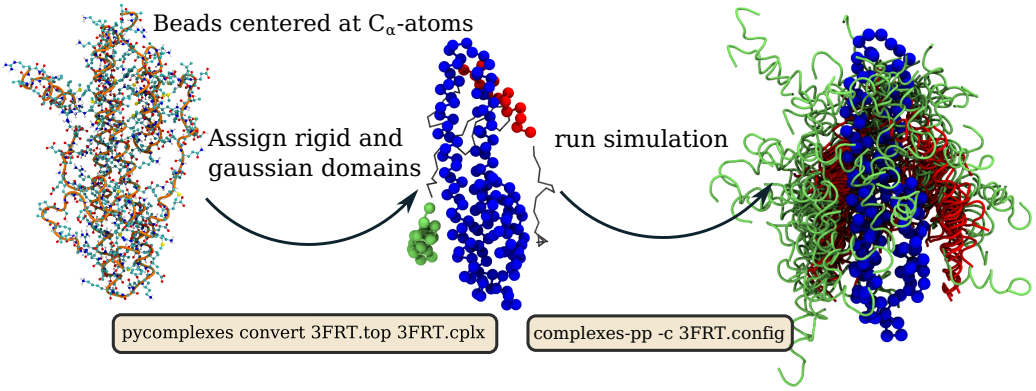

Figure 1: Example use case how to go from a single known structure to an ensemble of structures using complexes. To prepare the simulation domain types have to be assigned to amino acids and a CPLX file has to be generated.

domains. The last level is called a topology. It is a collection of connected domains. Topologies are useful to develop efficient sampling algorithms for simulations with multiple complexes. In this thesis, the general expression of the forcefield will be referred to as the complexes model, when specific values for forcefield parameters are given we refer to them as the KH model. In the following, the bead model and pair potential for different beads and different domains will be explained.

Chapter 1

Beads

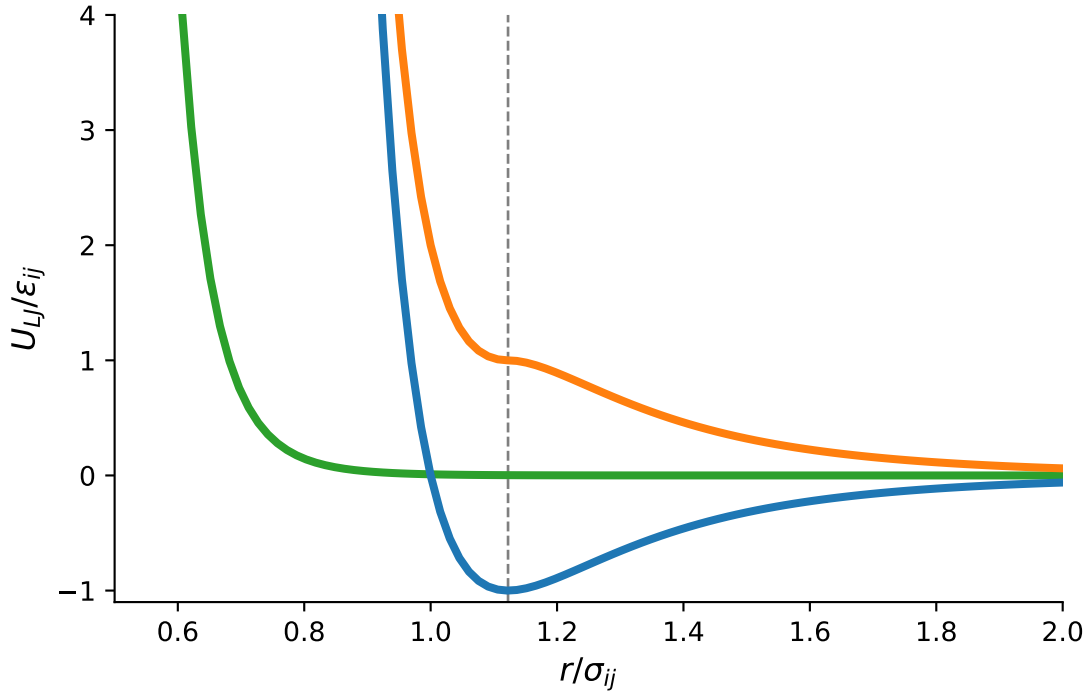

Figure 1.1: Modified Lennard Jones potential, eq 1.1, used in the complexes model in reduced units. The attractive branch $\epsilon_{ij} < 0$ is shown in blue, the repulsive part $\epsilon_{ij} > 0$ is shown in orange and the branch for $\epsilon_{ij} = 0$ in green. The gray dashed line shows the minima at $2^{1/6}$ of the attractive branch.

Beads are the interaction sites at which the force field and additional restraint potentials are evaluated. In the complexes model [1] amino acids are modeled as single beads centered on the C_α atoms. As energy function, a Lennard-Jones like potential is used for effective interactions of native and non-native contacts and a Coulomb term with an exponential screening term for the electrostatics. The potential energies are by convention calculated in units of $k_B T$, with $T = 300\,\mathrm{K}$ as the reference temperature.

The Lennard-Jones like potential U_{LJ}, between beads i and j, consists of four different

branches to model attractive and repulsive interactions

$$U_{LJ}(r,\sigma_{ij},\epsilon_{ij}) = \begin{cases} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] & \text{if } \epsilon_{ij} < 0 \\ 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] + 2\epsilon_{ij} & \text{if } \epsilon_{ij} > 0 \text{ and } r < 2^{1/6}\sigma_{ij} \\ -4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] & \text{if } \epsilon_{ij} > 0 \text{ and } r > 2^{1/6}\sigma_{ij} \\ .01\left(\frac{\sigma_{ij}}{r}\right)^{12} & \text{if } \epsilon_{ij} = 0, \end{cases} \tag{1.1}$$

with r the distance between the beads, the bead type pair parameters σ_{ij} and ϵ_{ij} for the contact distance and interaction energy, respectively. For $\epsilon_{ij} < 0$ this is the standard Lennard-Jones potential, see Figure 1.1. For $\epsilon_{ij} > 0$ this potential is purely repulsive, see Figure 1.1. In case of $\epsilon_{ij} = 0$ the potential is a hard wall slightly shorter than the Lennard-Jones minimum of $2^{1/6}\sigma_{ij}$ to avoid overlaps if additional potentials are attractive and have singularities at $r = 0$, for example electrostatic potentials. At contact $r = \sigma_{ij}$ this potential gives equal contributions from attractive and repulsive pairs with opposite sign. The parameters ϵ_{ij} are derived from the knowledge-based statistical contact potentials e_{ij} by Miyazawa and Jernigan (MJ) [29]. The MJ contact potentials have to be scaled to account for the added electrostatic interactions and the preference of residue-residue to residue-solvent interactions has to be balanced. In the complexes model this is done by scaling with a parameter λ for the electrostatic interaction and shifting the interaction with a parameter e_0 for the residue-residue to residue-solvent interactions with

$$\epsilon_{ij} = \lambda(e_{ij} - e_0). \tag{1.2}$$

For the KH model the values $\lambda = 0.159$ and $e_0 = -2.27\,\mathrm{k_B T}$ have been used, based on parametrizations to reproduce the experimental determined second virial coefficient of hen egg lysozyme and the dissociation constant K_d of the ubiquitin uim1 system [1]. The contact distances σ_{ij} are determined as weighted average $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ from the individual amino acids diameters, Table 1.1. Note that in the original paper these values have been incorrectly labeled as radii.

Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
5.0	6.6	5.7	5.6	5.5	6.0	5.9	4.5	6.1	6.2

Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
6.2	6.4	6.2	6.4	5.6	5.2	5.6	6.8	6.5	5.9

Table 1.1: Van-der-Waals diameters of amino acids in Å[1].

The electrostatic potential consists of Coulomb interactions with an ionic-screening term

$$U_{el}(r) = \underbrace{\frac{q_i q_j e^2}{4\pi\epsilon_0 D_{el} r}}_{\text{coulomb}} \underbrace{\exp\left(\frac{-r}{\xi}\right)}_{\text{ionic-screening}} \underbrace{\frac{1}{k_B T}}_{\text{scaling factor}}, \tag{1.3}$$

with e the elementary charge, ϵ_0 the vacuum permittivity, D_{el} the dielectric constant and ξ the Debye-length. The scaling factor of $1/k_B T$ is used to convert the electrostatic energy into units of $k_B T$. Bead charges are set according to amino acid type corresponding to a pH of 7. Arginine and lysine are charged with $+e$, histidine with $+\frac{1}{2}e$, due to its isoelectric point and aspartate and glutamine with $-e$. Charges for other amino acids are set to 0. The ionic-screening term is used to set the salt concentration of the environment with the

Debye-length

$$\xi = \sqrt{\frac{\epsilon_r \epsilon_0 k_\mathrm{B} T}{e^2 N_A 2I}}, \tag{1.4}$$

where ϵ_r the absolute permittivity, and I the ionic strength. For a typical salt concentration of 100 mM NaCl ξ is about 10 [Pleaseinsert\PrerenderUnicode{ÃĚ}intopreamble].

Chapter 2

Domains

Proteins and multiprotein complexes consist of multiple units that are connected together. The domains are the second abstraction level of the complexes model that connect beads. This abstraction is based on the assumption that a protein complex consists of different proteins that behave as single units for a short time span. There can be parts that are rigid during the lifespan of the protein complex and short stretches of unstructured peptide chains that serve to tether together as stiff parts. Domains are used to model this varied behavior. To model the two different behaviors the complexes model has rigid and flexible domains.

2.1 Rigid Domain

Rigid domains are the simplest form of a domain and the most versatile at the same time. As the name suggests in a rigid domain the internal coordinates of the beads in the domain do not change over time. Rigid domains are so versatile because they can be used to model very different things. The obvious cases are rigid protein parts like an α-helix or a β-sheet. While not described in the original complexes model it is possible to describe a rigid domain at an even coarser level by grouping together amino acids. Using such a CG description requires finding new forcefield parameters for the interactions but this versatility makes it possible to set up simulations that incorporate experimental data with an appropriate detailed given experimental uncertainties and prior knowledge.

2.2 Flexible Domain

A compelling feature of the original complexes implementation [1] was its explicit treatment of flexible chains in protein complexes. These chains served as an anchor to link rigid domains together. In the original paper [1] a peptide chain model using bond, angle and dihedral potentials similar to molecular dynamics (MD) forcefields [30] has been used. The advantage of such a model is that amino acids are modeled explicitly and can interact with the rigid domains. A drawback of using this model with a Monte-Carlo scheme is that only small movements of the whole chain can be made to have small energy differences and therefore good acceptance probabilities. As a side effect of this, the chain diffuses slowly through configuration space and the overall diffusion of attached rigid domains is limited by the linker instead of the translation and rotation step-size chosen for the rigid domains. While such an explicit model can work for smaller complexes [14] it would make simulations of larger complexes [15] difficult.

A linker model that has a limited influence on the diffusion of the rigid domains would be better suited to quickly sample configuration space. It has been shown that for a

linker only the length is important and not the exact dynamics [24]. We use this to make the assumption that the linker only has to hold two rigid domains together and has no other functional purpose and replace the explicit peptide chain with a potential of mean force (PMF) that only depends on the distance between the two rigid domains and the number of amino acids in the linker. This PMF potential acts as a restraint potential ensuring that the distance between two rigid domains is physically correct. Because no explicit beads are involved the diffusion of the rigid domains is only influenced by the selected translation and rotation step-size. We can use a Gaussian chain polymer model to calculate a suitable PMF. Gaussian chains are suitable models previously used to explain fluorescence resonance energy transfer (FRET) experiments [31].

In the Gaussian chain model the beads are point particles connected by harmonic springs. The average distance b between two beads determines the spring constant. For a three dimensional chain the distances $R_{ij} = |\vec{r}_j - \vec{r}_i|$ between beads i and j, irrespective of the direction, is distributed according to [32]

$$P(R_{ij}) = 4\pi R_{ij}^2 \left(\frac{3}{2\pi\langle R_{ij}^2\rangle}\right)^{3/2} \exp\left(-\frac{3R_{ij}^2}{2\langle R_{ij}^2\rangle}\right), R_{ij} > 0, \tag{2.1}$$

where $\langle R_{ij}^2\rangle = b^2(j-i)$ is the mean squared distances between beads i and j. The factor $4\pi R_{ij}^2$ is the volume element of a shell with width dR and radius R_{ij}. The end-to-end distance for a Gaussian chain with N beads is

$$\sqrt{\langle R_{1N}^2\rangle} = \sqrt{N}b. \tag{2.2}$$

To construct a PMF for the Gaussian chain model we look at the exponential term in eq 2.1. It is a similar expression as for the Boltzmann distribution for a harmonic oscillator

$$V(\vec{r}_i, \vec{r}_j) = \frac{3}{2b^2}\frac{1}{(j-i)}(\vec{r}_i - \vec{r}_j)^2, \tag{2.3}$$

with spring constant $3/(b^2(j-i))$. From this the potential of mean force between the ends of a Gaussian chain of length N follows as

$$PMF(\vec{r}_1, \vec{r}_N) = \frac{3}{2b^2}\frac{1}{N-1}(\vec{r}_1 - \vec{r}_N)^2. \tag{2.4}$$

To compare the distribution of end-to-end distances obtained from this PMF we run a Monte-Carlo simulation using eq 2.4 for a linker of length $N = 200$ and a bond length $b = 3.81\,\text{Å}$ [33], see Figure 2.1. The distribution of end-to-end distances created using the PMF agrees well with the expected distribution of the Gaussian polymer model. This PMF has been previously used to model proteins and ribonucleic acid (RNA) [34].

The final PMF that we use will be between two beads of the two connected rigid domains. These two beads have to be added to the length N of the linker. Therefore the final PMF for a linker of length N is

$$PMF(\vec{r}_0, \vec{r}_{N+1}) = \frac{3}{2b^2}\frac{1}{N+1}(\vec{r}_0 - \vec{r}_{N+1})^2, \tag{2.5}$$

with $\vec{r}_0$ and $\vec{r}_{N+1}$ being the two beads of the rigid domains that are connected by the linker.

