Coarse-grained molecular dynamics simulation of transport through the nuclear pore complex
Ghavami, Ali

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2014

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Chapter 2

Coarse-grained potentials for local interactions in unfolded proteins

Abstract

Recent studies have revealed the key role of natively-unfolded proteins in many important biological processes. In order to study the conformational changes of these proteins, a one-bead-per-amino-acid coarse grained (CG) model is developed and a method is proposed to extract the potential functions for the local interactions between CG beads. Experimentally obtained Ramachandran data for the coil regions of proteins are converted into distributions of pseudo-bond and pseudo-dihedral angles between neighboring alpha-carbons in the polypeptide chain. These are then used to derive bending and torsion potentials, which are residue and sequence specific. The validity of the developed model is testified by studying the radius of gyration as well as the hydrodynamic properties of chemically denatured proteins.

2.1 Introduction

Despite the classical view that a protein can attain its biological function only upon folding into a unique structure, there is increasing evidence that unfolded proteins play a key role in many important biological functions (Fink 2005, Tompa 2009). The basic functions of this class of proteins exploit the absence of a stable secondary structure in their polypeptide chain. Studies of the role of unfolded proteins in different molecular processes suggests that these proteins can be classified into four functional groups, involved in molecular recognition and signaling, protein modification, molecular assembly and entropic chain activities (such as linkers/spacers, bristles, springs) (Dunker et al. 2002, Radivojac et al. 2007, Tompa 2009). Rapid increase in our knowledge on structures and functions of unfolded proteins has disclosed that many proteins and protein domains are intrinsically unstructured and, more interestingly, their relative amount increases in more evolved and complex organisms (Tompa 2003). All of these findings indicate the importance of unfolded proteins and the necessity of developing new approaches to study them in more detail.
Although atomic-level molecular dynamics simulations have achieved a high level of reliability and accuracy, they are still not able to reach biologically interesting time and length scales due to the limitations in computational resources. The situation is even worse for unfolded proteins since by their nature they are more dynamic and therefore longer simulations are required in order to obtain statistically meaningful results. These limitations have drawn the interest of researchers towards the development of coarse-grained (CG) models to reduce the degrees of freedom and, thus, to increase the spatial and temporal domains of interest.

Available CG models can be categorized into different classes according to the level of coarse-graining, the treatment of the solvent environment, and the method used for the force field parametrization (Tozzini 2009). Among them, the CG models which describe each amino acid with one or a few beads have been successful in representing structural detail of proteins while maintaining computational efficiency (Tozzini and McCammon 2005). Coarser models have not gained much interest mainly because the secondary structure cannot easily be described and the force field cannot be informed from experimental data in a straightforward manner (Tozzini 2010). The solvent environment is of considerable importance in molecular simulations because many of the molecular driving forces like hydrophobicity, electrostatic interactions and hydrogen bonding are directly or indirectly affected by the properties of the solvent (Feig 2010). In order to decrease the computational time, many CG models do not explicitly include solvent molecules into the simulations (Sorenson and Head-Gordon 2002, Korkut and Hendrickson 2009, Tozzini and McCammon 2005, Basdevant et al. 2007). Instead, the solvent effects are implicitly included in their force field (Smith and Pettitt 1994). In all CG models, there is a trade-off between predictive power and accuracy which, to a great extent, is related to the parametrization strategy used in development of the force field. Depending on the functional form of the force field and available data for parametrization, several approaches such as elastic network (Tirion 1996), Boltzmann inversion (Tozzini and McCammon 2005), force matching (Ercolessi and Adams 1994) or a combination of these methods could be employed to determine the force field.

A short literature review on the available CG models and their features shows that there has not been much effort on developing CG models to study the disordered state of proteins. Elastic network and Go models are simple but their force fields are completely biased to a unique reference structure. In more complex CG models like MARTINI (Monticelli et al. 2008), the Head-Gordon model (Yap et al. 2007) and the model developed by Korkut and Hendrickson (Korkut and Hendrickson 2009), a-priori knowledge of the local secondary structure of the proteins is required to perform the simulations. On the other hand, CG models that can predict the folded structure of proteins to some extent, like the force fields developed by Tozzini et al. (Tozzini et al. 2006, Tozzini et al. 2007) and Bereau and Deserno (Bereau and Deserno 2009) are parametrized using databases of folded protein structures which do not guarantee to give the correct ensembles for unfolded proteins.

In this work, we propose a one-bead-per-amino-acid model for unfolded proteins. In order to obtain the local interaction potentials, experimentally obtained Ramachandran plots for the coil regions of proteins are converted into distributions of pseudo-bond and pseudo-dihedral angles between neighboring \( \alpha \)-carbons in the polypeptide chain. These distributions are used to derive bending and torsion potentials that are residue and sequence specific. Potentials with different levels of accuracy are proposed and an assessment study is then performed on the gen-
erated potential functions to select the best set of potentials in terms of simplicity and accuracy. Finally the model is used to study chemically denatured proteins and the effect of sequence composition on them.

2.2 Extraction method to obtain coarse-grained potentials

2.2.1 Mapping backbone internal degrees of freedom \((\phi, \psi)\) to pseudo bending and torsion angles \((\theta, \alpha)\)

A geometrical representation of a coarse-grained polypeptide chain together with the CG degrees of freedom is shown in Figure 2.1(b). In the all-atom representation of the backbone (Figure 2.1(a)), the bond lengths and bond angles display only a small variation from their average value (Finkelstein and Ptitsyn 2002) so they are assumed to remain fixed in the present work. The average bond lengths of \(C_N - N, C_C - C\), and \(C - N\) are 0.145 nm, 0.152 nm, 0.133 nm, respectively, with the average bond angles \(C_C - C - N = 116^\circ, C - N - C_C = 122^\circ\) and \(N - C_C - C = \tau = 111^\circ\) (Creighton 1993). A trans conformation is presumed for the peptide bond \((\omega = 180^\circ)\) and the rare possibility of cis conformation is ignored. The stated assumptions imply that the dihedral angles \(\phi\) and \(\psi\) (see Figure 2.1(a)) are the only degrees of freedom of the all-atom backbone.

