

Folding and Unfolding Simulations of a Three-Stranded Beta-Sheet Protein

Seung-Yeon Kim

School of Liberal Arts and Sciences, Korea National University of Transportation, Chungju, Republic of Korea
Email: sykimm@ut.ac.kr

Received 26 October 2015; accepted 5 January 2016; published 11 January 2016

Abstract

Understanding the folding processes of a protein into its three-dimensional native structure only with its amino-acid sequence information is a long-standing challenge in modern science. Two-hundred independent folding simulations (starting from non-native conformations) and two-hundred independent unfolding simulations (starting from the folded native structure) are performed using the united-residue force field and Metropolis Monte Carlo algorithm for betanova (three-stranded antiparallel beta-sheet protein). From these extensive computer simulations, two representative folding pathways and two representative unfolding pathways are obtained in the reaction coordinates such as the fraction of native contacts, the radius of gyration, and the root-mean-square deviation. The folding pathways and the unfolding pathways are similar each other. The largest deviation between the folding pathways and the unfolding pathways results from the root-mean-square deviation near the folded native structure. In general, unfolding computer simulations could capture the essentials of folding simulations.

Keywords

Protein, Folding, Unfolding, Computer Simulation

1. Introduction

Protein is a polypeptide in a living organism [1]. It is a linear polymer, built of twenty different amino acids (or residues), with a defined amino-acid sequence. Each amino acid consists of the intrinsic side-chain and the common backbone. Because different amino acids have a common backbone structure, their side-chains determine their nature—hydrophilic or hydrophobic. Hydrophobic amino acids of an unfolded protein in water drive it to fold into its three-dimensional native structure, determining its biological function. Two amino acids, glycine (G, Gly) and proline (P, Pro), are neutral. Usually, alanine (A, Ala), valine (V, Val), leucine (L, Leu), isoleucine (I, Ile), cysteine (C, Cys), methionine (M, Met), phenylalanine (F, Phe), tyrosine (Y, Tyr), and tryptophan (W, Trp) are classified as hydrophobic amino acids. Similarly, serine (S, Ser), threonine (T, Thr), asparagine (N, Asn), glutamine (Q, Gln), aspartic acid (D, Asp), glutamic acid (E, Glu), lysine (K, Lys), arginine (R, Arg), and histidine (H, His) are considered as hydrophilic amino acids. Among them, aspartic acid and glutamic acids are negatively charged, whereas lysine, arginine, and histidine are positively charged.

The primary structure of a protein is the one-dimensional amino-acid sequence of the protein, which is translated from the nucleic-acid sequence of the corresponding gene. For example, the primary structure of the brain neuropeptide Met-enkephalin, composed of five amino acids, is Tyr-Gly-Gly-Phe-Met [2]. The usual local conformations of backbones, such as alpha-helix and beta-sheet, of proteins correspond to the secondary structures of proteins. Alpha-helix, the most abundant secondary structure, is easily formed between amino acids neighboring in the sequence of a protein via hydrogen bonds between the backbone pairs, carboxyl oxygen and nitrogen with hydrogen. The right-handed alpha-helix has 3.6 residues per turn with the length 5.4 Å. The formation process of alpha-helix is well understood [3]. Each kind of protein has its unique three-dimensional folded structure, called the tertiary (native) structure. Information on the tertiary native structure of a protein is quite crucial in understanding its biological function and role.

Understanding the folding processes of a protein into its tertiary native structure only with its primary structure (amino-acid sequence information) is a long-standing challenge in modern science [4]. Understanding these folding processes is particularly important in this post-genomic era. Protein folding processes play the most important role in controlling a wide range of cellular functions. The failure of a proper protein folding results in the malfunction of biological systems, leading to various diseases. Although extensive experimental and computational studies on protein folding processes have been performed, many aspects of the processes are poorly understood [4].

Computer simulations have been carried out to study protein folding processes [5]. However, simulation of protein-folding processes with an atomistic model is a very difficult task. Usually, direct folding simulations have been mainly focused on simple models, such as lattice models, models where only native interactions are included (Go-type models), and a model with discrete energy terms whose parameters are optimized separately for each protein. Alternative indirect approaches have also been proposed including unfolding simulations starting from the folded state of a protein. Because protein folding simulation requires a very long time scale, protein unfolding simulation has been one of the most popular approaches. However, it is not clear whether we understand protein folding processes from unfolding simulations perfectly.

