Chapter 1
Towards a Coarse-Grained Model for Unfolded Proteins

Ali Ghavami, Erik Van der Giessen, and Patrick R. Onck

Abstract It is widely accepted that many biological systems benefit from the specific and unique properties of unfolded proteins. In order to study the conformational dynamics of these proteins, we propose an implicit solvent one-bead per amino-acid coarse-grained (CG) model. For the local backbone interactions, 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 CG chain. The obtained density plots are then used to derive bending and torsion potentials, which are residue- and sequence-specific. Our results show that the local interactions can be captured by specifically accounting for the presence of Proline and Glycine in the amino-acid sequence. An upper and lower bound is suggested for the radius of gyration of denatured proteins based on their specific sequence composition.

1.1 Introduction

In spite of the well-established relation between the biological function of proteins and their specific folded structure, the important role of unfolded proteins in many vital biological processes can not be ignored (Fink, 2005; Tompa, 2009). Rapid increase of our knowledge on the structure of proteins has revealed that many proteins and protein domains are intrinsically unstructured. The absence of a stable secondary structure in their polypeptide chain is the main reason behind the basic functions of unfolded proteins, which can be classified into four functional groups, namely molecular recognition, molecular assembly, protein modification and entropic chain activities (Dunker et al., 2002; Radivojac et al., 2007; Tompa, 2009).

Atomic-level molecular dynamics simulations provide detailed insight of the interactions and dynamics present in protein structures. However, because of the limitations in computational resources it is still not possible to reach biologically-interesting time and length scales. Unfolded proteins are even more dynamic and

A. Ghavami · E. Van der Giessen · P.R. Onck (✉)
Zernike Institute for Advanced Materials, University of Groningen, Groningen, The Netherlands
e-mail: p.r.onck@rug.nl

G.A. Holzapfel, E. Kuhl (eds.), *Computer Models in Biomechanics*,
DOI 10.1007/978-94-007-5464-5_1,

therefore even longer simulations are required in order to obtain statistically-meaningful results. The necessity to achieve biologically interesting time and length scales, have drawn the interest of researchers towards the development of coarse-grained (CG) models.

There is a limited set of available CG models that account for the disordered state of proteins. Simple models such as the elastic network and Go-models (Tirion, 1996) have been developed but their force fields are completely biased to a unique reference structure. In the more complex CG models like the MARTINI model (Monticelli et al., 2008), the Head-Gordon model (Yap et al., 2007) and the model developed by Korkut and Hendrickson (2009), a priori knowledge of the local secondary structure of the protein is required to perform the simulations. The CG models with more predictive power (Tozzini et al., 2006, 2007; Bereau and Deserno, 2009) are parametrized using databases of folded protein structures and therefore cannot be expected to reproduce the correct conformational dynamics of unfolded proteins.

In the present work, a one bead per amino-acid, implicit solvent model for unfolded proteins is proposed. Local interaction potentials are obtained by converting experimentally-obtained Ramachandran plots for the coil regions of proteins into distributions of pseudo-bond and pseudo-dihedral angles between neighboring α-carbons in the CG chain. These distributions are used to derive bending and torsion potentials, which are residue- and sequence-specific. As an example, the model is used to study the ensemble average gyration radius of denatured proteins as a function of the amino acid sequence.

1.2 Extraction Method to Obtain Coarse-Grained Potentials

In the following sections, the general methodology for extracting the CG potential functions from the Ramachandran data of coiled regions of proteins is summarized; more details can be found in (Ghavami et al., 2012).