Figure 2.1: All atom schematic of a polypeptide chain (a) and coarse-grained representation (b) of the backbone with pseudo-bending and torsion angles. In (b) the dashed lines represent the polypeptide chain and the solid lines are the pseudo-bonds between \(C_C\) carbons which represent the coarse-grained geometry.

Figure 2.1(b) demonstrates the CG representation of the polypeptide chain by connecting the \(\alpha\)-carbons through pseudo-bonds. With the assumptions above, the pseudo-bond lengths between subsequent \(C_C\)'s remain fixed at a distance of 0.38 nm as defined by the geometry. The pseudo-bending angle \(\theta\) and pseudo-dihedral angles \(\alpha\) for the CG chain are defined between three and four consecutive \(C_C\)'s. The relationship between the CG \((\theta, \alpha)\) and all-atom \((\phi, \psi)\) degrees of freedom can be established by explicit geometrical expressions (Levitt 1976, Tozzini et al. 2006). The pseudo-bending angle \(\theta_i\) of bead \(i\) of the coarse-grained chain is thus found to
be given by
\[
\cos \theta_i = \cos \tau (\cos \gamma_1 \cos \gamma_2 - \sin \gamma_1 \sin \gamma_2 \cos \phi_i \cos \psi_i) - \\
\sin \gamma_1 \sin \gamma_2 \sin \phi_i \sin \psi_i + \sin \tau (\cos \psi_i \sin \gamma_1 \cos \gamma_2 + \\
\cos \phi_i \cos \gamma_1 \sin \gamma_2)
\]
(2.1)
where \(\gamma_1 = 20.7^\circ, \gamma_2 = 14.7^\circ\) and \(\tau = 111^\circ\) are constant angles (see Figure 2.1(b)). The following approximate formula has been suggested for the pseudo-torsion angle (Tozzini et al. 2006):
\[
\alpha_i = 180^\circ + \psi_i + \phi_{i+1} + \gamma_1 \sin \psi_{i+1} + \gamma_2 \sin \phi_i.
\]
(2.2)
In contrast to the pseudo-bending angle, which depends only on one set of backbone dihedral angles \((\phi_i, \psi_i)\), the pseudo-torsion angle is a function of two consecutive sets of backbone dihedral angles \((\phi_i, \psi_i)\) and \((\phi_{i+1}, \psi_{i+1})\). It is important to note that in the force fields developed specifically for well-defined secondary structures (Levitt 1976, Tozzini et al. 2006), the simplifying assumption \(\phi_i = \phi_{i+1}, \psi_i = \psi_{i+1}\) is made for mapping \(\alpha\). However, for proteins without any regular structures, this assumption is not justified. This will be discussed in more detail in the following sections regarding the extraction of torsion potentials.

Besides the approximate formula discussed above, an exact method has been adopted to calculate the pseudo-torsion angle from the backbone dihedral angles. In this method, a small all-atom chain is built using four consecutive dihedral angles with the geometrical data mentioned earlier. The pseudo-torsion angles are then measured directly from the coordinates of the \(\alpha\)-carbons of the generated chain by means of simple vector algebra (see Flory (Flory 1974) for details). This will allow for more precise extraction of pseudo-torsion potentials from all-atom experimental data, and will be used subsequently to verify the accuracy of the use of Eq. (2.2).

### 2.2.2 Coil library

The backbone \(\phi\) and \(\psi\) values of proteins are often presented in two dimensional density plots, called Ramachandran plots. It turns out that the Ramachandran space \([-180^\circ, 180^\circ] \rightarrow [-180^\circ, 180^\circ]\) is divided into several regions, each referring to a specific secondary structure. The empty regions refer to unfavorable conformations which are mainly caused by the steric clash between neighboring side chains or steric hindrance to the formation of hydrogen bonds between peptide groups and water molecules (Avbelj et al. 2006). We use these density plots to generate the mean force potentials for local interactions in the unfolded state. For this purpose, we adopt the Boltzmann inversion method
\[
U(q) = -k_B T \ln(P(q)),
\]
(2.3)
where \(q\) is any desired degree of freedom, \(P(q)\) is the probability distribution for \(q\), \(T\) is the temperature and \(k_B\) is the Boltzmann constant.

Special care must be taken while choosing the appropriate Ramachandran plots. The required density plots should not include data related to any secondary structure, and also since we are interested in local interactions, long range effects must be absent or have a negligible
impact on the density plots. The data that satisfy these conditions best are from protein coil regions. Coil regions are those parts of proteins that cannot be classified to any kind of known secondary structure. This implies that their backbone conformations are not biased to any secondary structure. It has been shown that the intrinsic backbone preferences of di-peptides are strikingly similar to the backbone conformations of coil regions of proteins (Avbelj et al. 2006) confirming the assumption that long range hydrophobic or electrostatic interactions are absent or have a very small effect on the Ramachandran data obtained from this class of residues. The DASSD library is used to extract Ramachandran plots of the coiled regions of proteins. We extract the dihedral angles of the central residues of short fragments (of length 1, 3, and 5) from the coiled regions, which gives the possibility to extract meaningful potentials by considering the effect of neighboring residues on the obtained potentials. The database is generated from 5,227 non-redundant high resolution (less than 2 Å) protein structures (Dayalan et al. 2006).

