ChemPhysChem

Research Article
doi.org/10.1002/cphc.202300521

Chemistry Europe
European Chemical Societies Publishing

www.chemphyschem.org

Reverse Mapping of Coarse Grained Polyglutamine Conformations from PRIME20 Sampling

Thomas Kunze,[b] Christian Dreßler,[a] Christian Lauer,[b] Wolfgang Paul,[b] and Daniel Sebastiani*[b]

An inverse coarse-graining protocol is presented for generating and validating atomistic structures of large (bio-) molecules from conformations obtained via a coarse-grained sampling method. Specifically, the protocol is implemented and tested based on the (coarse-grained) PRIME20 protein model (P20/SAMC), and the resulting all-atom conformations are simulated using conventional biomolecular force fields. The phase space sampling at the coarse-grained level is performed with a stochastical approximation Monte Carlo approach. The method is applied to a series of polypeptides, specifically dimers of polyglutamine with varying chain length in aqueous solution. The majority (>70%) of the conformations obtained from the coarse-grained peptide model can successfully be mapped back to atomistic structures that remain conformationally stable during 10 ns of molecular dynamics simulations. This work can be seen as the first step towards the overarching goal of improving our understanding of protein aggregation phenomena through simulation methods.

1. Introduction

Proteins are one of the key constituents of life on our planet. Composed of specific amino acid sequences,[1,2] they perform a large part of bio-relevant functionality in all living organisms. On the other hand, protein malfunction is at the origin of numerous diseases, among many others, Alzheimer's,[3] Huntington's[4] and Parkinson's[5] disease. One of the problematic processes in this context is their unwanted aggregation, e.g., into amyloid fibers.[6,7]

This aggregation process, its local biochemical prerequisites, and also kinetic and mechanical aspects are the subject of an ongoing intense research effort.[8,9] In this context, computational methods are an important clue to the qualitative and quantitative understanding of the numerous individual elements of the aggregation process.[10] However, computational methods generally address only one particular step or one isolated question of the process, as there are no theoretical approaches that capture the vast complexity of the aggregation in a comprehensive way, i.e. with atomistic resolution on the picosecond timescale, chemical accuracy, hours of simulated times and including macroscopic effects like crowding.[11] There are continuously attempts made in the theory community to "bridge" computational scales, be it length scales, time scales, or accuracy and chemical resolution levels. These attempts normally consist of combining two or more established methods from different regions on those scales, and the theoretical challenge is to yield a consistent description of the system of interest across these methods, meaning that the two distinct methods must be enabled to "hand over" the system forth and back in a consistent manner.

In this context, we present here a protocol that enables the transfer of biomolecular systems of intermediate size between two specific simulation methods which are based on slightly different resolution levels (atomistic versus coarse-grained structures) and different interaction potentials (biomolecular force fields versus hard sphere-type potentials). Therefore, part of the representability and transferability problems of the quasi-global coarse-grained (CG) sampling gets addressed by the local spatio-temporal phase space coverage of the classical force field MD simulations.[12–18]

Specifically, we combine atomistic molecular dynamics simulations with a Monte-Carlo sampling scheme based on the coarse-grained PRIME20 protein model. The difficulty of this combination of simulation methods is the loss of atomistic resolution in the PRIME20 scheme which needs to be reverted and the partial simplification of repulsive and attractive interactions which need to reintroduce the energetic and entropic contributions of the neglected degrees of freedom into the coarse-grained potential. Especially the use of implicit solvent for biomolecules on aqueous solution may lead to a thermodynamically incorrect weighting of conformations of different nature.

Both Monte Carlo and Molecular Dynamics (MD) simulation were extensively used in the past to study bio molecules.[19–24] Several hybrid approaches already combine these two methods, because Monte Carlo and MD simulations are highly complementary techniques.[25–31] While Monte Carlo methods are a

[a] Prof. Dr. C. Dreßler
Institut für Physik, Ilmenau University of Technology
Weimarer Straße 32, 98693 Ilmenau, Germany

[b] T. Kunze, C. Lauer, Prof. Dr. W. Paul, Prof. Dr. D. Sebastiani
Faculty of Natural Sciences II,
Martin-Luther University Halle-Wittenberg
Von-Danckelmann-Platz 4, 06120 Halle, Germany
E-mail: daniel.sebastiani@chemie.uni-halle.de

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cphc.202300521

suitable tool to probe large parts of the conformational space of bio molecules, MD simulations are able to calculate the local structure fluctuations and dynamics of a given peptide configuration. In this work, we will combine the coarse-grained polymer model PRIME20 which has successfully been used in Stochastic Approximation Monte Carlo simulations (P20/SAMC) and an all atom MD simulation. The coarse-grained Monte Carlo model can be used to identify a set of low energy structures, which is not possible from a classical MD trajectory due to the limited length of the simulations. All atom MD simulations starting from the structures obtained from the Monte Carlo method will reveal the full atomistic picture including, e.g., solvation by explicit water molecules. The dynamical properties, such as the evolution of the hydrogen bond network, can be studied in that way and atomistic MD simulation will automatically incorporate entropic contributions of degrees of freedom which had been averaged over in the coarse grained description. In this way, the molecular dynamics simulations will act as validation and a posteriori correction tool for the thermodynamic weighting function for configurations delivered by the Monte Carlo simulations.

There are successful examples for the combination of MD and MC methods. The Inverse Monte Carlo approach[16] or the Iterative Boltzmann Inversion[32] can produce coarse-grained parameters fitted to MD simulation properties such as radial distribution functions. These and similar such methods were successfully improved and used to study a variety of topics.[15,17,33–40]

2. Coarse-Grained Model

The atomistic description of AMBER03[41] follows the general force field approach. In order to compare this already established technique, we have to introduce the characteristics of the PRIME20 model.

The PRIME20 model is a 4-bead model, where each amino acid is represented by 3 backbone beads and 1 side chain bead, as shown in Figure 1. The backbone beads refer to the NH bead, the C_α bead and the CO bead. They are located at the C_α position, the *C* position and the *N* position, respectively. The side chain bead R is located at the center of mass of the side chain, while its position and size is specific for the amino acid it represents. Here, we will focus on the parameters relevant for polyglutamine (PolyQ), which are obtained from the complete list of parameters for the PRIME20 model.[42]

Covalent bonds are represented as white sticks on the right side in Figure 1. They are modeled as infinite well potentials around an ideal bond length. The width of the well allows for bond length fluctuations Δ of 2.375 % from the ideal value:

$$V_{\text{bond}}(d) = \begin{cases} 0 & \text{if } d \in [d_{\text{ideal}} - \Delta, d_{\text{ideal}} + \Delta] \\ \infty & \text{otherwise} \end{cases} \quad (1)$$

Here d represents the distance between two bonds, d_{ideal} is the ideal bond length and $\Delta = 0.02375 d_{\text{ideal}}$. PRIME20 utilizes pseudo-bonds between beads separated by two covalent bonds to stabilize bond angles, and between consecutive C_α beads to keep the peptide in a *trans* configuration. Pseudo-bonds behave in the same way as covalent bonds and are represented by black and yellow sticks in Figure 1. Bond and pseudo-bond lengths for PolyQ are listed in Table 1.