2.3 Generating Explicit Beads for Linker Model

While the PMF, eq 2.4, restraint potential, without explicit modeling of the linker beads, is good for fast exploration of phase space some applications, like the comparison of simulations to small angle x-ray scattering (SAXS) measurements [14, 15], require to have

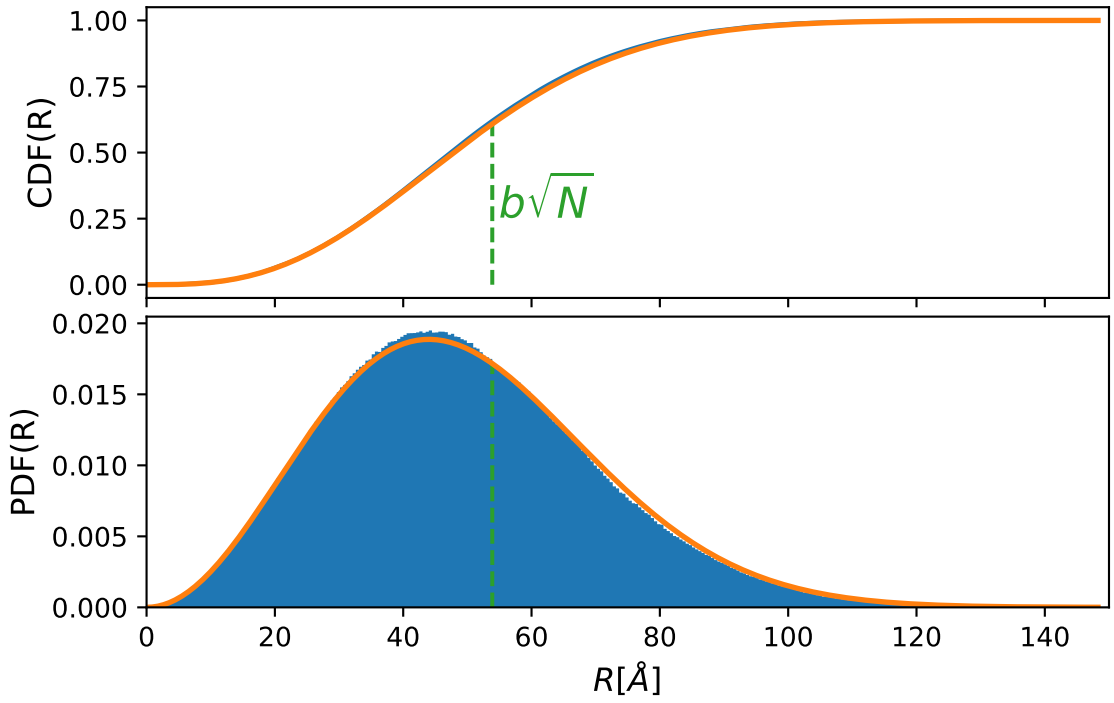

Figure 2.1: Cumulative distribution function (top) and probability density function (bottom) of the end-to-end distance R for a Gaussian chain of length $N = 200$ and bond length $b = 3.81$ Å. The distributions for the ideal Gaussian chain are shown in orange. Results from an ensemble of 1 million distances obtained from a Monte-Carlo simulation using the PMF eq 2.4 are shown in blue. The value of the mean end-to-end distance, eq 2.2, is marked in green.

explicit beads for the linker domains. We now describe an iterative algorithm to generate positions for the linker beads given fixed positions for the first and last bead of the linker. Given the positions of bead N and 1 a single bead $N-1$ can be generated with the following algorithm:

1. Randomly choose a distance d_{start} between bead N and $N-1$ distributed according to eq 2.1. This distance defines a sphere around bead N on which bead $N-1$ will be placed. (Figure 2.2, red circle)

2. Choose a distance d_{end} between bead 1 and $N-1$ so that the sphere around bead 1 intersects with the sphere calculated in step 1. (Figure 2.2, blue cone)

3. Choose a random point on the intersection of the red and blue sphere to place the bead $N-1$ (Figure 2.2, green bead). In three dimensions the intersection is a circle so a random angle θ has to be drawn.

To grow the next bead, with index $N-2$, simply repeat the algorithm with bead $N-1$ as new starting point to choose the random distance $R_{N-1,N-2}$. Repeat this algorithm until all beads are generated. This algorithm gives the three distance between the beads and orientation that uniquely determines where a new bead should be placed. To calculate the actual coordinates in the coordinate system of the simulation we are using the following algorithm.

1. Determine axis $\vec{z'}$ along the vector $\vec{r_N} - \vec{r_1}$.

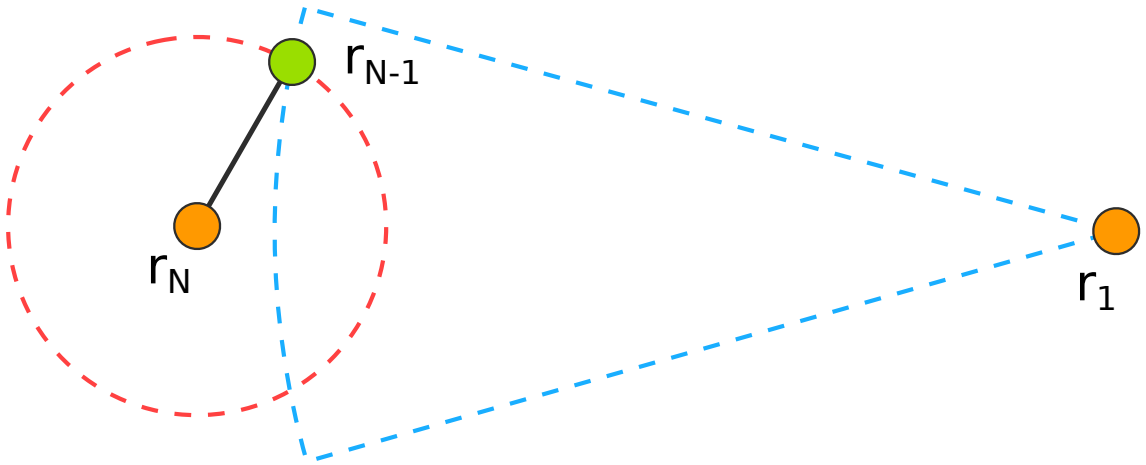

Figure 2.2: Example of distances drawn for a new bead $N-1$ (green) between to fixed endpoints (orange) of a Gaussian linker of length N in two dimensions. The red circle is the distance between the beads N and $N-1$. This distance is randomly chosen from eq 2.1. The bead $N-1$ can be placed anywhere on this circle. The distance between bead 1 and $N-1$ now has to be chosen so that circle (blue) drawn around bead 1 intersects with the first circle (red). In two dimensions this restricts the positions of the new bead to the two intersection points. In three dimensions it would be restricted to a sphere.