2.2.3 3-letter amino acid model

The current size of the coil library is not large enough to extract CG bending and torsion potentials for all 20 amino acids accounting for all possible neighbors. This would give a huge number of different potentials which is not in accordance with the aim of coarse graining and data reduction. To reduce the number of different potentials, the amino acids are categorized into several sub-groups based on the similarities between their Ramachandran data. Four basic types of Ramachandran plots have been reported in the literature depending on the stereo-chemistry of the amino acids: Glycine, Proline, ”Generic” (which refers to the remaining 18 amino acids), and “Pre-Proline” (which refers to residues preceding a Proline) (Ho and Brasseur 2005). Since the Ramachandran plots are the main input to generate the CG potentials, different potential functions are expected for Glycine (G), Proline (P), and the rest of the amino acids (X) depending on their neighbor residues.

2.2.4 Bending potentials

Pseudo-bending potentials are obtained by Boltzmann-inversion of the $\theta$ probability distribution. Initially, $\phi$ and $\psi$ dihedral angles for the central residue of different triple combinations of P, G, and X are extracted from the 3-residue-fragments in the coil library. A table with the number of datapoints for each combination is shown in the Appendix (Table 2.2). The extraction procedure is depicted schematically in Figure 2.2 for X-X-P combinations. In Figure 2.2(a) the $\phi$ and $\psi$ values are plotted for all X amino acids (i.e. those amino acids that are not P or G) that have an X preceding it and a P following it. In the next step, each data point in Ramachandran space is mapped to $\theta$-space using Eq. (2.1). Collecting all data points in data bins (here we use $m = 60$ data bins (Shimazaki and Shinomoto 2007)) gives the $\theta$ probability distribution (Figure 2.2(b)) which is then directly converted into the bending potential by Eq. (2.3) (Figure 2.2(c)).

In order to study the neighbor-residue effect on the obtained bending potentials, three levels of accuracy have been investigated. The level-1 bending potentials are obtained by picking out Ramachandran data for all 27 combinations of G, P, and X and then extracting bending potentials by the procedure explained in Figure 2.2.
2. Coarse-grained potentials for local interactions in unfolded proteins

Figure 2.2: Extraction procedure for the bending potential of X-X-P combinations. (a) Ramachandran data for X-X-P combinations extracted from the coil library. This plot for X-X-P contains 3559 points. (b) Normalized distribution of the bending angle $\theta$, which is obtained by mapping all the Ramachandran data from (a) to the pseudo-bending angle $\theta$ through Eq. (2.1) and counting the frequencies in each bin (number of bins $m = 60$) (Shimazaki and Shinomoto 2007). (c) Obtained bending potential, $U(\theta)$ after applying the Boltzmann-inversion method on the probability distribution $P(\theta)$ presented in (b) and fitted by a piecewise polynomial.

The level-1 bending potentials are shown in the Appendix, Figures 2.11 to 2.13. For special combinations the number of available data points is not sufficient to generate statistically-meaningful potentials, so that the data points from the closest potential from the other combinations were added to construct the potentials\(^1\). Level-2 potentials consist of a reduced version of level-1 in which we distinguish those central residues that precede and do not precede a Proline (P), resulting in 6 different potentials (O-Z-Y and O-Z-P, with $Z \in \{G, P, X\}$, $Y=X+G$,

\(^1\)In some special cases (GGG, GGP, PGG, GPG, GPP, PPG, and PPP) the number of Ramachandran data points is not sufficient. However, even for these limited number of data points, the obtained potential functions show a local minimum. In these cases, amongst the other potentials in the family with the same central residue, the data points from the one with the same local minimum (XGG, XGP, GXG, XGP, XPG, XPP, PPX and XPP, respectively) were added to make a combined data set in order to generate the potentials.
2.2. Extraction method to obtain coarse-grained potentials

O=X+G+P). The level-2 bending potentials are shown in Figure 2.9 of the Appendix. For the level-3 potentials no neighbor-residue effect is included, which means that the Ramachandran data are extracted for each residue regardless of the neighboring residues, resulting in three separate bending potentials for G, P, and X (see Appendix, Figure 2.14).

2.2.5 Torsion potentials

A similar methodology as described above is applied to derive the pseudo-torsion potentials. The main difference with the bending procedure is that two separate sets of Ramachandran data are required to convert the all-atom dihedral angles $\phi$ and $\psi$ to the coarse-grained dihedral angle $\alpha$ according to Eq. (2.2). As a result, torsion potentials are derived for all possible double-combinations of amino acids, giving nine different torsion potentials for the 4-residue fragments O-Z-Z'-O with \{Z,Z'\} $\in$ \{X,G,P\}. This allows to sample the first set of all-atom dihedral angles from the Ramachandran data of the first residue Z and sample the second set from the Ramachandran data of the next residue Z' (see Figure 2.3(a) and (b)). However, sampling out of two separate Ramachandran plots results in an enormous number of possible combinations, because each Ramachandran plot may contain more than $10^5$ data points.

Therefore, we first divide each Ramachandran plot into a limited number of cells using an $n \times n$ grid (see Figure 2.3(c) and (d)) and characterize each cell by the coordinate of the cell center, e.g. ($\phi$, $\psi$)$_I$, and a weight factor $N_I$ representing the number of data points in that cell. Then a pseudo-torsion angle $\alpha_k$ is obtained by picking two sets of dihedral angles (e.g. ($\phi$, $\psi$)$_I$ from the first Ramachandran plot and ($\phi$, $\psi$)$_J$ from the second Ramachandran plot where $I, J = 1, \ldots, n^2$) giving $n^4$ different pseudo-torsion angles $\alpha_k$ in the range (-180°, 180°). Each $\alpha_k$ value is assigned a weight factor $N_k = N_I \times N_J$.