1.2.1 Mapping Backbone Internal Degrees of Freedom (ϕ, ψ) to Pseudo Bending and Torsion Angles (θ, α)

A geometrical representation of the coarse-grained polypeptide chain together with the CG degrees of freedom are shown in Fig. 1.1. In the all-atom representation of the backbone (Fig. 1.1(a)), the bond lengths and bond angles display only a small variation from their average value so they are assumed to remain fixed in the present work (Finkelstein and Ptitsyn, 2002). The average bond lengths of C_α–N, C_α–C and C–N are 0.145 nm, 0.152 nm, 0.133 nm, respectively, with the average bond angles C_α–C–N $= 116°$, C–N–$C_\alpha = 122°$ and N–C_α–C $= \tau = 111°$. A Trans-conformation is presumed for the peptide bond ($\omega = 180°$) and the rare possibility of Cis-conformation is neglected. With the stated assumptions, it could be implied

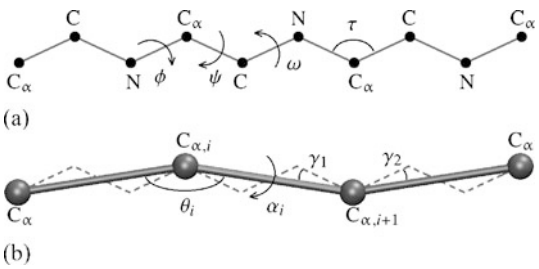

Fig. 1.1 All atom schematic of a polypeptide chain (**a**) and coarse-grained representation (**b**) of the backbone with pseudo-bending and torsion angles. Side chains are not shown in (**a**). In (**b**) the *dashed lines* represent the polypeptide chain and the bond angle θ and dihedral angle α represent the pseudo-bonds in the coarse-grained geometry

that ϕ and ψ dihedral angles are the only degrees of freedom of the all-atom backbone.

Figure 1.1(b) demonstrates the CG chain by connecting the α-carbons through pseudo-bonds. Since the all-atom bond lengths, bond angles and dihedral angle ω are not supposed to change, the pseudo-bond lengths between subsequent C_α's remain fixed at a distance of 0.38 nm. Also, the pseudo-bending angle θ and pseudo-dihedral angles α for the CG chain are defined between three and four consecutive C_α's, respectively. A geometrical expression can be established for the relationship between the CG (θ, α) and all-atom (ϕ, ψ) degrees of freedom as suggested by (Levitt, 1976; Tozzini et al., 2006). The pseudo-bending angle for the coarse-grained chain is given by

$$\begin{aligned}\cos\theta_i = {} & \cos\tau\,(\cos\gamma_1\cos\gamma_2 - \sin\gamma_1\sin\gamma_2\cos\phi_i\cos\psi_i) \\ & + \sin\tau\,(\cos\psi_i\sin\gamma_1\cos\gamma_2 + \cos\phi_i\cos\gamma_1\sin\gamma_2) \\ & + \sin\gamma_1\sin\gamma_2\sin\phi_i\sin\psi_i\,, \end{aligned} \tag{1.1}$$

where $\gamma_1 = 20.7^\circ$, and $\gamma_2 = 14.7^\circ$ are constant angles (see Fig. 1.1(b)). The exact formula for the pseudo-torsion angle is very complex, but the following approximate formula has been suggested by Tozzini et al. (2006):

$$\alpha_i = 180 + \phi_i + \psi_{i+1} + \gamma_1 \sin\psi_{i+1} + \gamma_2 \sin\phi_i\,. \tag{1.2}$$

It can be inferred from these equations that the pseudo-bending angle θ_i depends only on one set of backbone dihedral angles (ϕ_i, ψ_i), but the pseudo-torsion angle α_i is a function of two consecutive sets of backbone dihedral angles $(\phi_i, \psi_i, \phi_{i+1}, \psi_{i+1})$. It is worth noting that, in the force-fields developed specifically for well-defined secondary structures, the simplifying assumptions $(\phi_i = \phi_{i+1}, \psi_i = \psi_{i+1})$ are made for mapping α (Levitt, 1976; Tozzini et al., 2006). However, this assumption does not hold for proteins without any regular structure.

1.2.2 Coil Library

The ϕ and ψ dihedral angles of the backbone of protein structures are often presented in two-dimensional density plots, called Ramachandran plots. The Ramachandran space $[-180°, 180°) \rightarrow [-180°, 180°)$ is divided into several regions, each one referring to a specific secondary structure. The empty regions refer to the disfavored 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). Here, these density plots are used to generate the mean force potentials for local interactions in the unfolded state. For this purpose, we adopt the Boltzmann inversion method

$$\mathrm{U}(q) = -k_{\mathrm{B}} T \ln\big[\mathrm{P}(q)\big], \tag{1.3}$$

where q is any desired degree of freedom, $\mathrm{P}(q)$ is the probability distribution for q, T is the temperature and k_{B} is the Boltzmann constant.