Non-bonded bead interactions separate into two types. On the one hand, there are excluded volume interactions between multiple backbone beads and between backbone and side chain beads. They are modeled as hard-sphere (HS) repulsions. On the other hand, there are hydrophobic interactions between side chain beads as well as hydrogen bond formation between NH and the CO bead, which are modeled as semi-infinite square well potentials:

$$V_{\text{HS}}(d_{ij}) = \begin{cases} 0 & \text{if } d_{ij} > d_{ij}^{\text{HS}} \\ \infty & \text{otherwise} \end{cases} \quad (2)$$

$$V_{\text{SW}}(d_{ij}) = \begin{cases} 0 & \text{if } d_{ij} > d_{ij}^{\text{SW}} \\ \varepsilon_{ij} & \text{if } d_{ij}^{\text{HS}} < d_{ij} < d_{ij}^{\text{SW}} \\ \infty & \text{otherwise} \end{cases} \quad (3)$$

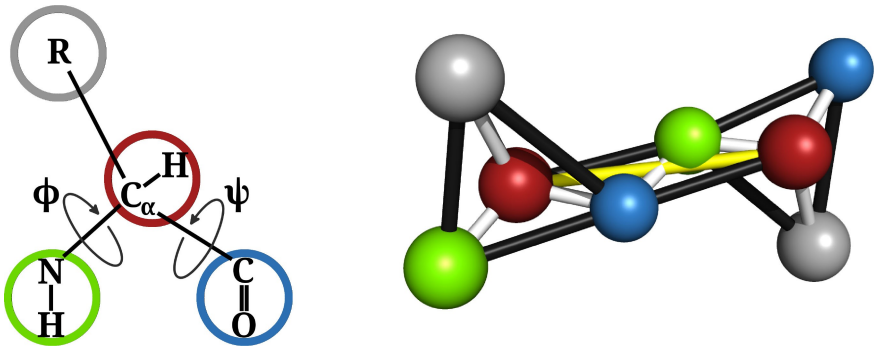

Figure 1. Geometry of the PRIME20 model. The backbone is represented by 3 beads: the NH group (green bead), the C_α carbon (red bead) and the CO group (blue bead). The side chain is represented by the fourth bead (gray bead). Its position and size is specific for the individual type of amino acid. On the left the assignment of atoms to beads and the dihedral angles are shown. On the right the geometry of a PRIME20 dimer is shown. White sticks represent covalent bonds. Black and yellow sticks represent pseudo-bonds that stabilize the structure. The size of the beads is not true to scale.

Table 1. Bond and pseudo-bond lengths between beads of PolyQ in PRIME20. Here, the index i represents beads of the (i)th residue and the index i+1 represents beads of the (i+1)th residue. Sizes in Å.

Bonds	NH_i-$C_{\alpha,i}$	$C_{\alpha,i}$-CO_i	CO_i-NH_{i+1}	R_i-$C_{\alpha,i}$		
	1.46	1.51	1.33	1.60		
Pseudo-bonds	NH_i-CO_i	$C_{\alpha,i}$-NH_{i+1}	CO_i-$C_{\alpha,i+1}$	NH_i-R_i	$C_{\alpha,i}$-$C_{\alpha,i+1}$	CO_i-R_i
	2.45	2.41	2.45	2.50	3.80	2.56

where d_{ij} is the distance between beads i and j, d_{ij}^{HS} is the hard-sphere diameter, d_{ij}^{SW} is the square-well interaction distance and ε_{ij} is the square-well depth. For interactions between side-chain beads, the 3 functional parameters (d_{ij}^{HS}, d_{ij}^{SW} and ε_{ij}) have specific values for each pair of interacting side-chain beads i and j. For hard-sphere repulsion interactions we use the Lorenz-Berthelot combining rule to calculate d_{ij}^{HS} from the beads d^{HS}. As side-chain diameters are only defined for side-chain-side-chain interactions, we use their self-interaction diameter for side-chain-backbone interactions. The self-interaction value of $d_{i,j}^{HS}$ and $d_{i,j}^{SW}$ are shown in Table 2.

For the formation of hydrogen bonds between NH and CO beads additional conditions, next to being within square-well interaction distance $d_{ij}^{SW} = 4.5$ Å, have to be satisfied. Firstly, both beads considered for the hydrogen bond formation are not already involved in another hydrogen bond, and secondly there is an angle constraint between the N–H and the C–O vector.

In the model described up to this point, beads in close proximity along the chain will overlap in a way that prevents the formation of certain protein structures found in nature. To solve this shortcoming, *squeeze parameters* are introduced, which reduce the effective diameters of beads in close proximity along the chain. There are squeeze parameters for 10 different bead interactions. These parameters applied to side chain beads are specific for each amino acid and the glutamine parameters are shown in Table 3. For a detailed description of hydrogen bond formation as well as squeeze parameter implementation in the PRIME20 model we refer to the following Refs. [42,43].

The energy scale in the model is defined by the hydrogen bond strength $\varepsilon_{HB} = -1$. Side-chain interaction energies are given relative to ε_{HB} (see ε in Table 2). Physical energies E' and temperatures T' can be retrieved from the reduced quantities (E and T) by assigning a value to ε_{HB}: $E' = \varepsilon_{HB}E$ and $T' = \varepsilon_{HB}T/k_B$.

Both, the coarse-grained MC as well as the MD approach are established techniques, which can be applied separately for the investigation of the polypeptide aggregation. The combination of these two methods requires the careful design of mutual interfaces. In the first part of the manuscript, we will present a possible pathway to transfer coarse-grained structures of two polyglutamine strands into all atom geometries by a general applicable protocol. In the second part, we will start from the converted all atom structures to perform molecular dynamics simulations and discuss the stability of the P20/SAMC structures. The importance and relevance of establishing protocols for the back- and forth-conversion of structures between the coarse-grained model and all atoms MD simulation was already shown in various applications, especially for biomolecular and micellar systems.[44–58]

Table 2. Bead diameters and square-well parameters of PolyQ in PRIME20. Sizes in Å.

	NH	C_α	CO	R
d^{HS}	3.3	3.7	4.0	3.6
d^{SW}	4.5	–	4.5	6.6
ε	−1.000	–	−1.000	−0.080

Table 3. Squeeze parameters and resulting reduced bead diameters for backbone bead interactions and interactions involving a polyglutamine side chain. Sizes in Å.

Interactions	$C_{\alpha,i}$-CO_{i+1}	$C_{\alpha,i}$-NH_{i-1}	CO_i-NH_{i+2}	NH_i-NH_{i+1}	CO_i-Co_{i+1}
original *d*	3.85	3.50	3.65	3.30	4.00
squeeze factor	1.1436	0.88	0.87829	0.8	0.7713
squeezed *d*	4.40286	3.08	3.2057585	2.64	3.0852
Interactions	$C_{\alpha,i-1}$-R_i	CO_{i-1}-R_i	NH_{i+1}-R_i	$C_{\alpha,i+1}$-R_i	CO_{i-2}-R_i
original *d*	3.65	3.8	3.45	3.65	3.8
squeeze factor	1.407	1.089	1.158	1.387	1.316
squeezed *d*	5.134	4.139	3.996	5.062	5.000

3. Results

3.1. Conversion of coarse-grained into all atom structures

Our goal is to develop a protocol for the back-conversion of conformations obtained from the coarse-grained peptide interaction model PRIME20 into atomistic structures. The concept of our protocol is illustrated in Figure 2. The PRIME20 scheme provides simulation data which contains coordinates for the backbone carbon and nitrogen atoms, as well as the center of mass (COM) coordinates of the side chain residues of the peptide. The illustration in Figure 3 indicates these with red circles. The atoms labeled with green circles are not provided by the PRIME20 scheme, and the center of mass of a residue R of course lacks the coordinates for the individual atoms.