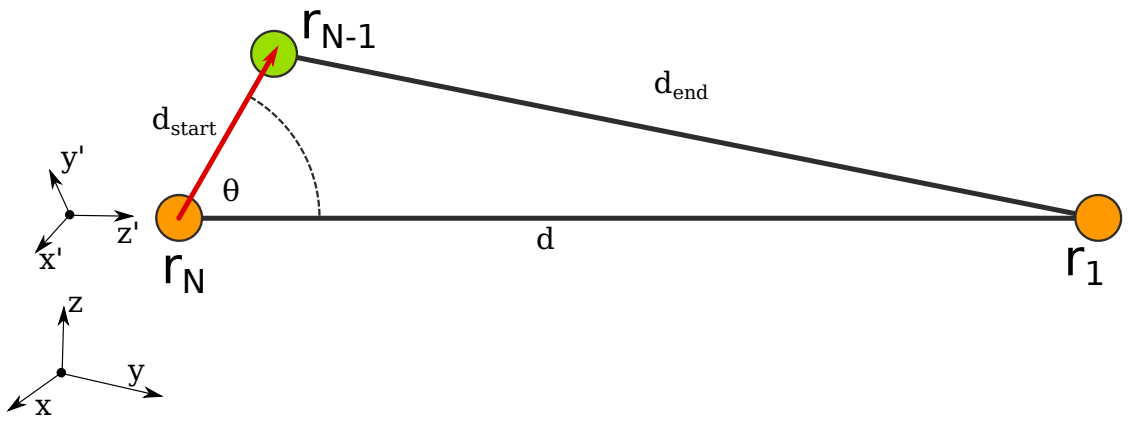

Figure 2.3: Schematic for calculating for calculating the position of bead $\vec{r}_{N-1}$ given the positions of $\vec{r}_N$, $\vec{r}_1$, the distances d, d_{start}, d_{end} and an angle ϕ (not shown). The coordinate system $(\vec{x}, \vec{y}, \vec{z})$ is our reference coordinate system. The coordinate system $(\vec{x'}, \vec{y'}, \vec{z'})$ is used to calculate $\vec{r}_{N-1}$.

2. Determine perpendicular axes $\vec{x'} = \left(\frac{-z_1'-z_2'}{z_0'}, 1, 1\right)^T$ and normalize. Permutate in elements of $\vec{x'}$ if $z_0' = 0$.
3. Determine $\vec{y'} = \vec{x'} \times \vec{z'}$, with $\times$ the cross product.
4. Determine angle ϕ using the law of cosines $\cos(\phi) = \frac{d_{\text{start}}^2 + d^2 - d_{\text{end}}^2}{2 d_{\text{start}} d}$, with $d = |\vec{r_n} - \vec{r_1}|$ see Figure 2.3.
5. Calculate $\vec{r_{N-1}}$ in the coordinate system spanned by $(\vec{x'}, \vec{y'}, \vec{z'})$ from the spherical coordinates given by $(d_{\text{start}}, \theta, \phi)$, see Figure 2.3.

6. Convert $r_{\vec{N-1}}$ into reference coordinate system $(\vec{x}, \vec{y}, \vec{z})$.

Linker configurations generated using the above two algorithms will include overlaps between neighboring beads, see Figure 2.4. To avoid overlaps and generate more extended configurations it's sufficient to add overlap checks in step 1 and 3 in the first algorithm algorithm, i.e. when two beads are to close to each other.

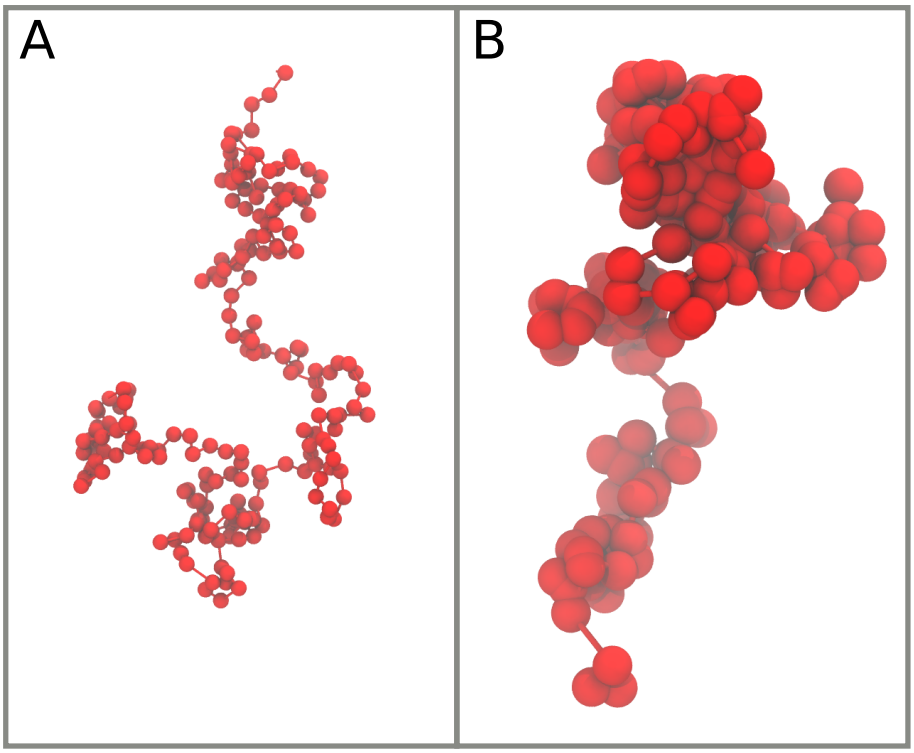

Figure 2.4: Gaussian chain with (left) and without (right) overlap check. Example configuration for a alanine chain of length 200 with a bond-length of b =3.81 Å. All beads are drawn with a diameter of 3.81 Å.

2.4 Relaxation of Gaussian Polymerchain

The structures produced by the Gaussian chain growing algorithms are not very physical. The distances between beads vary by a standard deviation of 1 Å with a mean of 5 Å in a single chain when the chain is grown with overlap checks. The fluctuation in typical protein structures is less than a hundredth of an [Pleaseinsert\PrerenderUnicode{ÃĚ}intopreamble] with a mean distance of 3.81 Å [33]. For the generation of more physical bead coordinates, it is, therefore, necessary to relax the structures generated by the previous algorithm. For the relaxation, the force field developed by Kim and Hummer [1] can be used.

The forcefield consists of bond potentials for pseudobonds for C_α–C_α, angle potentials for pseudo angles C_α–C_α–C_α and torsion potentials for pseudo torsion C_α–C_α–C_α–C_α. The bond potential is a harmonic potential

$$U_{\text{bond}} = \frac{1}{2}k(r - r_0)^2, \tag{2.6}$$

with r the C_α–C_α distance, $r_0 = 3.81$ Å the reference distance and $k = 378\,\text{kcal}\,\text{mol}^{-1}\text{Å}^{-2}$ the spring constant [35]. The pseudoangle potentials is given a double well potential [33]

$$\exp[-\gamma U_{\text{angle}}(\theta)] = \exp[-\gamma(k_\alpha(\theta - \theta_\alpha) + \epsilon_\alpha)] + \exp[-\gamma k_\beta(\theta - \theta_\beta)^2], \tag{2.7}$$

where θ is the angle between $C_\alpha - C_\alpha - C_\alpha$, the constants are $\gamma = 0.1\,\mathrm{mol\,kcal^{-1}}$, $\epsilon_\alpha = 4.3\,\mathrm{kcal\,mol^{-1}}$, $\theta_\alpha = 1.6\,\mathrm{rad}$, $\theta_\beta = 2.27\,\mathrm{rad}$, $k_\alpha = 106.4\,\mathrm{kcal\,mol^{-1}rad^{-2}}$ and $k_\beta = 23.6\,\mathrm{kcal\,mol^{-1}rad^{-2}}$. This potential accounts for the helical and extended pseudoangles. The torsion potential is given by [35]

$$U_{\mathrm{torsion}}(\phi) = \sum_{n=1}^{4} [1 + \cos(n\phi - \delta_n)]V_n, \tag{2.8}$$

where ϕ is the torsion angle of the middle two beads in $C_\alpha - C_\alpha - C_\alpha - C_\alpha$. The constants V_n and δ_n are chosen for alanine as $V_n = [0.936472, 2.307767, 0.131743, 0.613133]$ $\mathrm{k_B T}$ and $\delta_n = [287.354830, 271.691192, 180.488748, 108.041256]$ rad [35].