In order to obtain meaningful potentials for unfolded proteins, appropriate Ramachandran data must be extracted from the protein data bank. The required density plots should not be biased towards any secondary structure, while long-range effects (hydrophobic or electrostatic interactions) must be absent or have a negligible impact on the density plots. The data that satisfy these conditions best are for the coil regions of proteins. The coil regions are those parts of proteins that cannot be classified in any kind of known secondary structure. This implies that their backbone conformations are not biased to any regular structure. Also 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 negligible for this class of residues.

The DASSD library is used to extract Ramachandran plots of the coiled regions of proteins (Dayalan et al., 2006). This database contains dihedral angles of central residues of short amino-acid fragments (of length 1, 3 and 5), which gives the possibility to extract meaningful potentials by considering the effect of neighboring residues on the obtained potentials. The database is extracted from 5,227 non-redundant high resolution (less than 2 Å) protein structures and a secondary structure assignment is carried out using the STRIDE algorithm (Frishman and Argos, 1995).

1.2.3 Three-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. Since the Ramachandran plots are the main input for the extraction of the CG potentials, they provide the best reference to compare different amino-acids and to categorize them into several sub-groups based on the similarities in their Ramachandran data.

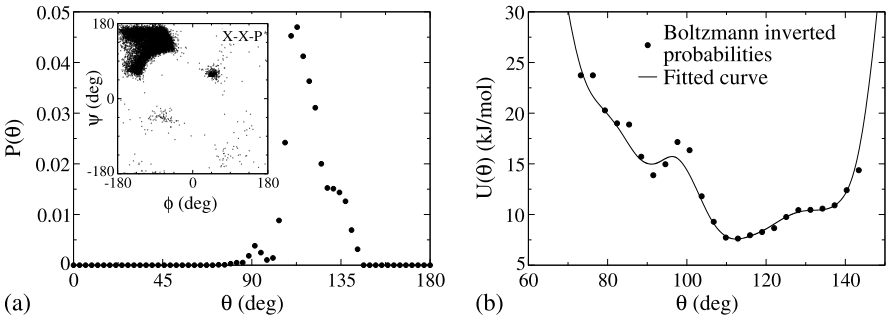

Fig. 1.2 Extraction procedure for the bending potential of X–X–P combinations. (**a**) Normalized distribution of the bending angle θ, which is obtained by mapping all the Ramachandran data (*inset*) to the pseudo-bending angle θ through Eq. (1.1). *Inset*: Ramachandran data for the central residue of X–X–P combinations extracted form the coil library. (**b**) Obtained bending potential, $U(\theta)$ after applying the Boltzmann-inversion method on the probability distribution $P(\theta)$ presented in (**a**)

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). As a result, different potential functions are expected for Glycine (G), Proline (P), and the rest of the amino acids (X) depending on their neighboring residues.

1.2.4 Extraction of Potential Functions

Bending potentials for the pseudo-bond angles are obtained by Boltzmann-inversion of the θ probability distribution. Initially, ϕ and ψ 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. The extraction procedure is depicted schematically in Fig. 1.2 for X–X–P combinations. In Fig. 1.2(a-inset) the ϕ and ψ 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 datapoint in the Ramachandran space is mapped to θ using Eq. (1.1). Collecting all datapoints in data bins gives the θ probability distribution (Fig. 1.2(a)) which is then directly converted to the bending potential by Eq. (1.3) (see Fig. 1.2(b)).

In order to develop the bending potentials, one can consider different levels of accuracy. This could range from developing 27 bending potentials for all combinations of G, P and X to just three sets of potentials for our three letter amino-acid alphabet ignoring any neighbor dependence. Studying proteins with different amino-acid contents shows that including the neighbor-residue effect is important only if the considered residue is preceding a Proline residue. Therefore, 6 sets of bending potentials are suggested in which we distinguish those central residues that do and do not precede a Proline.