The concept of our back-mapping scheme is to derive the coordinates of the carbonyl oxygens and the nitrogen protons from the peptide backbone directly from the backbone carbon coordinates, by assuming equilibrium bond distances and a planar geometry with respect to the two adjacent backbone atoms. For the other atoms in the amino acid residues R, the coordinate of the initial carbon atom is computed in the same way, and the orientation of the residue is defined by the connection vector from the backbone C_α atom to the center of mass from the PRIME20 simulation data (see Figure 3). For the initial conversion step, we assume the molecular equilibrium conformation for the amino acid residue as such, so that the anchor point (via the center of mass) and the orientation (via the C_α-COM vector) are sufficient to reconstruct the coordinates of the full residue.

The atomic coordinates computed in this way are tentative values, which lead to considerable misalignments in the peptide structure. The most common problem is that atoms from two adjacent amino acid residues are too close to each other. However, our protocol turns out to yield reasonable values for the start of a short geometry optimization cycle, in the sense that the standard optimization algorithms are able to respond to the close-proximity-misalignments and reorient the amino acid residues away from each other by maintaining the overall peptide structure as proposed by the coarse-grained scheme. It should be noted that while the resulting atomistic peptide geometry is technically possible, it is not for granted that this conformation is locally stable from a thermodynamical perspective. The latter aspect is addressed in a second stage within our back-mapping scheme.

To grasp the structural deviation from our back-mapping method, we have calculated the root-mean square displacement (RMSD) comparing the P20/SAMC resulted structure to the geometry optimized structures for the MD simulations. Similar to the structure conversion, we only compared the N,

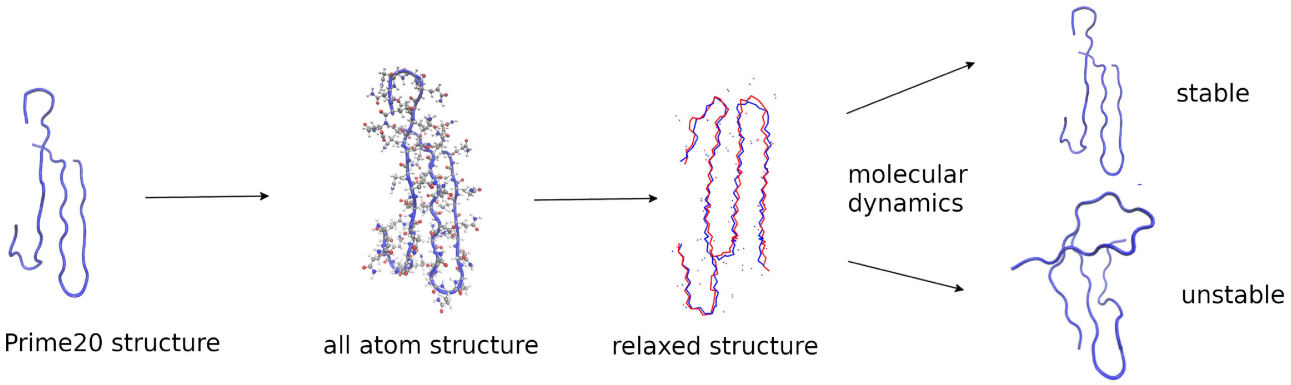

Figure 2. Visualization of the central process for the generation of data in this article.

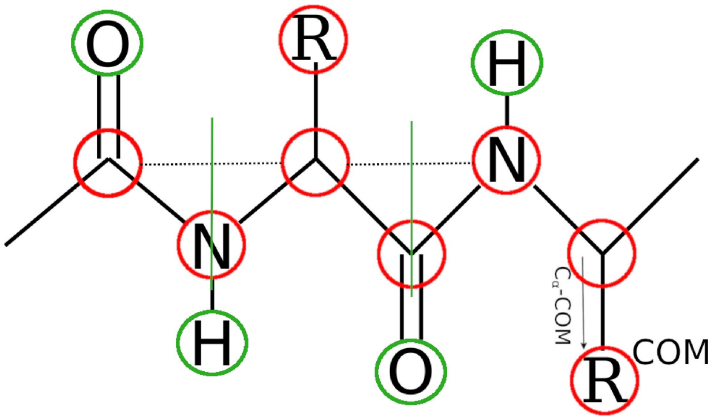

Figure 3. Scheme describing the conversion of the coarse-grained structures into all atom geometries. Red: atoms obtained from the coarse-grained PRIME20 model, green: atoms added by simple geometric considerations.

$C_{Carbonyl}$, C_α atoms and the sidechain COM to the respective beads of the PRIME20 model. The RMSD data for all conformations is given in Table 4, as well as visual examples for change in structure caused by the energy optimization in Figure 4. We observe relatively similar and actually quite small displacements for all calculated structures, which on the one hand shows the back-mapping technique is reliable, on the other hand shows the PRIME20 structures are near a local energy minimum instead of being geometries that will relax considerably upon energy minimization.

3.2. Relaxed geometry of the all atom peptide structures

In order to cover a broad variety of systems for the back-mapping protocol, we generated a series of 21 solvated peptide dimers $(Glu_n)_2$ with varying length (n ∈ 14, 16, 18, 20, 22, 24, 26, 28, 36) within the P20/SAMC simulation framework. For each of the nine dimer systems, up to four conformations were selected from the P20/SAMC scheme for inverse coarse-graining. In order to test the back-mapping protocol, we picked generally low-energy conformation along with a few extremely low-energy conformations, so that both "easy", in the sense of typical aggregated peptide conformations, and "difficult" conformations, in the sense of very uncommon peptide features, were processed and back-mapped to atomistic structures. The terms "easy" indicates typical aggregated conformations as was determined via analysis of the hydrogen bond contact probabilities within the PRIME20 scheme. "Difficult" conformations are of the lowest energy found in single P20/SAMC simulation runs. This makes them more likely to contain sterically demanding atomistic features such as highly rigid hydrogen bond networks. A complete list of the investigated peptide dimers including their chain lengths and energies calculated within the P20/SAMC model is given in Table 4. In the table, the canonical expectation value $\langle U\rangle_T = 1/Z_U \sum_U U g(U) e^{-\beta U}$ of the configuration energy at room temperature is given. It is derived from the density of states $g(U)$ of the PolyQ systems. One can see, that $\langle U\rangle_T$ increases when going to systems of longer chain lengths. Performing MD simulations at room temperature on conformations of configuration energies far below $\langle U\rangle_T$ has implications on the expected mechanical stability in MD. The further away from $\langle U\rangle_T$ a configurations energy is, the more likely it will be unstable in the MD simulation. However, for the MD simulation run lengths of 10 ns (see SI), possible meta-stability in configurations can be found.

In the next step, we added explicit solvent molecules to the all atom structures and performed geometry optimizations.

Table 4. Overview of all calculated systems with the canonical expectation value of the configurational energy $\langle U\rangle_T$ at room temperature. Furthermore, including MD energy properties, visual stability and a comparison of visual and ACF_{hb} stability. Green color shows agreement between both, red disagreement and black cases, where visual inspection was not fully distinguishable/ accessible (n.a.), for unstable (x) and stable (o) structures.