The relaxation with the above described potential is done using a Monte-Carlo algorithm. For trial moves the position of individual beads is changed. The start and end bead are treated as fixed. For a linker of length N the probability to pick a bead is uniform between all $N-2$ beads that are allowed to move. A sweep consists of $N-2$ trial moves. Because the structure is only supposed to be relaxed it isn't necessary to generate structures from an equilibrium distribution and therefore detailed balance does not need to be preserved. Here the move width is adjusted after each sweep to achieve a target acceptance ratio of 30 %. If after a sweep the acceptance ratio is larger than 30 % the step-width is increased by 10 % and decreased if the acceptance ration is below 30 %.

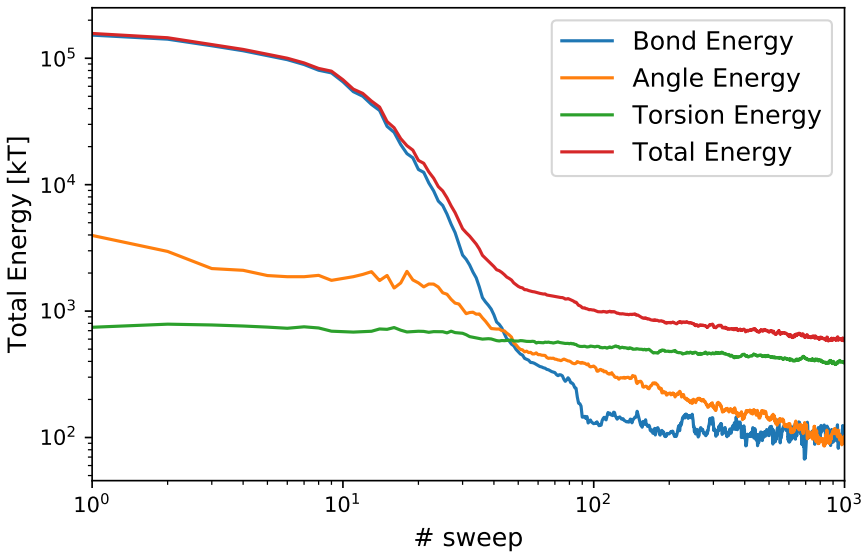

Figure 2.5: Energies of a nonoverlapping Gaussian chain during relaxation. The chain is 200 beads long and the initial structure was generated using the Gaussian chain model with no overlaps. The acceptance rate during the Monte Carlo simulation was set to target 30 %. The bond energy is shown in blue, the angle energy in orange, the torsion energy in green and the total energy in red.

Energy contributions for each potential from a relaxation run for a chain of 200 beads is shown in Figure 2.5. In the beginning, the energy is dominated by the bond potential, this is due to the fact that average bond-length is larger than 3.81 Å. The bond lengths are fully relaxed after around 200 sweeps. The angle potentials start to relax around sweep 50 when the bond energy has already dropped by a factor of two. The angle potential is fully relaxed after around 1000 sweeps. The last potential to relax are the torsion angles. After about 500 sweeps the torsion angles start to see a more pronounced decrease after 3000

sweeps when the other two potentials haven been fully relaxed. The energy difference in the torsion potential from the beginning of the simulation to the final structure is significantly less than for the other two terms in the energy. The reason for this could be that the starting structures generated by the Gaussian chain algorithm are particularly ill chosen or that the simple spatial trial moves of the beads are not good for relaxing this potential function.

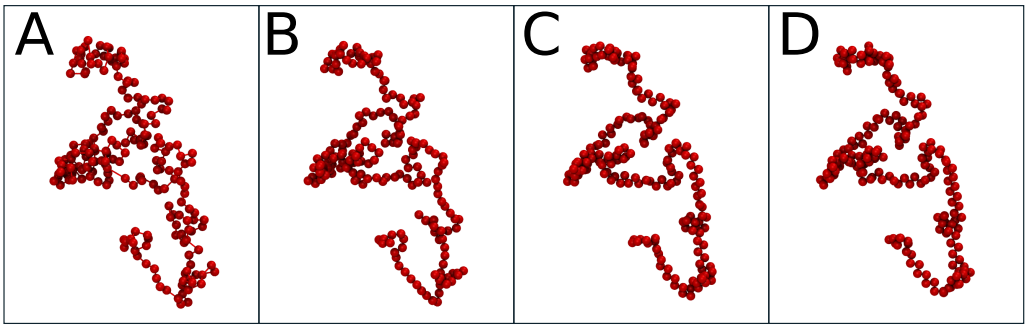

Figure 2.6: Linker configuration before relaxation (A) and after using only the bond potential (B), using the bond and angle potential (C), and using the bond, angle and torsion potential (D). All relaxation runs used the same initial structure (A).

The structures generated by relaxing an initial configuration from the Gaussian chain algorithm can be seen in Figure 2.6. For comparison, single structures have been generated with the full potential, only the bond potential, and the bond and angle potential. The structures have all been generated from the same initial structure and relaxation runs where 1000 sweeps long. The bond potential alone has the biggest visual influence on the structure by achieving a more uniform bond distance. The addition of the angle potential also gives a visual improvement. Differentiating the bond and angle potential structure from the structure with the full potential is difficult as both are very similar.