The same methodology is also applied to derive the pseudo-torsion potentials. The main difference with the bending procedure is that in Eq. (1.2) two separate sets of Ramachandran data (e.g., ϕ_i, ψ_i, ϕ_{i+1} and ψ_{i+1}) are required to convert the all-atom dihedral angles ϕ and ψ to the coarse-grained dihedral angle α. Studying different levels of accuracy resulted in torsion potentials for all possible double-combinations of X, P and G amino-acids, giving 9 different torsion potentials. The reader is referred to (Ghavami et al., 2012) for more background information.

1.3 Application to Denatured Proteins

High temperature, pressure or the presence of a chemical denaturing agent can break down the native structure of folded proteins. As a result, 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 disrupts the native hydrogen bonds and weakens the hydrophobic forces in the protein (Das and Mukhopadhyay, 2008; Lim et al., 2009; Zangi et al., 2009). After denaturation, only local interactions that restrict the polypeptide backbone to limited regions of the conformational space are retained (Creamer, 2008). Experimental studies have revealed that the ensemble-average radius of gyration of denatured proteins follows a power-law scaling:

$$R_g = R_0 N^{\nu}, \tag{1.4}$$

where N is the number of residues, R_0 is a constant related to the persistence length of the polymer and ν is a scaling exponent.

The obtained potentials are used to study the effect of the composition of the protein sequence on the R_g of denatured proteins, with special emphasis on the role of Proline in enlarging and that of Glycine in reducing the conformational radius. A survey of protein sequences of folded and unfolded proteins shows that the amount of Proline and Glycine residues never exceeds 15 percent. 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 15 percent of Glycine and Proline residues randomly distributed along the chain. As expected, the sequences rich in Proline residues lead to a higher R_g compared to the chains rich in Glycine, producing an upper and lower bound for the R_g of denatured proteins (see Fig. 1.3). Any chain with less than 15 percent Proline or 15 percent Glycine falls within this band. The experimental results of low-charge crosslink-free chemically unfolded proteins with sizes ranging from 16 to 549 residues, shows that R_g can be well fitted by the power-law relationship in Eq. (1.4) with $R_0 = 0.202 \pm 0.041$ nm and $\nu = 0.588 \pm 0.037$ (Kohn et al., 2004), which indeed falls in between the computed bounds in Fig. 1.3.

Recently, many studies have been conducted on poly-Proline proteins showing that these proteins form elongated left-handed helices with a very stiff backbone structure (Adzhubei and Sternberg, 1993). The current model is able to capture the helix conformation of poly-Proline proteins with a rise of 2.97 Å per Proline, which

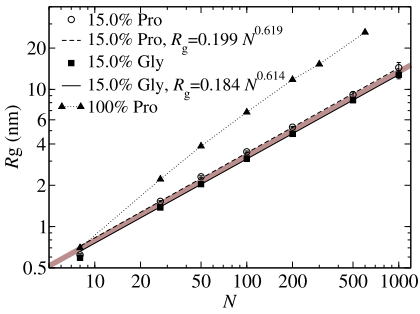

Fig. 1.3 Simulation results for chains rich in Proline and Glycine residues

is comparable to the 3.0 Å rise per Proline from atomistic simulations and 3.1 Å per Proline from experimental data (Dolghih et al., 2009). The dependence of the radius of gyration of poly-Proline proteins on the sequence length is also studied in Fig. 1.3. It clearly shows the higher dimensions of these synthetic proteins compared to natural proteins. It should be noticed that since Proline is considered a hydrophobic amino acid and the influence of solvent is not included in the present model, the predicted gyration radius is overestimating the R_g of poly-Proline proteins in aqueous solution.