System	$\langle U\rangle_T (T = 300\ K)$	Visual Stability	Stability hb_{inter}	RMSD
−23.92	−1.68	unstable	x	0.94
−18.88	−3.01	stable	o	0.79
−21.88	−3.01	half-stable (n.a.)	x	0.74
−25.00	−3.40	stable	o	0.65
−26.36	−3.40	stable	o	0.78
−30.00	−3.40	unstable	x	0.83
−27.00	−8.05	stable	o	0.72
−30.00	−8.05	half-stable (n.a.)	o	0.92
−29.36	−20.48	stable	o	0.76
−30.00	−20.48	stable	o	0.57
−37.48	−20.48	stable	o	0.90
−30.32	−19.86	stable	o	0.72
−31.40	−19.86	stable	o	0.76
−33.92	−19.86	stable	o	0.73
−44.04	−19.86	unstable	x	0.90
−35.00	−31.48	stable	o	0.79
−44.96	−31.48	unstable	x	0.85
−36.12	−20.63	unstable	x	0.72
−38.24	−20.63	unstable	o	0.76
−40.00	−37.23	stable	o	0.75
−40.00	−37.23	stable	o	0.67

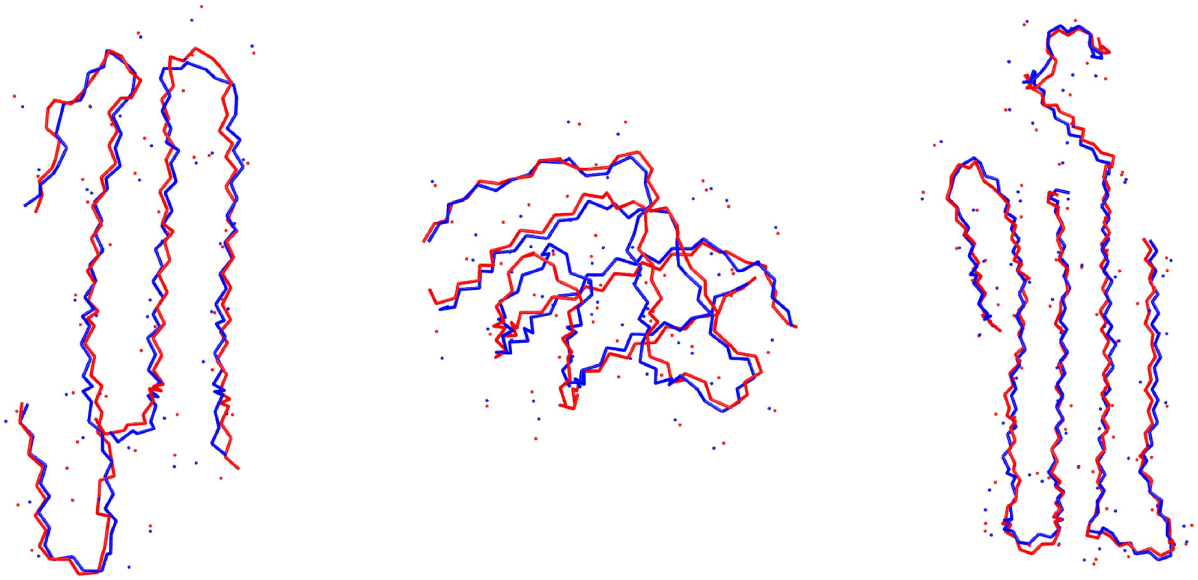

Figure 4. Comparison of the PRIME20 structure (red) and the backmapped, geometry optimized all atom structures (blue) used as starting point for the MD simulation. All PRIME20 beads are visualized and their respective MD atoms: C_α, N, C and the sidechain centre of mass COM_{Side}).

To this purpose the all atom peptide dimer structures were centered in a 4 nm×4 nm×4 nm simulation box, and water molecules were added until a density between 1.00–1.07 g/cm^3 was reached. After solvating the peptide dimers, we performed geometry optimizations of the all atom structures using the program package GROMACS and the force field AMBER03.[41]

The calculation of force field energies was successful, in the sense that all calculations converged rapidly, for each of the converted all atom systems and the atomic positions of the peptide dimers were relaxed with respect to the minimization of the energy. For the comparison to the initial coarse grained P20/SAMC structures, the geometry optimized all atom peptide dimer structures were again reconverted into the coarse grained structures. The root mean square deviation of the coarse grained peptide dimers between before and after geometry optimization is given in Table 4.

In Figure 4, we show for three selected examples initial coarse-grained structures from the P20/SAMC calculations and the relaxed and back mapped all atom structures. The initially obtained coarse-grained peptide dimer structures and the geometry optimized all atom peptide dimer geometries are in good agreement.

In conclusion, both the back-conversion of the coarse-grained peptide dimer structures into atomic configurations and the subsequent local geometry optimizations with explicit aqueous solvation were successful and resulted in structurally acceptable conformations with a very good structural similarity to the original (coarse-grained) configurations.

All individual steps within our backmapping protocol are summarized in Figure 2. Our approach can be used for the automatic generation of fully solvated initial structures for all atom molecular dynamics simulations from coarse-grained P20/SAMC model geometries. In the future, we plan to extend our approach to peptide structures formed by other amino acids than glutamine.

3.3. Molecular dynamics simulation of initial dipeptide configurations obtained from the P20/SAMC calculations

We have visually inspected the peptide dimer structures provided by the P20/SAMC sampling before and after the molecular dynamics relaxations in order to characterize the structures on an empirical level as “stable” or “unstable”. We have focused on the strength of structural changes within the stronger hydrogen-bonded central regions of the peptides. The hydrogen bonding can be either at the peptide backbone level (NHO, both intramolecular and intermolecular, corresponding to beta-hairpin structures and collinear peptide strand conformations, respectively) or between amino acid sidechains (mainly intermolecular), see Figure 5.

Regarding the visual discrimination between “stable” and “unstable”, we have started by defining a “core” and a “peripheral” part of the dimer (green and yellow shaded areas in Figure 5). The core region is the part that contains direct peptide contacts, and would be the nucleation area for further aggregation of additional peptides. The peripheral regions are peptide segments that are fully solvated and/or localized outside the direct attachment region for additional peptides. The classification “stable” vs. “unstable” is now applied based on the structural integrity of the core region, i.e. its persistence after the short MD simulation.