2.5 Comparison of Unfolded Proteins and Linker Model

To understand how well the model describes unfolded protein regions I will compare the radius of gyration R_G, as a measure of compactness, of our model with experimental data. For unfolded proteins the R_G has been determined experimentally in dependence on the protein length with denaturants [36]

$$\langle R_G \rangle = R_0 N^{\nu}, \tag{2.9}$$

with $R_0 = 1.927^{+0.271}_{-0.238}$Å and $\nu = 0.598 \pm 0.028$. The radius of gyration of the Gaussian chain model is

$$\langle R_G \rangle = \sqrt{\frac{1}{6}\frac{N(N+2)}{N+1}}b \approx \sqrt{\frac{1}{6}}N^{1/2}b. \tag{2.10}$$

This scaling behavior is slightly different with $\nu = .5$ and $R_0 = 1.555\,$Å. Therefore it is unlikely that the Gaussian polymer without overlap checks reproduces the R_G values of a denatured protein for any number of beads. To compare the R_G scaling behavior of the linker growth algorithm to the scaling of denatured proteins we generate 1000 different structures for a linker of 25 to 200 beads length. To get the values of a truly free chain for

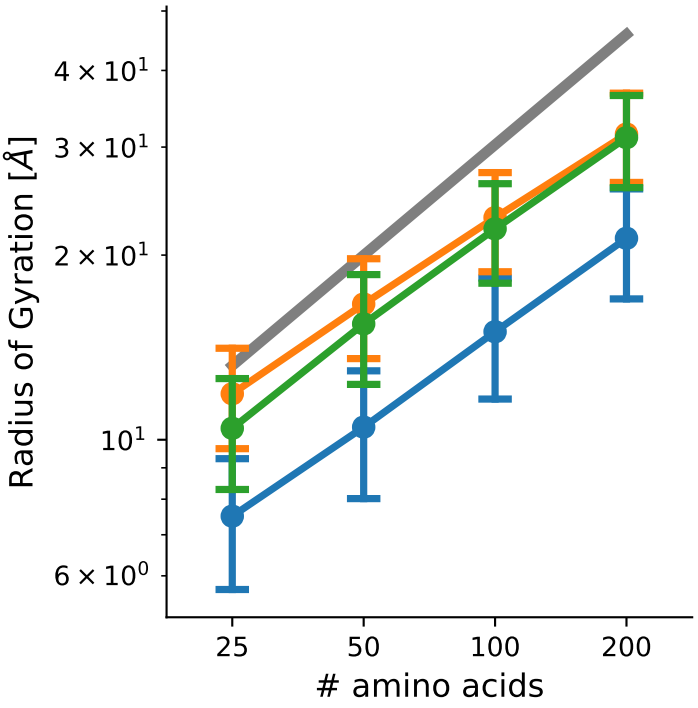

Figure 2.7: Radius of Gyration computed for a linker with of different number of amino acids. For each length 1000 structures have been generated with overlap check (orange), without (blue), and fill relaxation (green). Error bars donate the standard deviation. The experimental R_G scaling law for denatured proteins [36] is shown as gray line.

each simulation the linker was padded with 25 beads in the front and end so that only the middle N beads have been used to calculate the R_G. The bond-length was set to 3.81 Å [33]. Start and end points have been placed at the optimal end-to-end distance, eq 2.2. Results are shown in Figure 2.7. As anticipated without an overlap check the R_G values are systematically different. But with overlap checks enabled the Gaussian polymer is within one standard of the experimental values if $N < 50$.

It should be noted that for intrinsically disordered proteins it has been shown that the R_G with denaturants is larger than of the protein observed in natural conditions [37–40]. Therefore our Gaussian polymer model is a good enough description for the flexible domains. A similar model has been employed to study intrinsically disordered proteins [7].

Chapter 3

Topologies

Topologies are a collection of connected domains. They can be used to model large protein complexes that consist of multiple domains. The connection is typically modeled as a distance based potential between two beads of the connected domains. The connection potential can be chosen according to the Gaussian chain model eq 2.4 or as a normal harmonic spring.

Chapter 4

Pair Interactions Potentials

In Complexes++ pair potentials are called pair kernels following a common terminology used in computer science. The parameters of all available pair kernels are given in the forcefield class. The parameters σ_{ij} and ϵ_{ij} are set according to bead type and shared between the pair-kernels. During a simulation pair-kernels can be chosen for each individual domain type pair present in the simulation, allowing to fine tune the interactions. The Lennard-Jones like potential eq 1.1 in combination with the electrostatic potential eq 1.3, and several other potentials have been implemented. One additional potential is the Weeks-Chandler-Anderson (WCA) potential [41]

$$U_{\mathrm{WCA}}(r, \sigma_{ij}, \epsilon_{ij}) = \begin{cases} -\epsilon_{ij} & \text{if } r < 2^{1/6}\sigma_{ij} \\ 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] & \text{if } \epsilon_{ij} > 0. \end{cases} \tag{4.1}$$

An alternative version of the Lennard-Jones like potential that smoothly decays to zero is implemented with the following smoothing term

$$F_{\mathrm{smooth}}(r, a, b) = \begin{cases} 1 & \text{if } r/\sigma_{ij} < a \\ 0 & \text{if } r/\sigma_{ij} > b \\ \frac{(b^2-(r/\sigma_{ij})^2)^2(b^2+2(r/\sigma_{ij})^2-3a^2)}{(b-a)^3} & \text{otherwise ,} \end{cases} \tag{4.2}$$

with $a = 1.4\,\text{Å}$ and $b = 1.8\,\text{Å}$ being the bound in which the potential decays to 0. The value for a and b are hard coded. A purely repulsive potential with

$$U_{\mathrm{repulsive}} = \left(\frac{\sigma_{ij}}{r}\right)^{12} \tag{4.3}$$

is also implemented. Also implemented is a soft-core potential [42] that allows to tune how soft the beads are and therefore they can potentially overlap. The soft-core potential is a modification of the Lennard-Jones like potential without the explicit repulsion branch,

$$U_{\mathrm{SC}}(r_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}^6}{\alpha\sigma_{ij}^6 + (r_{ij}-s)^6}\right)^2 - \left(\frac{\sigma_{ij}^6}{\alpha\sigma_{ij}^6 + (r_{ij}-s)^6}\right)\right], \tag{4.4}$$

with α the parameter to tune the softness of the beads, and $s = \left(\sqrt[6]{2} - \sqrt[6]{2-\alpha}\right)\sigma_{ij}$ a shift parameter to ensure that the minimum is always at $\sqrt[6]{2}$ independent of α. The other branches of the Lennard-Jones like potential can be obtained by applying the same modifications. α can be changed in the range of zero to one, with $\alpha = 1$ allowing full overlap of the beads as $U_{\mathrm{SC}}(r = 0) = 0$ and $\alpha = 0$ recovering the Lennard-Jones like potential $U_{\mathrm{SC}}(r = 0) = \infty$, eq 1.1. Because this potential explicitly allows overlaps the

electrostatics potential also has to be changed to remove the divergence at $r = 0$. For this we use the potential between two Gaussian charge distributions [43]

$$U_{\mathrm{el}}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 D} \frac{\mathrm{erf}\left(r_{ij}\sqrt{\lambda_{ij}}\right)}{r_{ij}} \exp\left(-\frac{r_{ij}}{\zeta}\right) \frac{1}{k_{\mathrm{B}}T} \tag{4.5}$$

with $\lambda_{ij} = \frac{\lambda_i \lambda_j}{\lambda_i + \lambda_j}$ and λ_i the charge radius of bead i. The charge radii are specified in a forcefield for every bead type. In the standard Complexes++ forcefield all radii are set to one.

Chapter 5

Replica Exchange Algorithms

For enhanced sampling, Complexes++ implements replica exchange algorithms [44–46]. In replica exchange simulations N independent copies of a simulation, which are further refereed to as replicas, are run simultaneously and exchanged periodically. Implemented are Temperature Replica Exchange [47], Hamiltonian replica exchange [48] and pressure replica exchange [49].