One of the most regular conformations adopted by proteins is the beta-sheet whose basic unit is the beta-strand. It is not stable itself whereas a single alpha-helix is stable itself. Frequently, this unstability of a single beta-strand results in formation of amyloid fibrils and various fatal diseases. It is difficult to understand the formation processes and stability of proteins with beta-strands. Betanova [6] is a monomeric, beta-sheet protein consisting of three antiparallel beta-strands. Betanova has twenty amino acids, and its primary structure is given by Arg-Gly-Trp-Ser-Val-Gln-Asn-Gly-Lys-Tyr-Thr-Asn-Asn-Gly-Lys-Thr-Thr-Glu-Gly-Arg. In this article, we perform and compare both folding simulations (starting from non-native conformations) and unfolding simulations (starting from the tertiary native structure) for betanova using the united-residue force field [7] and the most popular computer simulation method, Metropolis Monte Carlo algorithm [8].

2. Computational Methods

In the united-residue force field [7], a protein chain is represented by a sequence of alpha-carbon (C^α) atoms connected by virtual bonds with attached united side-chains (SC) and united peptide group located in the middle of the C^α - C^α virtual bonds. All the virtual bonds are fixed in length; the C^α - C^α length is set to 3.8 Å, and the C^α -SC lengths are given for each amino-acid type. The energy function of a protein in the united-residue force field is given by

$$E = U_{dis} + \sum_{i < j} [U_{el-loc}^{(4)}(i, j) + U_{ss}(i, j)] + \sum_{i \neq j} U_{sp}(i, j) + \sum_{i < j-1} U_{pp}(i, j) + \sum_i [U_b(i) + U_t(i) + U_r(i)], \quad (1)$$

where U_{dis} denotes the energy term to form a disulfide bridge and $U_{el-loc}^{(4)}$ is the four-body interaction term. $U_{ss}(i, j)$ represents the mean free energy of the hydrophobic (hydrophilic) interaction between the side-chains i and j , $U_{sp}(i, j)$ corresponds to the excluded-volume interaction between the side-chain i and the peptide group j , and $U_{pp}(i, j)$ accounts for the electrostatic interaction between the peptide groups i and j . The terms $U_b(i)$, $U_t(i)$ and $U_r(i)$ denote the energies of virtual angle bending, virtual dihedral angle torsions, and side-chain rotamers, respectively. The parameters of the united-residue force field were optimized simultaneously for four proteins; betanova, zink-finger based beta-beta-alpha motif 1fsd (28 amino acids, one beta-hairpin and one alpha-helix), villin headpiece protein subdomain HP-36 (36 amino acids, three alpha-helix bundle), and fragment B of sta-

phylococcal protein A (46 amino acids, three alpha-helix bundle). The parameters were adjusted in such a way that the native-like conformations are more favored than the others energetically. In [7], the procedures to obtain the optimized parameters used in this article are described in detail.

The perturbed conformation is accepted according to the change in the energy, following Metropolis Monte Carlo algorithm [8]. For betanova, 100 independent unfolding simulations (starting from the tertiary native structure) with 10^5 Monte Carlo steps (shortly, MCS) were run at a fixed temperature. We divided 10^5 MCS into 28 intervals (first ten 10^2 MCS, subsequent nine 10^3 MCS, and the next nine 10^4 MCS), and took average in each interval. These averages were again averaged over 100 independent unfolding simulations at a given temperature. Also, 200 independent folding simulations (10^6 MCS for each run) were performed at a fixed temperature for betanova [9]. For folding simulations starting from non-native conformations, we divided 10^6 MCS into 28 intervals (first ten 10^3 MCS, subsequent nine 10^4 MCS, and the next nine 10^5 MCS), and the averages are taken over the whole conformations for each interval. These averages are averaged again over 100 independent computer simulations starting from random conformations. The same procedure is repeated for 100 independent computer simulations starting from fully extended conformations.

3. Computational Results

During all Monte Carlo simulations the values of the root-mean-square deviation (RMSD) from the experimental structure and the radius of gyration R_g were calculated using C^α coordinates. The fraction of the native contacts ρ is also measured during all computer simulations [10]. The values of $\rho = 1$ and $\rho = 0$ mean the experimental structure and a completely disordered conformation, respectively. RMSD, the radius of gyration, and the fraction of native contacts are the most popular reaction coordinates in understanding the folding and unfolding processes between the primary structure (one-dimensional amino-acid sequence) and the tertiary (three-dimensional) native structure.