The empirical classification of all 21 peptide dimer conformations in terms of “stable” or “unstable” is given in Table 4. The atomic coordinates of the first and last frame are also reported as raw data in the SI. A qualitative observation from

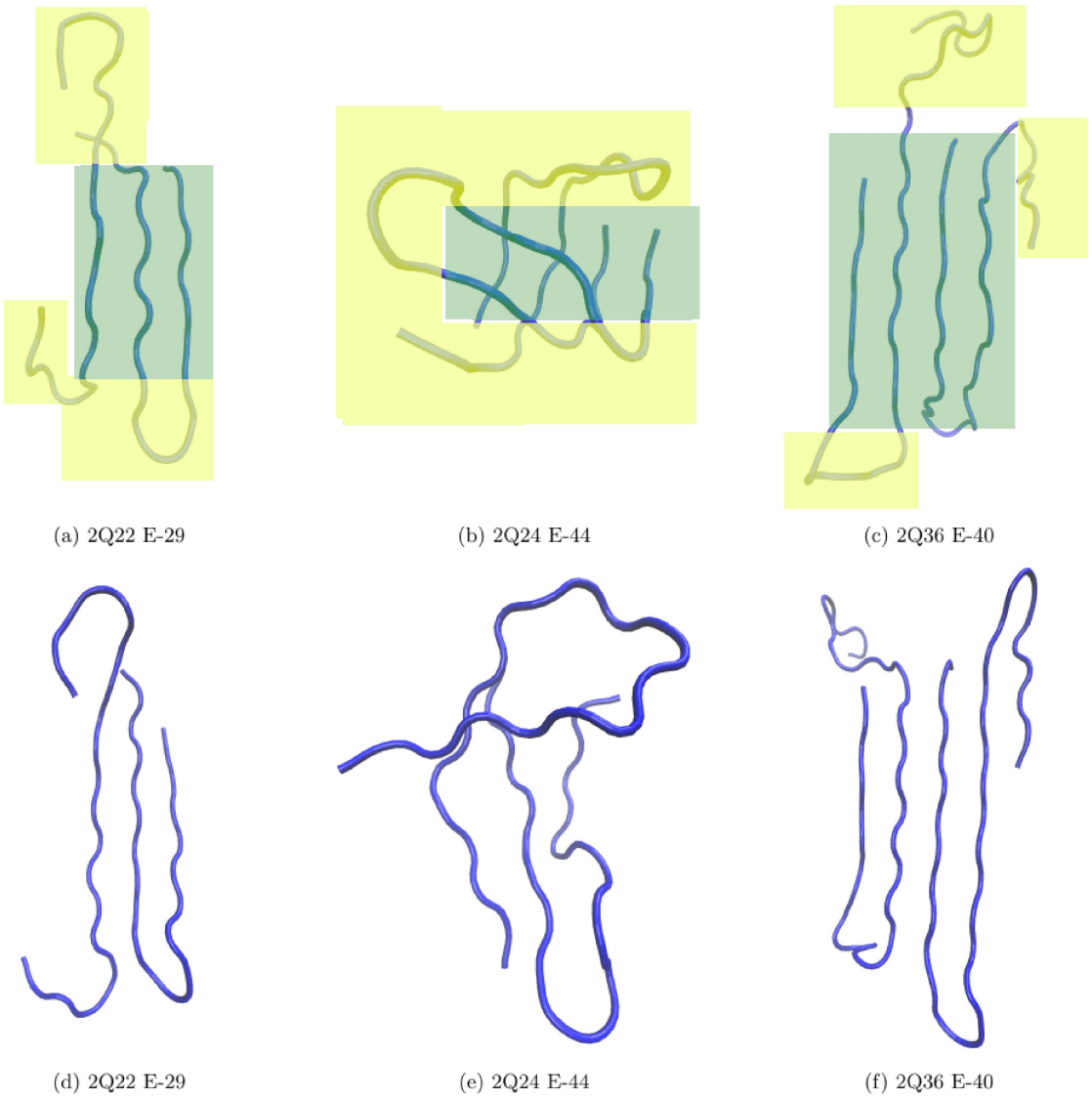

Figure 5. Visualization of the peptides at the start (a-c) and end (d-f) of the simulations.

this first visual classification step is that if the initial P20/SAMC structure incorporates parallel peptide strands (either intramolecular, in a hairpin conformation, or intermolecular by just parallel backbone segments) connected through hydrogen bonds, then the structure of the aggregated peptide strands remained stable during the atomistic relaxation run. Examples for such configurations are the structures Figure 5(a)/(d) and (c)/(f).

Another empirical observation from our visual inspections is that as soon as one of the coarse-grained peptides adopts a quasi-spherical shape, the dimer interaction is inhibited and the resulting configuration turns out to be unstable under molecular dynamics equilibration. A typical example for such a structure is represented in Figure 5(b)/(e), where both peptides lose their initial P20/SAMC conformation (Figure 5(b)) after relaxation 5(e). A possible explanation for this observation is the implicit treatment of solvation within the PRIME20 model. More compact (i.e. rather spherical) conformations tend to maximize the intramolecular contacts of the peptide and to minimize the surface area towards the solvent. Within the explicit solvation used for the atomistic molecular dynamics simulations, the enthalpic benefit of peptide-solvent interactions is stronger, and thus the tendency to form compact structures is weaker. Independently of the solvation influence, the P20/SAMC calculations produces very low-energy structures, which are associated to exist at lower temperatures. However, the model P20 model is optimized for proteinogenic structures at room temperature, which could lead to unphysical structures at the low temperature range. This behaviour is reflected in comparing the $< U >_T$ to the actual system energy. In most cases, only the lowest energy was not stable during the MD simulations.

3.4. Hydrogen bond dynamics

As a complementary perspective regarding the dimer stability, we now look for a physical property that can be quantified a bit better compared to a mere visual inspection. We chose to look at the intermolecular hydrogen bonds between the peptides, in particular considering their temporal stability. Therefore, we calculated the autocorrelation function of all intermolecular hydrogen bonds and its time evolution. This function indicates how many of the initial hydrogen bonds (at $t = 0$) have remained intact after a given time (e.g. during the full simulation of 10 ns). The data is shown in Figure 6 for a selection of dimer configurations. Each line corresponds to a given starting structure from the PRIME20 sampling, and those structures that have been visually characterized as "stable" are represented as full lines, while "unstable" structures are shown as dashed lines.

While there is a certain amount of numerical noise, a plateau value is reached for most of the dimers after around 3 ns. Afterwards, we observe fluctuations around those plateaus, which corresponds to hydrogen bond breaking and reformation processes. Interestingly, our inital empirical assessment in terms of stability is fully confirmed by this semi-quantitative analysis: all "stable" structures yield a highly preserved hydrogen bond network (i.e. little decay of the autocorrelation function), while the "unstable" structures all exhibit a rapid decay and large fluctuations. The average values of the autocorrelation functions are listed in Table 5.

Table 5. Averages of all unstable or stable autocorrelation functions of the intermolecular hydrogen bonds.

Stability	Average
stable	0.84
unstable	0.51

Here, we have looked at the hydrogen bonding autocorrelation functions merely with a qualitative eye, as a complementary semi-quantitative tool in addition to the classification of stable/unstable structures as discussed above. We have explicitly avoided to fit the hydrogen bond autocorrelation functions shown in Figure 6 to exponentials (yielding a numerical hydrogen bond lifetime), as we believe this would imply a quantitative relaxation time measure, which, however, is simply not reflected by the raw data (to our belief).

Additionally, we have also calculated the same autocorrelation functions but for intramolecular and intra/inter hydrogen bonds combined (all data given in the Supporting Information). However, with our focus on the peptide dimer stability, the intermolecular hydrogen bonds had most significance.