In order to find the probability distributions, the computed $\alpha_k$'s are distributed over $m$ bins ranging from -180° to 180° and the frequency of occurrence of $\alpha_p$ in each bin $p \in \{1, m\}$ is obtained by $N_p = \sum N_k$. The resulting normalized distribution function $P$ (Figure 2.3(e)) is then Boltzmann inverted using Eq. (2.3) and the torsion potential is determined by fitting a goniometric polynomial (Figure 2.3(f)) (The polynomial and coefficients can be found in Eq. (2.5) and Table 2.3 of the Appendix).

Torsion potentials have also been developed using two levels of accuracy: For level-1 torsion potentials, two sets of Ramachandran data have been extracted, taking into account the effect of sequence. For example, in order to obtain the torsion potential for the X-P combination, the first set of Ramachandran data corresponds to the central residue (X) of the 3-residue segment O-X-P and the second one corresponds to P from X-P-O (see Figure 2.3(a) and (b)). It should be noted that the size of the current coil library is insufficient to include the effect of other neighbors (e.g. subdividing O into G, P, and X and generating potentials for different combinations). Torsion potentials of level-2 are less accurate by construction, as two sets of Ramachandran data are selected regardless of the sequence. For example, to obtain the torsion potential for the X-P combination, the first set of Ramachandran data corresponds to the X residue from O-X-O segments and the second to the P residue from O-P-O, also resulting in 9 torsion potentials.
Figure 2.3: Extraction procedure for the torsion potentials for X-P combinations. (a) Ramachandran data for O-X-P combinations in the coil library, i.e. φ and ψ values for the central residue X that has an O (any residue) preceding it and a Proline following it. (b) Ramachandran data for X-P-O combinations. (c), (d) Frequency plots for Ramachandran data of O-X-P and X-P-O combinations in the coil library using a $n^2 = 40 \times 40$ grid. (e) Normalized probability distribution for the torsion angle $\alpha$ after mapping Ramachandran data from (c) and (d) to the pseudo-torsion angles and counting the frequencies in each bin (number of bins $m = 60$). (f) Obtained torsion potential after applying Boltzmann inversion on the probability distribution presented in (e).
2.3. Assessment of the potential functions

<table>
<thead>
<tr>
<th>Bending Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-1 Ben-1</td>
<td>27 potentials for all possible combinations of G, P, and X</td>
</tr>
<tr>
<td>Level-2 Ben-2</td>
<td>6 potentials for G, P, X and their pre-proline variants</td>
</tr>
<tr>
<td>Level-3 Ben-3</td>
<td>3 potentials for G, P, and X, regardless of the neighbor residues</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Torsion Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-1 Tor-1</td>
<td>The neighbor residue effect is included and mapping is done by Eq. (2.2)</td>
</tr>
<tr>
<td>Level-1 Tor-1-ex</td>
<td>The neighbor residue effect is included and mapping is done by vector algebra</td>
</tr>
<tr>
<td>Level-2 Tor-2</td>
<td>The neighbor residue effect is not included and mapping is done by Eq. (2.2)</td>
</tr>
</tbody>
</table>

Table 2.1: Bending and torsion potentials developed with different levels of accuracy.

In order to study the effect of the accuracy of the mapping scheme of Eq. (2.2) on the results, the level-1 potentials have been extracted using both the approximate Eq. (2.2) and the exact vector algebra method (see the last paragraph of Section 2.2.1). A summary of the developed bending and torsion potentials is given in Table 2.1.

2.3 Assessment of the potential functions

In this section, we will report on molecular dynamics (MD) simulations performed to study the effect of the different levels of accuracy for the developed potentials on the radius of gyration $R_g$ of arbitrary CG chains. The simulations are performed with GROMACS version 4.0.7 (Hess et al. 2008) in canonical ensemble at a temperature of $T = 300$ K using a Langevin thermostat. The equations of motion are integrated with a time-step of 0.02 ps for at least $10^7$ MD steps and the first $10^6$ steps are not considered in the averaged $R_g$ and $R_s$ values reported in this study (see Figures 2.15 and 2.16 for a convergence study and time-step size selection). The translation of the center of mass of the chain is removed at each time-step but the chain is allowed to freely rotate. All CG beads have a mass of 124 amu (i.e. the average mass of all amino acids) and the bond lengths are kept fixed by a stiff harmonic potential of the form $U_{\text{bond}} = K(r - b)^2$ with $K = 8038$ kJ/nm$^2$/mol (Ashbaugh and Hatch 2008) and $b = 0.38$ nm and $r$ is the distance between two consecutive beads. The excluded volume effect is incorporated using a purely repulsive potential with a repulsion radius of $R_{\text{rep}} = 0.38$ nm. This is chosen based on the proposed values in the literature for short charge-free polypeptides denatured in Urea and GdmCl (Soranno et al. 2009).

The simulations have been carried out for chains composed of 100 CG beads with different amounts of G, P, and X residues. The results are presented in terms of diagrams (see Figure 2.4 and Figure 2.5), in which each curve represents the $R_g$ of a two-phase chain characterized by the pair (phase 1, phase 2) of the number of beads in phase 1 and phase 2, respectively. When the chains are composed of mixed phases, the residues are randomly distributed along the chains. Each data point shows the average of 8 simulations in which the specific sequence is chosen at random.

In Figure 2.4, the bending potentials of level-1, 2, and 3 are evaluated for the level-1 torsion potential (indicated by Tor-1, see Table 2.1). As can be seen from the three combinations of phases (P,X), (G,X) and (G,P), a higher value of $R_g$ is predicted as the amount of P residues in the chain is increased, while increasing the amount of G residues always results in a lower
2. Coarse-grained potentials for local interactions in unfolded proteins

$R_g$. This is due to the fact that the P residues induce a higher local stiffness and the G residues induce a lower local stiffness than the rest of the amino acids (Rauscher et al. 2006), which is captured in the developed potentials. The results for the level-2 bending potentials (Ben-2) show only a slight difference compared to the level-1 potentials (Ben-1), while the level-3 potentials profoundly underestimate the radius of gyration. This clearly demonstrates that accounting for the pre-Proline variants is essential, whereas accounting for all possible G, P and X combinations does not add additional accuracy.