Figure 1 shows two different folding pathways at $T = 40$ (arbitrary units) and the unfolding pathways at two different temperatures $T = 100$ and $T = 200$, between the folded native structure and unfolded conformations, in the reaction coordinates ρ and R_g (in units of \AA) for betanova. The folding pathways are obtained from 200 independent computer simulations at $T = 40$. The green triangles represent the averages of 100 folding pathways starting from random conformations. The red inverted triangles are the averages over 100 folding pathways starting from fully extended conformations (with $\rho = 0$, $R_g = 19.9 \text{ \AA}$, and $\text{RMSD} = 16.4 \text{ \AA}$). The unfolding pathways are obtained from 100 independent computer simulations, starting from the folded native structure (with $\rho = 0.97$, $R_g = 7.6 \text{ \AA}$, and $\text{RMSD} = 1.6 \text{ \AA}$), at a fixed temperature. The black circles represent the averages of 100 unfolding pathways at $T = 100$, and the blue squares are the averages over 100 unfolding pathways at $T = 200$. As shown in **Figure 1**, the unfolding pathways are similar even for quite different temperatures $T = 100$ and $T = 200$. The unfolding pathways are almost identical from the point of $\rho = 0.97$ and $R_g = 7.6 \text{ \AA}$ to the point of $\rho = 0.75$ and $R_g = 8 \text{ \AA}$, corresponding to a native-like conformation of betanova, still maintaining the antiparallel three-stranded beta-sheet. Two different folding pathways converge at the point of $\rho = 0.3$ and $R_g = 9.5 \text{ \AA}$, corresponding to the collapsed unfolded conformations. From this point, they are almost identical. Finally, the folding pathways and the unfolding pathways meet at the point of $\rho = 0.75$ and $R_g = 8 \text{ \AA}$. Between $\rho = 0.3$ and $\rho = 0.75$, the unfolding pathways lie slightly above the folding pathways for the same ρ value.

Figure 2 shows the folding pathways and the unfolding pathways between the folded native structure and the unfolded conformations in the reaction coordinates, ρ and RMSD (in units of \AA), for betanova. Around $\rho = 0.2$, four different pathways converge. Between $\rho = 0.2$ and $\rho = 0.5$, the folding pathways and the unfolding pathways are similar. For $\rho > 0.55$, the folding pathways lie above the unfolding pathways for the same ρ value, and the difference in RMSD values between the folding pathways and the unfolding pathways may become larger as ρ increases.

Figure 3 shows the folding pathways and the unfolding pathways between the folded native structure and the unfolded conformations in the reaction coordinates RMSD (in units of \AA) and R_g (in units of \AA) for betanova. Between $\text{RMSD} = 7 \text{ \AA}$ and 9 \AA , the folding pathways starting from fully extended conformations are nearly identical to two different unfolding pathways. For $\text{RMSD} < 7 \text{ \AA}$, the folding pathways lie slightly below the unfolding pathways for the same RMSD value. **Figure 3** suggests that the folding pathways meet the unfolding pathways around the point of $\text{RMSD} = 1.6 \text{ \AA}$ and $R_g = 7.6 \text{ \AA}$.

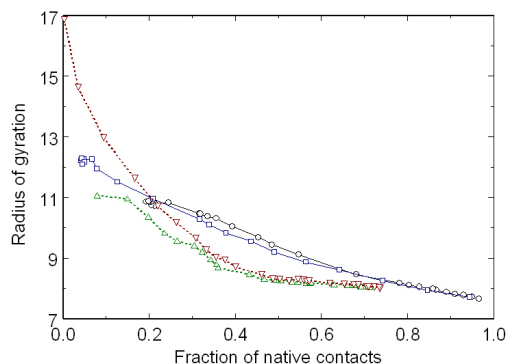


Figure 1. Folding and unfolding pathways between the folded native structure and unfolded conformations in the reaction coordinates, the fraction of native contacts ρ and the radius of gyration, for betanova. The values of $\rho = 1$ and $\rho = 0$ mean the experimental structure and a completely disordered conformation, respectively. The folding pathways are obtained from 200 independent computer simulations at $T = 40$. The (green) triangles represent the averages of 100 folding pathways starting from random conformations. The (red) inverted triangles are the averages over 100 folding pathways starting from fully extended conformations. The unfolding pathways are obtained from 100 independent computer simulations, starting from the folded native structure, at a fixed temperature. The (black) circles represent the averages of 100 unfolding pathways at $T = 100$. The (blue) squares are the averages over 100 unfolding pathways at $T = 200$.

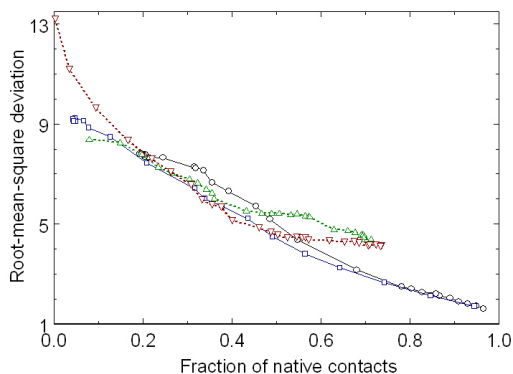


Figure 2. Folding and unfolding pathways in the reaction coordinates, the fraction of native contacts and the root-mean-square deviation (RMSD), for betanova. The (green) triangles, the (red) inverted triangles, the (black) circles, and the (blue) squares represent the folding pathways starting from random conformations at $T = 40$, the folding pathways starting from fully extended conformations at $T = 40$, the unfolding pathways at $T = 100$, and the unfolding pathways at $T = 200$, respectively.