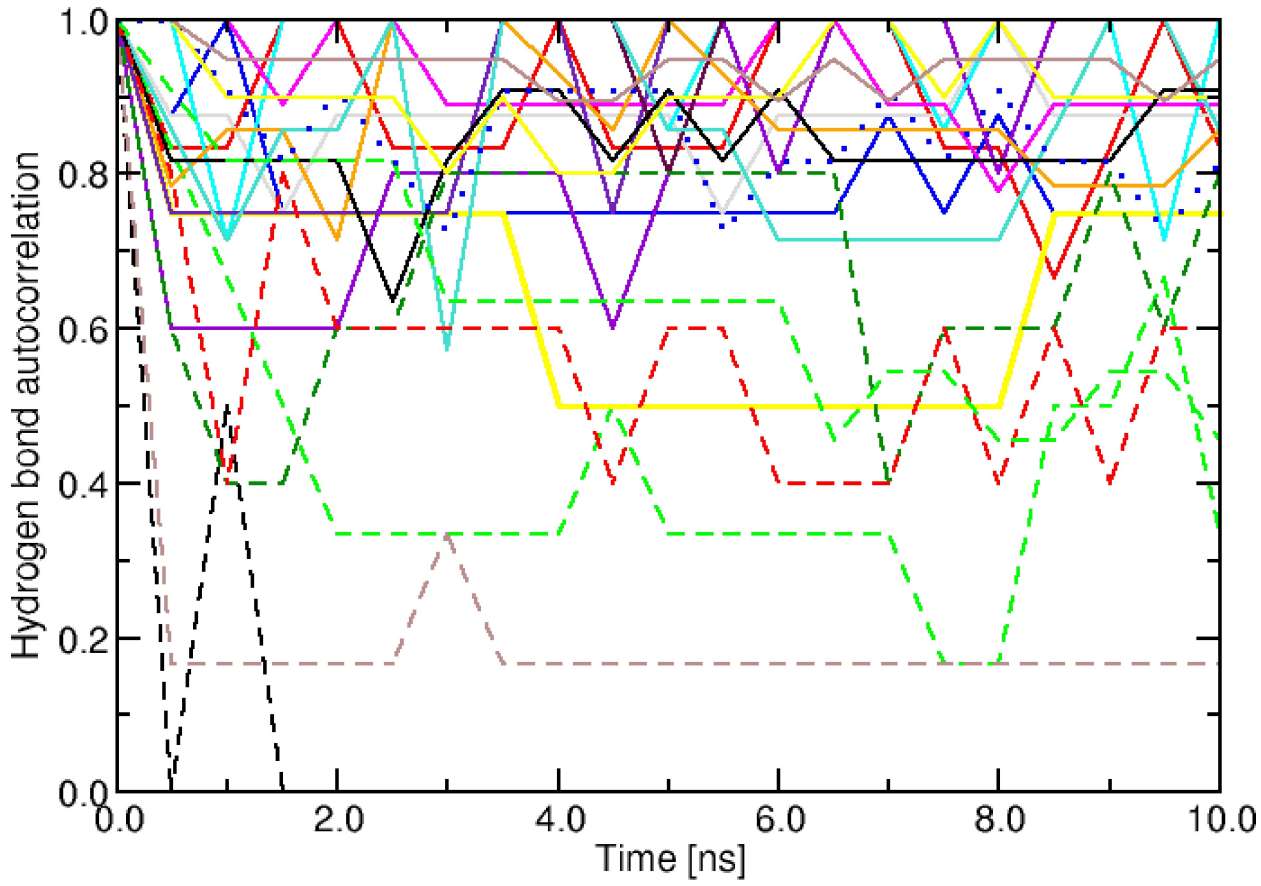

Figure 6. Intermolecular backbone hydrogen bonding autocorrelation function (percentage of hydrogen bonds of initial structure that are preserved) over the whole MD simulation time for all simulated peptides. Straight lines resemble stable structures, dashed lines unstable structures and dotted lines show the structure, which is visually unstable but hydrogen bonding suggests a stable structure.

3.5. Discussion and Outlook

The overall picture of our simulations confirms the reliability of the P20/SAMC method. From the thermodynamically most representative conformations generated from the P20/SAMC approach, physical meaningful configurations remained stable during the MD simulations, while unphysical peptide dimer structures were unstable. A particularly characteristic shape of highly unstable structures resembles a sphere, and these structures can be caused by the specific way in which the solvent is represented within the PRIME20 model. Instead of a chemically specific solvent interaction (which would depend on the actual chemical environment, i.e. whether there are actual particules in the vicinity), the PRIME20 model incorporates solvent effects by reducing the interaction strength between actual particles. As an example, the energetic strength of a hydrogen bond is chosen considerably lower than the normal chemical value of around 20 kJ/mol. Since the side chains are normally more solvent exposed than the peptide backbone, those hydrogen bonds carry an even lower energy contribution. The stable peptide aggregates formed in most cases extended hydrogen bond patterns between parallel peptide strands. While the categorization of coarse-grained structures into "stable" and "unstable" types is nontrivial from a quantitative point of view, it turned out that a more qualitative perspective is (in our opinion) sufficient to capture whether the conformation are essentially chemically reasonable.

Thus, the combination of the coarse-grained MC and MD simulations is suited to identify and investigate the local dynamics of stable aggregates of peptide strands. The P20/SAMC model allows to sample efficiently the entire phase space, while the all atom molecular dynamics simulations enable the probing of the geometric as well as the dynamic properties of the local minimum energy structures. As a side effect, molecular dynamics helps to validate the reliability of the P20/SAMC low energy structures by exclusion of unstable geometries from further analysis.

In the next step, we plan to extend our protocol for the conversion of coarse-grained into all atom structures towards peptides composed of other amino acids compared to glutamine.

The algorithm for the conversion of coarse-grained into all atom structures could be applied to all 21 PRIME20 polyglutamine structures without any changes. Subsequent relaxation of the coordinates of the fully solvated peptide dimer structures was possible using the GROMACS program package. The resulting geometry optimized structures were in good agreement with the initial P20/SAMC geometries. This is in particular remarkable, because the transferability to an all atom approach was not explicitly intended during the development of the PRIME20 model.

This work can be seen as the first step towards the overarching goal of improving the understanding of peptide aggregation using the PRIME20 model. In this development step, we have demonstrated how to convert coarse-grained P20/SAMC structures into all atom structures for MD simulations. Although the back mapping was possible, the resulting coarse-grained structures could not be used for energy calculations within the PRIME20 model. The reason being the use of square-well potential and many cutoff values for inter- and intramolecular distances that have to be fulfilled by a peptide geometry to be a valid PRIME20 structure. Fluctuating configurations from finite temperature molecular dynamics simulations do often not fulfill these strict cut off criterions.

4. Conclusions

We have designed and implemented a reverse coarse-graining approach for the back-mapping of atomistic structures into conformations obtained from a united-atom scheme (PRIME20 approach) that is suitable for large-scale Monte-Carlo based conformational sampling. The reverse coarse-graining method is straightforward to implement for regular proteins/peptides and allows for a subsequent exploitation of atomistic molecular structures generated from the extensive conformational search done at the coarse-grained level.

We have validated the approach with a series of shorter peptide dimers via a conformational stability analysis using molecular dynamics simulations. It turns out that the majority – but not all – of the conformations delivered from the large-scale conformational sampling are "good" structures that remain stable for at least 10 ns of simulation. As a side result, we have found that a visual empirical assessment of the conformations yields stability estimates which are in good agreement with a more quantitative analysis in terms of the persistence of the intermolecular hydrogen bond network. All structures that were visually assessed as "unphysical" turned out to be unstable during the molecular dynamics simulations.

Our approach provides a further layer of atomistic detail to the coarse grained simulation of structurally challenging systems, combining the large-scale phase space sampling capability of the coarse-grained Monte Carlo method with the better accuracy and the atomistic resolution available at the molecular dynamics level.