Figure 2.4: The average radius of gyration $R_g$ of proteins with different combinations of X, P and G residues (P,X), (G,X) and (G,P) is used to evaluate the accuracy of the obtained bending potentials. The horizontal axis of the diagram shows the number of residues of each phase. The error bars are the standard deviation from the mean values of 8 simulations.

A similar procedure is pursued for evaluating the torsion potentials. Figure 2.5 shows the same diagram for the level-1 and level-2 torsion potentials (Tor-1 and Tor-2), while keeping the bending potentials fixed (level-2, Ben-2). An additional set of simulations is carried out using the level-1 torsion potentials generated by the exact mapping scheme described earlier (Tor-1-ex, see Table 2.1). The graph clearly depicts the big effect of neighboring residues on the overall dimensions of the CG chain. It is interesting to see that the level-2 torsion potentials predict a lower $R_g$ compared to the more detailed level-1 torsion potentials for chains composed of P and G residues, but on the other hand, predict a larger $R_g$ for chains mainly composed of X residues. The reason could be explained in the context of pre-Proline residues which have a stiffer and extended backbone conformation (Rauscher et al. 2006). Since no neighbor effect is accounted for in extracting the Tor-2 potentials, the pre-Proline conformations are also included in potentials with central residue X (resulting in a higher $R_g$ for purely X-containing chains). On the other hand, the stiffness of the potentials for pre-Proline residues is reduced by including other less-stiff conformations (resulting in lower $R_g$ for purely P-containing chains). The close agreement between Tor-1 and Tor-1-ex also demonstrates that Eq. (2.2) gives a good approximation for the pseudo-torsion angles ($\alpha$) as a function of the all-atom dihedral angles.

To estimate the accuracy of the level-1 torsion potentials, relative to level-2, we have investigated an even more accurate torsion potential for one point of Figure 2.5, namely (P,X)=(100,0) for which the difference between level-1 and level-2 is the largest. The generated torsion potential (using two sets of PPP Ramachandran data instead of using OPP and PPO) results in
2.4 Application to denatured proteins

When the native state of a protein breaks down due to high temperature, pressure or the presence of a denaturing agent, it will turn to a dynamic set of complex conformations which is called the denatured state of a protein (Rose 2002). The addition of denaturants weakens the hydrophobic forces and disrupts the native hydrogen bonds in the protein (Zangi et al. 2009, Lim et al. 2009, Das and Mukhopadhyay 2008) and as a result, only local interactions that restrict the polypeptide backbone to limited regions of conformation space are the dominant source
of structure in the denatured state of the proteins (Creamer 2008). Henceforth, we make use of the radius of gyration and Stokes radius for denatured proteins to benchmark the potential functions for bonded interactions developed in this study.

2.4.1 The radius of gyration $R_g$ of denatured proteins

Experimental studies have revealed that the ensemble average radius of gyration of denatured proteins follows a power law scaling of the type

$$R_g = R_0 N^{\nu},$$  \hspace{1cm} (2.4)

where $N$ is the number of residues, $R_0$ is a constant related to the persistence length of the polymer and $\nu$ is a scaling exponent. More specifically, it has been found that for crosslink-free low-charge chemically unfolded proteins with sizes ranging from 16 to 549 residues, $R_0 = 0.202 \pm 0.041$ nm and $\nu = 0.588 \pm 0.037$ (Kohn et al. 2004).

We have performed MD simulations using the exact amino acid sequence of 20 denatured proteins as extracted from the protein data bank and using the developed bending and torsion potentials of our 3-letter amino acid model (a list of the simulated proteins is provided in the Appendix, see Table 2.5). Repulsive interactions between non-neighboring beads are included to take into account the excluded volume effect using $R_{\text{rep}} = 0.38$ nm.

![Figure 2.6: Comparison between the simulated and experimentally obtained $R_g$ for 20 different non-crosslinked chemically denatured proteins (Kohn et al. 2004). Experimental error bars are taken from Kohn et.al (Kohn et al. 2004) and the numerical error bars represent the standard deviation of 8 independent simulations. The studied proteins are listed in Table 2.5 of the Appendix.](image)

The predicted values of $R_g$ for 20 denatured proteins are shown in Figure 2.6. For proper sampling, each protein sequence has been simulated for $10^7$ MD steps and each data point corresponds to the average of 8 simulations obtained for different initial velocity distributions. The length of the chosen proteins spans two orders of magnitude. The simulation results show that $R_g$ for chains longer than $N \sim 30$ beads follows a power-law scaling but the shorter chains do not. This suggests that for short chains, the local stiffness of the chain is the dominant factor.
2.4. Application to denatured proteins

in determining the structure rather than the entropic motion. For longer chains, it can be seen
that the $R_g$ of denatured proteins is well predicted by the developed potentials within 10% error (except two outliers). The obtained simulation results can be fitted to the power law in
Eq. (2.4) with $\nu = 0.589 \pm 0.009$ and $R_0 = 0.215 \pm 0.011$ nm (95% confidence bounds), which
slightly overestimates the experimental values. This small deviation could be explained by the
formation of hydrophobic clusters or the presence of residual secondary structures which can
affect the mean persistence length of the polymer under experimental conditions (Baldwin and