1.4 Conclusion

In this paper, we have presented an implicit solvent, one-bead per amino-acid coarse-grained model to study the unfolded state of proteins. To ensure that the CG bending and torsion potential functions for bonded interactions are not biased to any specific secondary structure, the obtained potentials were extracted from Ramachandran data of the coil regions of proteins. The potential functions have been developed by accounting for the effect of neighboring residues, rendering the model to be residue- and sequence-specific. The model has been used to study the correlation between sequence composition and dimension of denatured proteins. Based on the Proline and Glycine content of the protein sequence, an upper and lower bound is constructed for the ensemble average R_g of denatured proteins, which is in agreement with the available experimental data. The developed model sets the stage for further developments towards the inclusion of electrostatic and hydrophobic interactions to study the characteristics of natively unfolded proteins under physiological conditions.

References

Adzhubei AA, Sternberg MJ (1993) Left-handed polyproline II helices commonly occur in globular proteins. J Mol Biol 229:472–493

Avbelj F, Grdadolnik SG, Grdadolnik J, Baldwi RL (2006) Intrinsic backbone preferences are fully present in blocked amino acids. Proc Natl Acad Sci USA 103:1272–1277

Bereau T, Deserno M (2009) Generic coarse-grained model for protein folding and aggregation. J Chem Phys 130:235106
Creamer T (2008). Unfolded proteins: from denatured to intrinsically disordered. Nova Publishers, Hanover
Das A, Mukhopadhyay C (2008) Atomistic mechanism of protein denaturation by urea. J Phys Chem B 112:7903–7908
Dayalan S, Gooneratne ND, Bevinakoppa S, Schroder H (2006) Dihedral angle and secondary structure database of short amino acid fragments. Bioinformation 1:78–80
Dolghih E, Ortiz W, Kim S, Krueger BP, Krause JL, Roitberg AE (2009) Theoretical studies of short polyproline systems: recalibration of a molecular ruler. J Phys Chem A 113:4639–4646
Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z (2002) Intrinsic disorder and protein function. Biochemistry 41:6573–6582
Fink AL (2005) Natively unfolded proteins. Curr Opin Struct Biol 15:35–41
Finkelstein AV, Ptitsyn O (2002) Protein physics: a course of lectures. Academic Press, San Diego
Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23:566–579
Ghavami A, Van der Giessen E, Onck PR (2012) Coarse-grained potentials for local interactions in unfolded proteins (submitted for publication)
Ho BK, Brasseur R (2005) The Ramachandran plots of glycine and pre-proline. BMC Struct Biol 5:14
Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, Hasan MZ, Pande VS, Ruczinski I, Doniach S, Plaxco KW (2004) Random-coil behavior and the dimensions of chemically unfolded proteins. Proc Natl Acad Sci USA 101:12491–12496
Korkut A, Hendrickson WA (2009) A force field for virtual atom molecular mechanics of proteins. Proc Natl Acad Sci USA 106:15667–15672
Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104:59–107
Lim WK, Rösgen J, Englander SW (2009) Urea, but not guanidinium, destabilizes proteins by forming hydrogen bonds to the peptide group. Proc Natl Acad Sci USA 106:2595–2600
Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink S-J (2008) The MARTINI coarse-grained force field: extension to proteins. J Chem Theory Comput 4:819–834
Radivojac P, Iakoucheva LM, Oldfield CJ, Obradović Z, Uversky VN, Dunker AK (2007) Intrinsic disorder and functional proteomics. Biophys J 92:1439–1456
Rose GD (2002) Advances in protein chemistry, vol 62. Academic Press, San Diego
Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. Phys Rev Lett 77:1905–1908
Tompa P (2009) Structure and function of intrinsically disordered proteins. Chapman & Hall/CRC, London
Tozzini V, Rocchia W, McCammon JA (2006) Mapping all-atom models onto one-bead coarse-grained models: general properties and applications to a minimal polypeptide model. J Chem Theory Comput 2:667–673
Tozzini V, Trylska J, Chang CE, McCammon JA (2007) Flap opening dynamics in HIV-1 protease explored with a coarse-grained model. J Struct Biol 157:606–615
Yap EH, Fawzi NL, Head-Gordon T (2007) A coarse-grained alpha-carbon protein model with anisotropic hydrogen-bonding. Proteins, Struct Funct Bioinform 70:626–638
Zangi R, Zhou R, Berne BJ (2009) Urea's action on hydrophobic interactions. J Am Chem Soc 131:1535–1541