Computational Details

We used the PRIME20 model to perform coarse grained Monte-Carlo simulations for dimers of polyglutamine with chains length n between 14 and 36 amino acids. The simulation method we used is the Stochastic Approximation Monte Carlo (SAMC) method. It is an advanced flat-histogram Monte Carlo method which aims for a flat visitation histogram of energy states. In achieving this, it avoids getting stuck in local energy minima as can be the case with conventional Monte Carlo methods. SAMC achieves the even visitation of energy states by approximating the density of states (DOS) $g(U)$ with respect to the potential energy U. The DOS describes the number of states in the system that belong to a given energy interval $[U, U+\Delta U]$. It then uses the DOS in its acceptance criterion: for an SAMC move from configuration x with the energy $U(x)$ to configuration x' with the energy $U(x')$, the move is accepted with the probability $\min(1, \tilde{g}(U(x'))/\tilde{g}(U(x)))$. $\tilde{g}(U)$ is the current estimate for the DOS. After the move is rejected or accepted, $\tilde{g}(U)$ is updated according to $\tilde{g}(U(x_{new})) = \tilde{g}(U(x_{new})) + \gamma_t$, where $x_{new} = x'$ if the move was accepted and $x_{new} = x$ if the move was

rejected. The modification factor γ_t goes to 0, for time $t \to \infty$. t is measured in MC steps. Additional conditions have to be met in order for the DOS to converge.[59–61] After a sufficiently accurate $g(U)$ was obtained, further MC runs with a fixed DOS were performed. With the flat visitation histogram of energy states, snapshots at various energies were collected in multiple simulation runs of 10^9 MC steps.

Four different MC move types are used in the SAMC simulations. A local displacement move, which moves a single bead in a randomly chosen direction by a random distance, with a maximal displacement of 0.02 Å. A pivot rotation move, which randomly chooses a residue and rotates either its Ψ or Φ angle by a random amount and direction. Furthermore, two moves are implemented to manipulate the relative position of the two chains in the system: a whole-chain rotation and a whole-chain translation move. After every move, the new configuration must be in agreement with the PRIME20s constraints on bond-lengths and excluded volumes. Similar to already successful calculations,[62] we simulated polyglutamine dimer systems with chain lengths $N \in$ (14, 16, 18, 20, 22, 24, 26, 28, 36). N refers to the number of residues in a single chain. For shorter chains ($N \in$ (14, 16, 18, 20, 22, 24, 26)) the cubic simulation box was of length $L = 112.5$ Å and for longer chains ($N \in$ (28, 36)) the box was of length $L = 150$ Å. The simulation box was periodic in all directions. This translates to a milli-molar concentration, which is close to in vitro experiments on polyglutamine aggregation.

The coarse-grained low-energy structures resulting from the PRIME20 simulations listed in Table 4 were translated into all-atom structures with both termini charged and were directly suitable for calculations. These structures were then explicitly solvated using the standard GROMACS[63,64] solvation tool; it should be noted that this solvation algorithm resulted in varying numbers of water molecules for different geometries. After an initial energy minimization (emtol = 100; emstep = 0.1; niter = 20) for all atoms, a 10 ns NVT MD simulation with a 0.5 fs time step was performed at 300 K using velocity rescaling with 0.1 ps time constant, Lincs 4th order constraint[65] for covalent hydrogen bonds and the AMBER03[41] force field, while water interactions were represented by the TIP3P[66] water model. The Verlet cutoff-scheme and periodic-boundary conditions were used, and electrostatics were calculated with PME using potential-shift-Verlet for the coulomb modifier.

The energies and radii of gyration R_G were calculated by GROMACS tools,[63,64] and visualization was performed with VMD.[67] The first 2 ns were treated as initial equilibration and not used for GROMACS analysis. The hydrogen bond autocorrelation functions were calculated with a python script; the persistence of all hydrogen bonds determined in the initial structure was checked every 1 ns along the trajectory, by means of a combined distance/angle criterion. Note that we explicitly checked for temporary ruptures of hydrogen bonds, i.e. the autocorrelation function can increase again if a hydrogen bond is only shortly broken.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 189853844 – TRR 102. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Reverse mapping of coarse grained polyglutamine conformations from PRIME20 sampling at https://github.com/thomascookies/Reverse-mapping-of-coarse-grained-polyglutamine-conformations-from-PRIME20-sampling, reference number 0.

Keywords: backmapping · coarse-grained · molecular dynamics simulations · monte carlo simulation · peptide secondary structure · PRIME20

[1] M. Huntley, G. B. Golding, *J. Mol. Evol.* **2000**, *51* (2), 131.
[2] C. E. Pearson, R. R. Sinden, *Curr. Opin. Struct. Biol.* **1998**, *8* (3), 321.
[3] R. H. N. Kalaria, S. I. Harik, *J. Neurochem.* **1989**, *53* (4), 1083.
[4] H. Y. Zoghbi, H. T. Orr, *Annu. Rev. Neurosci.* **2000**, *23*, 217.
[5] C. M. Lill, C. Klein, *Nervenarzt* **2017**, *88* (4), 345.
[6] C. Soto, *FEBS Lett.* **2001**, *498* (2–3), 204.
[7] P. H. Nguyen, A. Ramamoorthy, B. R. Sahoo, J. Zheng, P. Faller, J. E. Straub, L. Dominguez, J. E. Shea, N. V. Dokholyan, A. de Simone, B. Ma, R. Nussinov, S. Najafi, S. T. Ngo, A. Loquet, M. Chiricotto, P. Ganguly, J. McCarty, M. S. Li, C. Hall, Y. Wang, Y. Miller, S. Melchionna, B. Habenstein, S. Timr, J. Chen, B. Hnath, B. Strodel, R. Kayed, S. Lesné, G. Wei, F. Sterpone, A. J. Doig, P. Derreumaux, *Chem. Rev.* **2021**, *121* (4), 2545.
[8] A. M. Morris, M. A. Watzky, R. G. Finke, *Biochim. Biophys. Acta Proteins Proteomics* **2009**, *1794* (3), 375.
[9] J. A. Housmans, G. Wu, J. Schymkowitz, F. Rousseau, *FEBS J.* **2023**, *290* (3), 554.
[10] S. Navarro, S. Ventura, *Curr. Opin. Struct. Biol.* **2022**, *73*, 102343.
[11] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, D. E. Shaw, *Annu. Rev. Biophys.* **2012**, *41* (1), 429.
[12] D. Rosenberger, M. Hanke, N. F. Van der Vegt, *Eur. Phys. J. Spec. Top.* **2016**, *225* (8–9), 1323.
[13] H. J. Risselada, S. J. Marrink, *Phys. Chem. Chem. Phys.* **2009**, *11* (12), 2056.
[14] E. Brini, V. Marcon, N. F. Van der Vegt, *Phys. Chem. Chem. Phys.* **2011**, *13* (22), 10468.
[15] D. Reith, M. Pütz, F. Müller-Plathe, *J. Comput. Chem.* **2003**, *24* (13), 1624.
[16] A. P. Lyubartsev, A. Laaksonen, *Phys. Rev. E* **1995**, *52* (4), 3730.
[17] S. Izvekov, G. A. Voth, *J. Phys. Chem. B* **2005**, *109* (7), 2469.
[18] J. W. Mullinax, W. G. Noid, *J. Phys. Chem. C* **2010**, *114* (12), 5661.
[19] M. Karplus, J. Kuriyan, *Proc. Natl. Acad. Sci. USA* **2005**, *102* (19), 6679.
[20] M. Bendahmane, K. P. Bohannon, M. M. Bradberry, T. C. Rao, M. W. Schmidtke, P. S. Abbineni, N. L. Chon, S. Tran, H. Lin, E. R. Chapman, J. D. Knight, A. Anantharam, *Mol. Biol. Cell* **2018**, *29* (7), 834.
[21] S. Sharma, M. Lindau, *Proc. Natl. Acad. Sci. USA* **2018**, *115* (50), 12751.
[22] R. M. Henry, C. H. Yu, T. Rodinger, R. Pomés, *J. Mol. Biol.* **2009**, *387* (5), 1165.
[23] L. K. Scarbath-Evers, S. Jähnigen, H. Elgabarty, C. Song, R. Narikawa, J. Matysik, D. Sebastiani, *Phys. Chem. Chem. Phys.* **2017**, *19* (21), 13882.
[24] F. Hoffmann, J. Adler, B. Chandra, K. R. Mote, G. Bekçioğlu-Neff, D. Sebastiani, D. Huster, *J. Phys. Chem. Lett.* **2017**, *8* (19), 4740.
[25] I. Kurisaki, S. Tanaka, *Phys. Chem. Chem. Phys.* **2022**, *24* (17), 10575.
[26] M. S. Barhaghi, B. Crawford, G. Schwing, D. J. Hardy, J. E. Stone, L. Schwiebert, J. Potoff, E. Tajkhorshid, *J. Chem. Theory Comput.* **2022**, *18* (8), 4983.
[27] H. J. Woo, A. R. Dinner, B. Roux, *J. Chem. Phys.* **2004**, *121* (13), 6392.
[28] I. Y. Ben-Shalom, C. Lin, T. Kurtzman, R. C. Walker, M. K. Gilson, *J. Chem. Theory Comput.* **2019**, *15* (4), 2684.
[29] M. S. Bodnarchuk, M. J. Packer, A. Haywood, *ACS Med. Chem. Lett.* **2020**, *11* (1), 77.
[30] G. A. Ross, E. Russell, Y. Deng, C. Lu, E. D. Harder, R. Abel, L. Wang, *J. Chem. Theory Comput.* **2020**, *16* (10), 6061.
[31] S. Pylaeva, A. Böker, H. Elgabarty, W. Paul, D. Sebastiani, *ChemPhysChem* **2018**, *19* (21), 2931.
[32] D. Reith, M. Pütz, F. Müller-Plathe, *J. Comput. Chem.* **2003**, *24* (13), 1624.
[33] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, H. C. Andersen, *J. Chem. Phys.* **2008**, *128*, 24.
[34] E. Brini, N. F. Van der Vegt, *J. Chem. Phys.* **2012**, *137*, 154113.