The generated conformations can also be used to study the hydrodynamic properties of
denatured proteins. In the HYDRO computer program (Garcia de la Torre et al. 1994, Carrasco
and Garcia de la Torre 1999) the protein is represented by spherical frictional elements centered
at the $\alpha$-carbons of the polypeptide chain and the Stokes radius is calculated based on the
hydrodynamic theory of interacting beads. In the hydrodynamic calculation the bead radius
of each frictional element is chosen to be 0.51 nm and the Stokes radius calculations with the
suggested value gives reasonable results for folded proteins with rigid structures (Zhou 2002).
The simulations are performed on each protein sequence for $2.5 \times 10^7$ MD steps. The Stokes
radius is calculated every 5000 steps and the average value is reported. Fourteen denatured
proteins (see Table 2.4) with lengths ranging from $N = 16$ to $N = 605$ amino acids are studied
by the developed model and the obtained Stokes radii are fitted to the power law $R_s = 0.259 \pm
0.019 N^{0.515 \pm 0.013}$ nm (95% confidence bounds) which is in good agreement with the reported
experimental data $R_s = 0.252 N^{0.522}$ nm (Zhou 2002) (see Figure 2.7).

![Figure 2.7: Comparison of the simulated and experimental Stokes radius for 14 chemically
denatured proteins (Zhou 2002). The error bars indicate the standard deviation. The studied
proteins are listed in Table 2.4 of the Appendix.](image)

The bending and torsion potentials in our model are uncoupled, so that the correlation
between these degrees of freedom is lost (Alemani et al. 2009, Trovato and Tozzini 2012). This
can affect the local conformations of the simulated proteins as well as the ensemble average
properties such as gyration radius. To investigate the local conformations, we have collected
the pseudo bending and torsion angles $\alpha$ and $\theta$ at regular intervals during the simulations of
the proteins studied in Figure 2.6. The results are plotted in Figure 2.18 (a). To check the validity
of the obtained density maps, a direct mapping from the Ramachandran plot of the coil regions of proteins (Figure 2.18 (b)) to the \((\alpha, \theta)\) space is constructed (Figure 2.18 (c)). Comparison between the obtained \((\alpha, \theta)\) density plots (i.e., comparing Figures 2.18 (a) and (c)) shows that the proteins can sample conformations in \((\alpha, \theta)\) space that are not realistic due to the absence of bending-torsion coupling in the potentials. To investigate this effect on the predicted results for the gyration radius, we have used an approximate method to account for the coupling effect by introducing a “1-4” distance potential between beads \(i\) and \(i + 3\). The procedure to obtain this potential is explained in section 2.G in the Appendix. We have repeated the simulations for the denatured proteins of Figure 2.6 with this additional potential, and it can be seen from Fig. 2.18 (d) that including the bending-torsion coupling effect, even in this very primitive form, pushes the system to sample more realistic conformations. It can also be observed that the most populated area in Figure 2.18 (d) corresponds to the poly-Proline II region which is experimentally confirmed to be the dominant backbone conformation in denatured proteins (Shi et al. 2002, Whittington et al. 2005). Next, the influence of the bending-torsion coupling on the average gyration radius is studied by comparing the results for \(R_g\) with and without the 1-4 distance potentials in Figure 2.19. This clearly shows that using the 1-4 distance potential only results in a minor change in the reported \(R_g\) values (maximum error of 3.1%, see Figure 2.19). This implies that the correlation between bending and torsion potentials does not have a considerable effect on the ensemble average \(R_g\) of the denatured proteins studied in this work, despite the fact that additional regions are sampled in \((\alpha, \theta)\) space.

### 2.4.2 Sequence and composition effect

The polymer chains studied in Section 2.3 clearly show the effect of amino acid composition of the chain on \(R_g\), highlighting the role of Pro in enlarging and that of Gly in reducing the conformational radius. A survey on the protein sequences studied in the previous section (Figure 2.6) and the proteins in the DASSD library shows that the amount of Pro and Gly residues never exceeds 15 percent.

![Figure 2.8: Simulation results for chains rich in Proline and Glycine residues.](image)

In order to study the effect of sequence and composition on the \(R_g\) of unfolded proteins, a series of simulations has been performed on protein chains with different lengths, containing
either 15 percent of Gly or 15 percent of Pro residues randomly distributed along the chain (the rest of the beads are of type X and the same simulation parameters as in section 2.3 and 2.4.1 are used). As expected, the sequences rich in Pro residues adopt a higher $R_g$ compared to the chains rich in Gly, producing an upper and lower bound for the $R_g$ of denatured proteins (see Figure 2.8). Indeed, the simulation results for the specific denatured proteins of Figure 2.6 fall in between the bounds of Figure 2.8.

2.5 Conclusion

We have proposed an implicit solvent one-bead-per-amino-acid coarse-grained model to study the unfolded state of proteins. Bending and torsion potentials for the bonded interactions are extracted from Ramachandran data of the coil regions of proteins in the protein data bank and hence the obtained potentials are not biased to any specific secondary structure. The accuracy of the potential functions is a compromise that maintains residue and sequence specificity, and at the same time allows for data reduction. For the bending potentials, the neighbor residue effect does not play an important role except in the case of pre-proline residues. On the other hand, accounting for the neighboring residue effect in the extracted Ramachandran density plots improved the torsion potentials to a great extent. The validity of the model was demonstrated by comparing the ensemble average $R_g$ of a series of denatured proteins with experimental values. It was shown that the developed potential functions for bonded interactions can successfully predict the scaling relations of denatured proteins. Future work will be focused on extending the current model with electrostatic and hydrophobic interactions in order to study natively unfolded proteins under physiological conditions.
2. Coarse-grained potentials for local interactions in unfolded proteins

2.A Appendix: Detailed Potentials

The bending and torsion potentials resulting from the methodology in Section 2.3 are presented in Figures 2.9 and 2.10.