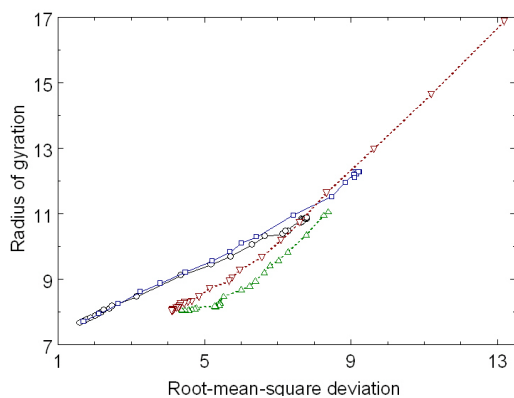


Figure 3. Folding and unfolding pathways in the reaction coordinates, the root-mean-square deviation and the radius of gyration, for betanova. The (green) triangles, the (red) inverted triangles, the (black) circles, and the (blue) squares represent the folding pathways starting from random conformations at $T = 40$, the folding pathways starting from fully extended conformations at $T = 40$, the unfolding pathways at $T = 100$, and the unfolding pathways at $T = 200$, respectively.

4. Conclusion

We have performed 200 independent folding simulations (starting from non-native conformations) and 200 independent unfolding simulations (starting from the tertiary native structure) using the united-residue force field and Metropolis Monte Carlo algorithm for betanova (three-stranded antiparallel beta-sheet protein). From these extensive computer simulations, we have obtained two representative folding pathways and two representative unfolding pathways in the reaction coordinates such as the fraction of native contacts, the radius of gyration, and the root-mean-square deviation. The folding pathways and the unfolding pathways are similar each other. The largest deviation between the folding pathways and the unfolding pathways results from the root-mean-square deviation near the folded native structure. Therefore, we may conclude that unfolding computer simulations capture the essentials of folding simulations.

Acknowledgements

This was supported by Korea National University of Transportation in 2015.

References

- [1] Creighton, W.E. (1993) *Proteins: Structures and Molecular Properties*. 2nd Edition, W.H. Freeman and Company, New York.
- [2] Hughes, J., Smith, T.W., Kosterlitz, H.W., Fothergill, L.A., Morgan, B.A. and Morris, H.R. (1975) Identification of Two Related Pentapeptides from the Brain with Potent Opiate Agonist Activity. *Nature*, **258**, 577-579. <http://dx.doi.org/10.1038/258577a0>
- [3] Pauling, L., Corey, R.B. and Branson, H.R. (1951) The Structure of Proteins: Two Hydrogen-Bonded Configurations of the Polypeptide Chain. *Proceedings of the National Academy of Sciences of the United States of America*, **37**, 205-211. <http://dx.doi.org/10.1073/pnas.37.4.205>
- [4] Pain, R.H. (2000) *Mechanisms of Protein Folding*. 2nd Edition, Oxford University Press, New York.
- [5] Friesner, R.A. (2002) *Computational Methods for Protein Folding*. John Wiley & Sons, New York. <http://dx.doi.org/10.1002/0471224421>
- [6] Kortemme, T., Ramirez-Alvarado, M. and Serrano, L. (1998) Design of a 20-Amino Acid, Three-Stranded Beta-Sheet Protein. *Science*, **281**, 253-256. <http://dx.doi.org/10.1126/science.281.5374.253>
- [7] Lee, J., Kim, S.-Y. and Lee, J. (2004) Design of a Protein Potential Energy Landscape by Parameter Optimization. *Journal of Physical Chemistry B*, **108**, 4525-4534. <http://dx.doi.org/10.1021/jp037076c>
- [8] Fishman, G.S. (1996) *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag, New York. <http://dx.doi.org/10.1007/978-1-4757-2553-7>
- [9] Kim, S.-Y., Lee, J. and Lee, J. (2005) Folding Simulations of Small Proteins. *Biophysical Chemistry*, **115**, 195-200. <http://dx.doi.org/10.1016/j.bpc.2004.12.040>
- [10] Kim, S.-Y., Lee, J. and Lee, J. (2004) Folding of Small Proteins Using a Single Continuous Potential. *Journal of Chemical Physics*, **120**, 8271-8276. <http://dx.doi.org/10.1063/1.1689643>