[35] P. Ganguly, N. F. A. Van der Vegt, *J. Chem. Theory Comput.* **2013**, *9* (12), 5247.
[36] L. C. Jacobson, R. M. Kirby, V. Molinero, *J. Phys. Chem. B* **2014**, *118* (28), 8190.
[37] J.-w. Shen, C. Li, N. F. Van der Vegt, C. Peter, *J. Chem. Theory Comput.* **2011**, *7* (6), 1916.
[38] M. Langeloth, T. Sugii, M. C. Böhm, F. Müller-plathe, *J. Chem. Phys.* **2015**, *143*, 243158.
[39] S. Jain, S. Garde, S. K. Kumar, *Ind. Eng. Chem. Res.* **2006**, *45* (16), 5614.
[40] C.-C. Fu, P. Kulkarni, S. Shell, G. Leal, *J. Chem. Phys.* **2012**, *137*, 164106.
[41] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, P. Kollman, *J. Comput. Chem.* **2003**, *24* (16), 1999.
[42] A. Böker, W. Paul, *J. Phys. Chem. B* **2022**, *126* (38), 7286.
[43] A. Böker, Ph.D. thesis, Martin-Luther-University Halle-Wittenberg, **2019**.
[44] J. Peng, C. Yuan, R. Ma, Z. Zhang, *J. Chem. Theory Comput.* **2019**, *15* (5), 3344.
[45] S. D. Peroukidis, D. G. Tsalikis, M. G. Noro, I. P. Stott, V. G. Mavrantzas, *J. Chem. Theory Comput.* **2020**, *16* (5), 3363.
[46] M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, C. L. Brooks III, *Proteins Struct. Funct. Genet.* **2000**, *41* (1), 86.
[47] B. Hess, S. Leo, N. Van der Vegt, K. Kremer, *Soft Matter* **2006**, *2* (5), 409.
[48] A. P. Heath, L. E. Kavraki, C. Clementi, *Proteins Struct. Funct. Bioinf.* **2007**, *68* (3), 646.
[49] C. Peter, K. Kremer, *Soft Matter* **2009**, *5* (22), 4357.
[50] S. M. Gopal, S. Mukherjee, Y.-m. Cheng, M. Feig, *Proteins Struct. Funct. Bioinf.* **2009**, *78* (5), 1266.
[51] A. J. Rzepiela, L. V. Schäfer, N. Goga, H. J. Risselada, A. H. De Vries, S. J. Marrink, *J. Comput. Chem.* **2010**, *31* (6), 1333.
[52] P. J. Stansfeld, M. S. P. Sansom, *J. Chem. Theory Comput.* **2011**, *7* (4), 1157.
[53] P. Brocos, P. Mendoza-Espinosa, R. Castillo, J. Mas-Oliva, Á. Piñeiro, *Soft Matter* **2012**, *8* (34), 9005.
[54] T. A. Wassenaar, K. Pluhackova, R. A. Bo, S. J. Marrink, D. P. Tieleman, *J. Chem. Theory Comput.* **2014**, *10* (3), 676.
[55] L. E. Lombardi, M. A. Martí, L. Capece, *Bioinformatics* **2016**, *32* (8), 1235.
[56] M. Machado, S. Pantano, *Bioinformatics* **2016**, *32* (10), 1568.
[57] S. Poblete, S. Bottaro, G. Bussi, *Biochem. Biophys. Res. Commun.* **2018**, *498* (2), 352.
[58] M. Shimizu, S. Takada, *J. Chem. Theory Comput.* **2018**, *14* (3), 1682.
[59] F. Liang, *J. Stat. Phys.* **2006**, *122* (3), 511.
[60] F. Liang, C. L. Liu, R. J. Carroll, *J. Am. Stat. Assoc.* **2007**, *102* (477), 305.
[61] T. Shakirov, S. Zablotskiy, A. Böker, V. Ivanov, W. Paul, *Eur. Phys. J. Spec. Top.* **2017**, *226* (4), 705.
[62] C. Lauer, W. Paul, *Macromol. Theory Simul.* **2023**, *2200075*, 1.
[63] H. J. C. Berendsen, D. Van der Spoel, R. Van Drunen, *Comput. Phys. Commun.* **1995**, *91* (1–3), 43.
[64] D. Van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. Berendsen, *J. Comput. Chem.* **2005**, *26* (16), 1701.
[65] B. Hess, *J. Chem. Theory Comput.* **2008**, *4* (1), 116.
[66] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79* (2), 926.
[67] W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graphics* **1996**, *14* (1), 33.

Manuscript received: November 10, 2023
Revised manuscript received: February 1, 2024
Accepted manuscript online: February 5, 2024
Version of record online: March 28, 2024