**Figure 2.9:** Level-2 bending potentials for triple-combinations with G, P, and X as central residues that (a) does not have a Proline following them and (b) those which have a Proline following them. O represents any type of amino acid while Y represents any type of amino acid except P.

**Figure 2.10:** Torsion potentials for: (a) G-X, G-G, and G-P combinations. (b) P-X, P-G, and P-P combinations. (c) X-X, X-G, and X-P combinations.
2.B Appendix: Ramachandran data points of the 3-residue-fragments

<table>
<thead>
<tr>
<th>Combination</th>
<th># Data points</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGG *</td>
<td>259</td>
</tr>
<tr>
<td>GGP *</td>
<td>111</td>
</tr>
<tr>
<td>GGX</td>
<td>1733</td>
</tr>
<tr>
<td>PGG *</td>
<td>87</td>
</tr>
<tr>
<td>PGP *</td>
<td>91</td>
</tr>
<tr>
<td>PXG</td>
<td>55</td>
</tr>
<tr>
<td>PXP</td>
<td>89</td>
</tr>
<tr>
<td>XGP</td>
<td>1730</td>
</tr>
<tr>
<td>XPP</td>
<td>981</td>
</tr>
<tr>
<td>XGX</td>
<td>16279</td>
</tr>
</tbody>
</table>

Table 2.2: The number of Ramachandran datapoints for all 3-residue-fragments extracted from the coil library. The combinations indicated by an asterisk are those special cases for which the number of Ramachandran datapoints is not sufficient to extract Level-1 bending potentials.

2.C Appendix: Bending potentials of level-1 and level-3

Figure 2.11: Bending potentials of level-1 for GLY with different neighbor residues.
Figure 2.12: Bending potentials of level-1 for PRO with different neighbor residues.

Figure 2.13: Bending potentials of level-1 for X with different neighbor residues.
2.D Appendix: Fitting torsion potentials

In order to have the obtained periodic potentials and their first derivatives continuous at the whole region of $0 \leq \alpha \leq 360^\circ$, the torsion potentials are fitted by the function

$$V_{\text{tor}}(\alpha) = a_1 \sin(\alpha) + a_2 \cos(\alpha) + a_3 \sin(2\alpha) + a_4 \cos(2\alpha) +
\quad a_5 \sin(3\alpha) + a_6 \cos(3\alpha) + a_7 \sin(4\alpha) + a_8 \cos(4\alpha) + a_9 \sin(5\alpha) +
\quad a_{10} \cos(5\alpha) + a_{11} \sin(6\alpha) + a_{12} \cos(6\alpha) + a_{13}.$$  \hspace{1cm} (2.5)

The coefficients obtained from fitting to the datapoints for the level-1 torsion potentials are given in Table 2.3.

### Table 2.3: Coefficients of Eq. 2.5 obtained by fitting the torsion potentials of Level-1.
2. Coarse-grained potentials for local interactions in unfolded proteins

2.E Appendix: MD Simulations

Figure 2.15: Convergence study on one of the longest protein sequences in the present study (GroEL, \(N = 549\)). The chain is simulated for \(2.5 \times 10^7\) MD steps (with time-step 0.02 ps). The evolution of gyration radius (orange line) and the time-averaged radius of gyration (black solid line) show that \(10^7\) MD steps (time = \(2 \times 10^5\) ps) are sufficient for a converged sampling. The black solid line is the time average \(R_g\) neglecting the first \(10^6\) steps, \(R_g = \frac{1}{t-t_0} \int_{t_0}^{t} R_g(t')dt'\) (used to obtain the reported \(R_g\) values in the main text). The black dashed line is the radius of gyration averaged from \(t_0 = 0\) to time \(t\).

Figure 2.16: Convergence study on the time-step size of the MD simulations for a chain with \(N = 100\) beads (all X residues). The average temperature of the system during the whole simulation time is calculated for different time-step sizes. For the time-steps larger than 0.02 ps the average temperature deviates more than 2 K from the specified 300 K.
2.F Appendix: List of the denatured proteins

<table>
<thead>
<tr>
<th>Name</th>
<th>Length (N)</th>
<th>Solvent condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ribonuclease A</td>
<td>124</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>hemoglobin</td>
<td>144</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>myoglobin</td>
<td>153</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>immunoglobulin(fragment I)</td>
<td>230</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>chymotrypsinogen A</td>
<td>245</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>G3P dehydrogenase</td>
<td>331</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>pepsinogen</td>
<td>376</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>aldolase</td>
<td>370</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>serumalbumin</td>
<td>605</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>phosphoglycerate kinase</td>
<td>415</td>
<td>6M GuHCl</td>
</tr>
<tr>
<td>lysozyme (4964)</td>
<td>16</td>
<td>8M Urea</td>
</tr>
<tr>
<td>BPTI</td>
<td>58</td>
<td>-</td>
</tr>
<tr>
<td>lysozyme</td>
<td>129</td>
<td>8M Urea</td>
</tr>
<tr>
<td>triosephosphate isomerase</td>
<td>247</td>
<td>2.5M GuHCl</td>
</tr>
</tbody>
</table>

Table 2.4: The proteins used in studying $R_g$ and their denaturing condition (Zhou 2002).