The implemented replica exchange algorithms differ by the acceptance function. For temperature replica exchange the acceptance function for two configurations i and j is [47]

$$W(i \to j) = \min(1, \exp\left((\beta_j - \beta_i)(U(x_j) - U(x_i))\right)), \tag{5.1}$$

with U being the energy function, x_i the configuration of replica i and β_i the temperature of replica i. For Hamiltonian replica exchange the acceptance function is [48]

$$W(i \to j) = \min\left(1, \exp\left(\frac{1}{\beta_i}(U_i(x_i) - U_i(x_j)) + \frac{1}{\beta_j}(U_j(x_j) - U_j(x_i))\right)\right), \tag{5.2}$$

with U_i the energy function of replica i. For pressure replica exchange the acceptance function is [49]

$$W(i \to j) = \min\left(1, \exp\left((\beta_i - \beta_j)(U(x_i) - U(x_j)) + (\beta_i P_i - \beta_j P_j)(V_i - V_j)\right)\right), \tag{5.3}$$

with P_i the pressure of replica i and V_i the volume of replica i.

Bibliography

[1] Young C. Kim and Gerhard Hummer. Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. *J. Mol. Biol.*, 375(5): 1416–1433, 2008. ISSN 00222836. doi: 10.1016/j.jmb.2007.11.063.

[2] Fangqiang Zhu and Bo Chen. Monte Carlo Simulations of HIV Capsid Protein Homodimer. *J. Chem. Inf. Model.*, 55(7):1361–1368, 2015. ISSN 15205142. doi: 10.1021/acs.jcim.5b00126.

[3] Hao Sha and Fangqiang Zhu. Parameter Optimization for Interaction between C-Terminal Domains of HIV-1 Capsid Protein. *J. Chem. Inf. Model.*, 57(5):1134–1141, 2017. ISSN 15205142. doi: 10.1021/acs.jcim.7b00011.

[4] Young C. Kim and Jeetain Mittal. Crowding induced entropy-enthalpy compensation in protein association equilibria. *Phys. Rev. Lett.*, 110(20):1–5, 2013. ISSN 00319007. doi: 10.1103/PhysRevLett.110.208102.

[5] Young C. Kim, Robert B. Best, and Jeetain Mittal. Macromolecular crowding effects on protein-protein binding affinity and specificity. *J. Chem. Phys.*, 133(20), 2010. ISSN 00219606. doi: 10.1063/1.3516589.

[6] J Rosen and Young C. Kim. Modest Protein - Crowder Attractive Interactions Can Counteract Enhancement of Protein Association by Intermolecular Excluded Volume Interactions. *J. Phys. Chem. B*, I:2683–2689, 2011.

[7] Gregory L. Dignon, Wenwei Zheng, Young C. Kim, Robert B. Best, and Jeetain Mittal. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.*, 14(1):1–23, 2018. ISSN 15537358. doi: 10.1371/journal.pcbi.1005941.

[8] Tatjana Skrbić, Pietro Faccioli, and Cristian Micheletti. The role of non-native interactions in the folding of knotted proteins: insights from molecular dynamics simulations. *PLoS Comput. Biol.*, 8(6):1–12, 2012. ISSN 2218273X. doi: 10.3390/biom4010001.

[9] Silvio a Beccara, Tatjana Škrbić, Roberto Covino, Cristian Micheletti, and Pietro Faccioli. Folding Pathways of a Knotted Protein with a Realistic Atomistic Force Field. *PLoS Comput. Biol.*, 9(3), 2013. ISSN 1553734X. doi: 10.1371/journal.pcbi.1003002.

[10] Anshul Sirur and Robert B. Best. Effects of interactions with the groel cavity on protein folding rates. *Biophys. J.*, 104(5):1098–1106, 2013. ISSN 00063495. doi: 10.1016/j.bpj.2013.01.034.

[11] Roberto Covino, Tatjana Skrbić, Silvio A. Beccara, Pietro Faccioli, and Cristian Micheletti. The role of non-native interactions in the folding of knotted proteins: insights from molecular dynamics simulations. *Biomolecules*, 4(1):1–19, 2014. ISSN 2218273X. doi: 10.3390/biom4010001.

[12] Anshul Sirur, David De Sancho, and Robert B. Best. Markov state models of protein misfolding. *J. Chem. Phys.*, 144(7), 2016. ISSN 00219606. doi: 10.1063/1.4941579.

[13] Roberto Covino. *Investigating Protein Folding Pathways at Atomistic Resolution : from a Small Domain to a Knotted Protein*. Phd, UNIVERSITÀ DEGLI STUDI DI TRENTO, 2013.

[14] Bartosz Różycki, Young C. Kim, and Gerhard Hummer. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure*, 19(1):109–116, 2011. ISSN 09692126. doi: 10.1016/j.str.2010.10.006.

[15] Jürgen Köfinger, Michael J. Ragusa, Il-Hyung Lee, Gerhard Hummer, and James H. Hurley. Solution Structure of the Atg1 Complex: Implications for the Architecture of the Phagophore Assembly Site. *Structure*, 23(5):809–818, 2015. ISSN 09692126. doi: 10.1016/j.str.2015.02.012.

[16] A. Baumlova, D. Chalupska, B. Rozycki, M. Jovic, E. Wisniewski, M. Klima, A. Dubankova, D. P. Kloer, R. Nencka, T. Balla, and E. Boura. The crystal structure of the phosphatidylinositol 4-kinase II. *EMBO Rep.*, 15(10):1085–1092, 2014. ISSN 1469-221X. doi: 10.15252/embr.201438841.

[17] Bartosz Rózycki, Marek Cieplak, and Mirjam Czjzek. Large conformational fluctuations of the multi-domain xylanase Z of Clostridium thermocellum. *J. Struct. Biol.*, 191(1):68–75, 2015. ISSN 10958657. doi: 10.1016/j.jsb.2015.05.004.

[18] Dominika Chalupska, Andrea Eisenreichova, Bartosz Różycki, Lenka Rezabkova, Jana Humpolickova, Martin Klima, and Evzen Boura. Structural analysis of phosphatidylinositol 4-kinase IIIβ (PI4KB) – 14-3-3 protein complex reveals internal flexibility and explains 14-3-3 mediated protection from degradation in vitro. *J. Struct. Biol.*, 200(1):36–44, 2017. ISSN 10958657. doi: 10.1016/j.jsb.2017.08.006.

[19] Kei Ichi Okazaki, Takato Sato, and Mitsunori Takano. Temperature-enhanced association of proteins due to electrostatic interaction: A coarse-grained simulation of actin-myosin binding. *J. Am. Chem. Soc.*, 134(21):8918–8925, 2012. ISSN 00027863. doi: 10.1021/ja301447j.

[20] Duccio Malinverni, Alfredo Jost Lopez, Paolo De Los Rios, Gerhard Hummer, and Alessandro Barducci. Modeling Hsp70/Hsp40 interaction by multi-scale molecular simulations and coevolutionary sequence analysis. *Elife*, 6:1–20, 2017. ISSN 2050084X. doi: 10.7554/eLife.23471.

[21] Krishnakumar M. Ravikumar, Wei Huang, and Sichun Yang. Coarse-grained simulations of protein-protein association: An energy landscape perspective. *Biophys. J.*, 103(4):837–845, 2012. ISSN 00063495. doi: 10.1016/j.bpj.2012.07.013.