<table>
<thead>
<tr>
<th>Name</th>
<th>Length (N)</th>
<th>Solvent condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>GroEL</td>
<td>549</td>
<td>4 M Urea</td>
</tr>
<tr>
<td>yPGK</td>
<td>416</td>
<td>2 M GuHCl</td>
</tr>
<tr>
<td>Creatine Kinase</td>
<td>380</td>
<td>6 M GuHCl</td>
</tr>
<tr>
<td>a-1S</td>
<td>268</td>
<td>6 M GuHCl</td>
</tr>
<tr>
<td>Carbonic anhydrase</td>
<td>260</td>
<td>6 M GuHCl</td>
</tr>
<tr>
<td>Apomyoglobin</td>
<td>154</td>
<td>6 M GuHCl</td>
</tr>
<tr>
<td>Snaase</td>
<td>149</td>
<td>6 M GuHCl</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>129</td>
<td>4 M GuHCl</td>
</tr>
<tr>
<td>CheY</td>
<td>129</td>
<td>5.7 M Urea</td>
</tr>
<tr>
<td>Rnase A</td>
<td>124</td>
<td>3.2-6 M GuHCl</td>
</tr>
<tr>
<td>mAcP</td>
<td>98</td>
<td>6.5-8 M Urea</td>
</tr>
<tr>
<td>ctACP</td>
<td>98</td>
<td>5.5-6.8 M Urea</td>
</tr>
<tr>
<td>Ubiquitin</td>
<td>76</td>
<td>4.9-6 M GuHCl</td>
</tr>
<tr>
<td>drK SH3</td>
<td>59</td>
<td>2 M GuHCl</td>
</tr>
<tr>
<td>Cytochrome C</td>
<td>39</td>
<td>4 M Urea</td>
</tr>
<tr>
<td>AK-37</td>
<td>37</td>
<td>6 M GuHCl</td>
</tr>
<tr>
<td>AK-32</td>
<td>32</td>
<td>6 M GuHCl</td>
</tr>
<tr>
<td>AK-27</td>
<td>27</td>
<td>6 M GuHCl</td>
</tr>
<tr>
<td>AK-16</td>
<td>16</td>
<td>4 M Urea</td>
</tr>
<tr>
<td>Angiotensin</td>
<td>8</td>
<td>4 M Urea</td>
</tr>
</tbody>
</table>

Table 2.5: The proteins used in studying $R_g$ and their denaturing condition (Kohn et al. 2004).
2.G Appendix: Bending-torsion coupling

In this section we develop a 1-4 distance potential (between beads $i$ and $i + 3$) to take into account the coupling between the bending and torsion potentials in an approximate manner. We assume that the potential energy of the distance between the $i$ and $i + 3$ beads from the simulations with the un-coupled bending and torsion potentials, $U_{\text{un-coupled}}^{i,i+3}(r)$, can be corrected by applying an additional 1-4 distance potential, $U_{1-4}^{i,i+3}(r)$, i.e.,

$$U_{\text{coupled}}^{i,i+3}(r) = U_{\text{un-coupled}}^{i,i+3}(r) + U_{1-4}^{i,i+3}(r).$$

(2.6)

By using Boltzmann-inversion for the coupled and un-coupled terms, we can write

$$U_{1-4}^{i,i+3}(r) = -k_B T \ln \left( \frac{P_{\text{coupled}}^{i,i+3}(r)}{P_{\text{un-coupled}}^{i,i+3}(r)} \right),$$

(2.7)

where $P_{\text{coupled}}^{i,i+3}(r)$ is the probability distribution of the distance between the $C_{\alpha}$ atoms at locations $i$ and $i + 3$ from the coil library (coupled bending-torsion), and $P_{\text{un-coupled}}^{i,i+3}(r)$ is the probability distribution obtained directly from the bead coordinates of the simulated denatured proteins using the uncoupled potentials. To obtain $P_{\text{coupled}}^{i,i+3}(r)$ first the Ramachandran data are converted to the pseudo bending and torsion angles and then the $i$ and $i + 3$ distances are calculated for each pair of pseudo bending and torsion angles by means of vector algebra, assuming a fixed bond length of 0.38 nm between the $\alpha$-carbons. The obtained potential is fitted with a polynomial of the form

$$U_{1-4}^{i,i+3}(r) = K_{1-4} \left( \frac{r}{b} - 1.32 \right)^4 \left( \frac{r}{b} - 2.63 \right)^2 - C$$

(2.8)

where $b = 0.38$ nm, $K_{1-4} = 45.16$ kJ/mol and $C = 1.5$ kJ/mol (see Fig. 2.17).

![Figure 2.17: The developed 1-4 distance potential $U_{1-4}^{i,i+3}(r)$ to include the bending-torsion coupling effect in an approximate sense.](image)
2.H Appendix: \((\alpha, \theta)\) density maps

Figure 2.18: (a) The normalized \((\alpha, \theta)\) density map of the 20 denatured proteins listed in Table 2.5 obtained from the simulations shown in Figure 6 of the main text. For each protein, the pseudo bending and torsion angles for all residues are calculated every 100 ps. Then the \((\alpha, \theta)\) space is discretized by \(2.5^\circ \times 2.5^\circ\) cells and the number of data-points in each cell is counted and normalized to obtain the density map. (b) The normalized distribution of the atomistic dihedral angles from the amino acids in the coil regions of proteins taken from (Dayalan et al. 2006). (c) The \((\alpha, \theta)\) density plot obtained by mapping (b) to the \((\alpha, \theta)\) space using equations (1) and (2) of the main text. The populated region at \((\alpha \sim -130, \theta \sim 116)\) corresponds to the poly-Proline II region in the Ramachandran plot. (d) The normalized \((\alpha, \theta)\) density map of the 20 denatured proteins listed in Table 2.5 obtained from the simulations carried out with the additional 1-4 distance potential shown in Figure 2.17.
2. Coarse-grained potentials for local interactions in unfolded proteins

2.1 Appendix: Effect of bending-torsion coupling on $R_g$

![Graph showing $R_g$ comparison](image)

**Figure 2.19:** Comparison of the average $R_g$ of proteins in table 2.5, obtained from simulations with and without using 1-4 potential. The maximum error is 3.1 %.