[22] Shilpa Yadahalli, V. V. Hemanth Giri Rao, and Shachi Gosavi. Modeling Non-Native Interactions in Designed Proteins. *Isr. J. Chem.*, 576104, 2014. ISSN 00212148. doi: 10.1002/ijch.201400035.

[23] Nicole Fortoul, Pankaj Singh, Chung Yuen Hui, Maria Bykhovskaia, and Anand Jagota. Coarse-grained model of SNARE-mediated docking. *Biophys. J.*, 108(9): 2258–2269, 2015. ISSN 15420086. doi: 10.1016/j.bpj.2015.03.053.

[24] Bartosz Różycki, Pierre-André Cazade, Shane O'Mahony, Damien Thompson, and Marek Cieplak. The length but not the sequence of peptide linker modules exerts the primary influence on the conformations of protein domains in cellulosome multienzyme complexes. *Phys. Chem. Chem. Phys.*, 19(32):21414–21425, 2017. ISSN 1463-9076. doi: 10.1039/C7CP04114D.

[25] M Fortoul, Maria Bykhovskaia, and Anand Jagota. Coarse-Grained Model of the SNARE Complex Shows that Quick Zippering Requires Partial Assembly. *bioRxiv*, 2018. doi: 10.1101/294181.

[26] Brandon G Horan, Dimitrios Vavylonis, and Young C Kim. Computational modeling highlights disordered Formin Homology 1 domain's role in profilin-actin transfer. *bioRxiv*, 2018. doi: 10.1101/263566.

[27] Bartosz Rozycki and Marek Cieplak. Stiffness of the C-terminal disordered linker affects the geometry of the active site in endoglucanase Cel8A. *Mol. BioSyst. Mol. BioSyst*, 12(I):1–2, 2016. ISSN 1463-9076. doi: 10.1039/C6MB00606J.

[28] Oren Ben-Kiki, Clark Evans, and Ingy döt Net. http://yaml.org/spec/1.2/spec.html, 2018.

[29] Sanzo Miyazawa and Robert L. Jernigan. Residue – Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.*, 256(3):623–644, 1996. ISSN 00222836. doi: 10.1006/jmbi.1996.0114.

[30] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins*, 65:712–725, 2006.

[31] Edward P. O'Brien, Greg Morrison, Bernard R. Brooks, and D. Thirumalai. How accurate are polymer models in the analysis of Förster resonance energy transfer experiments on proteins? *J. Chem. Phys.*, 130(12), 2009. ISSN 00219606. doi: 10.1063/1.3082151.

[32] Hiromi Yamakawa. *Modern Theory of Polymer Solutions*. Harper & Row, NewYork, 1971.

[33] Robert B. Best, Yng Gwei Chen, and Gerhard Hummer. Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of Arc repressor. *Structure*, 13:1755–1763, 2005. ISSN 09692126. doi: 10.1016/j.str.2005.08.009.

[34] C. Hyeon, G. Morrison, and D. Thirumalai. Force-dependent hopping rates of RNA hairpins can be estimated from accurate measurement of the folding landscapes. *Proc. Natl. Acad. Sci.*, 105(28):9604–9609, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0802484105.

[35] John Karanicolas and Charles L Brooks. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.*, 11(10):2351–2361, 2002. ISSN 0961-8368. doi: 10.1110/ps.0205402.

[36] Jonathan E. Kohn, Ian S. Millett, Jaby Jacob, Bojan Zagrovic, Thomas M. Dillon, Nikolina Cingel, Robin S. Dothager, Soenke Seifert, P. Thiyagarajan, Tobin R. Sosnick, M. Zahid Hasan, Vijay S. Pande, Ingo Ruczinski, Sebastian Doniach, and Kevin W. Plaxco. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 101(34):12491–6, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0403643101.

[37] Gustavo Fuertes, Niccolò Banterle, Kiersten M. Ruff, Aritra Chowdhury, Davide Mercadante, Christine Koehler, Michael Kachala, Gemma Estrada Girona, Sigrid Milles, Ankur Mishra, Patrick R. Onck, Frauke Gräter, Santiago Esteban-Martín, Rohit V. Pappu, Dmitri I. Svergun, and Edward A. Lemke. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci.*, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1704692114.

[38] Jianhui Song, Gregory Neal Gomes, Tongfei Shi, Claudiu C. Gradinaru, and Hue Sun Chan. Conformational Heterogeneity and FRET Data Interpretation for Dimensions of Unfolded Proteins. *Biophys. J.*, 113(5):1012–1024, 2017. ISSN 15420086. doi: 10.1016/j.bpj.2017.07.023.

[39] Wenwei Zheng, Alessandro Borgia, Karin Buholzer, Alexander Grishaev, Benjamin Schuler, and Robert B. Best. Probing the Action of Chemical Denaturant on an Intrinsically Disordered Protein by Simulation and Experiment. *J. Am. Chem. Soc.*, 138(36):11702–11713, 2016. ISSN 15205126. doi: 10.1021/jacs.6b05443.

[40] Alessandro Borgia, Wenwei Zheng, Karin Buholzer, Madeleine B. Borgia, Anja Schüler, Hagen Hofmann, Andrea Soranno, Daniel Nettels, Klaus Gast, Alexander Grishaev, Robert B. Best, and Benjamin Schuler. Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.*, 138(36):11714–11726, 2016. ISSN 15205126. doi: 10.1021/jacs.6b05917.

[41] John D. Weeks, David Chandler, and Hans C. Andersen. Role of repulsive forces in determining the equilibrium structure of simple liquids. *J. Chem. Phys.*, 54(12): 5237–5247, 1971. ISSN 00219606. doi: 10.1063/1.1674820.

[42] Iris Antes. DynaDock: A now molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. *Proteins Struct. Funct. Bioinforma.*, 78(5):1084–1104, 2009. ISSN 08873585. doi: 10.1002/prot.22629.

[43] David R Yarkony. *Modern electronic structure theory Part II*. World Scientific Publishing, Singapore, 1995.

[44] Rh Swendsen and Js Wang. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.*, 57(21):2607–2609, 1986. ISSN 1079-7114. doi: 10.1103/PhysRevLett.57.2607.

[45] Charles H. Bennett. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.*, 22(2):245–268, 1976. ISSN 10902716. doi: 10.1016/0021-9991(76)90078-4.

[46] Mark Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, 2010.

[47] Ulrich H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281(1-3):140–150, 1997. ISSN 00092614. doi: 10.1016/S0009-2614(97)01198-6.

[48] Giovanni Bussi. Hamiltonian replica exchange in GROMACS: a flexible implementation. *Mol. Phys.*, 112(3-4):379–384, 2014. ISSN 0026-8976. doi: 10.1080/00268976.2013.824126.

[49] Tsuneyasu Okabe, Masaaki Kawata, Yuko Okamoto, and Masuhiro Mikami. Replica-exchange Monte Carlo method for the isobaric-isothermal ensemble. *Chem. Phys. Lett.*, 335(5-6):435–439, 2001. ISSN 00092614. doi: 10.1016/S0009-2614(01)00055-0.