Biochimica et Biophysica Acta 1804 (2010) 2003–2015

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Folding and unfolding characteristics of short beta strand peptides under different environmental conditions and starting configurations

Astrid Maaß [a,1], Emine Deniz Tekin [b,c,1], Anton Schüller [a,1], Ahmet Palazoglu [b], Dirk Reith [a], Roland Faller [b,*]

[a] Fraunhofer Institute for Algorithms and Scientific Computing, Department of Simulation Engineering, Schloss Birlinghoven, 53754 St. Augustin, Germany
[b] University of California Davis, Department of Chemical Engineering and Materials Science, One Shields Avenue, Davis, CA 95616, USA
[c] Department of Physics, Middle East Technical University, 06531 Ankara, Turkey

ARTICLE INFO

Article history:
Received 2 March 2010
Received in revised form 31 May 2010
Accepted 21 June 2010
Available online 13 July 2010

Keywords:
Protein folding
Principal component analysis
Beta strands

ABSTRACT

We analyze the effect of different environmental conditions, sequence lengths and starting configurations on the folding and unfolding pathways of small peptides exhibiting beta turns. We use chignolin and a sequence of peptide G as examples. A variety of different analysis tools allows us to characterize the changes in the folding pathways. It is observed that different harmonic modes dominate not only for different conditions but also for different starting points. The modes remain essentially very similar but their relative importance varies. A detailed analysis from diverse viewpoints including the influence of the particular amino acid sequence, conformational aspects as well as the associated motions yields a global picture that is consistent with experimental evidence and theoretical studies published elsewhere. Patterns of modes that remain stable over a range of temperatures might serve as an additional diagnostic to identify conformations that have reliably adopted a native fold. This could aid in reconstructing the folding process of a complete protein by identifying conformationally determined regions.

1. Introduction

One of the most important unsolved problems in molecular biology is protein folding, i.e., understanding and predicting the three dimensional native structure of a protein and the sequence of intermediates which are assumed based on the information of its amino acid sequence. To understand protein folding and unfolding dynamics, molecular dynamics (MD) simulation provides a powerful tool that can generate detailed information about the interactions between the atoms of the protein and the solvent [1]. In atomistic MD simulations unfolding, the reverse process of folding is normally the more viable route [2–10] as proteins fold on timescales between microseconds and minutes. Even with the fastest computers, such a simulation would take years of computer time. Therefore, one has to speed up the simulation which is possible by studying unfolding instead at elevated temperatures. Another reason to favor the simulation of unfolding is that one starts with the well-defined native state (usually taken from the Protein Data Bank, PDB) in contrast to folding where one may have difficulty in finding proper initial states out of virtually infinitely many. There are, however, a few cases of short chain peptides where folding can be observed. Ho and Dill simulated the folding of a number of very short peptides [11] which provide a good starting point for selecting fragments that reliably fold into well-defined structures. There are at least two reasons for choosing a short peptide: The nanosecond simulation time scale is expected to match the real life folding and unfolding of a peptide and, some peptides tend to mimic proteins as far as the folding/unfolding properties are concerned [12–14].

One important structural motif in proteins is the β-hairpin. To form a β-hairpin a peptide has to provide two sufficiently long potential β-strands and a turn-sequence, thus requiring a minimum length. We chose chignolin (sequence: GYDPETGTWG), a synthetic peptide that has experimentally been shown to fold reliably into a well-defined native β-hairpin structure in water [15,16]. The folding and unfolding processes have been examined in several MD-simulation studies, confirming that the folding can be observed in silico. These studies focused on the folding free-energy landscape [17] the main interaction partners [15,18], reproducing the experimentally observed conformation [15], folding kinetics [19] or the key steps of the folding mechanism [18], respectively. Here, we are mainly interested in identifying the harmonic modes and changes in these modes related to folding and unfolding events. We especially focus on the dependence of the folding sequence on different non-native initial conditions.

A second peptide we are studying here is a related, short sequence (DDATKTFT) of the immunoglobulin G-binding protein G that has been reported to fold into a hairpin-turn structure [11] although

* Corresponding author. Tel.: +1 530 752 5839; fax: +1 530 752 1032.
E-mail addresses: astrid.maass@scai.fraunhofer.de (A. Maaß), dcalisir@metu.edu.tr (E.D. Tekin), anton.schueller@scai.fraunhofer.de (A. Schüller), anpalazoglu@ucdavis.edu (A. Palazoglu), dirk.reith@scai.fraunhofer.de (D. Reith), rfaller@ucdavis.edu (R. Faller).
[1] These authors contributed equally.

1570-9639/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbapap.2010.06.019

there is some debate [20]. Here, the two C-terminal residues are dangling, i.e., do not appear to have interaction partners in other sections of the strand marking the minimum of sequence required for inducing a stable reversal of the peptide backbone. We again identify the harmonic modes and additionally study their dependence on temperature and sequence length from 6 to 8 amino acids. Thus, we are studying not only this sequence but also two shorter variants.

Our work is in part based on the Karhunen–Loeve (KL) expansion [21–23] which is a special type of principal component analysis (PCA) to reduce the dimensionality of the data. PCA is an extremely useful tool to extract meaningful spatiotemporal information from the data. In our case, PCA helps us identify the relevant, small number of modes or degrees of freedom of the atomic fluctuations of the protein out of the entire simulation. In the literature, there have been several studies along this line [10,24–27].

2. Methodology

2.1. Simulations

We perform atomistic MD simulations of the two peptides using the Gromacs code [28] with the united atom GROMOS96 43a1 force field [29] for the peptides and the SPC water model [30]. Simulations use periodic boundary conditions in all three directions with a 1 or 2 fs time step. Bond lengths were constrained by the linear constraint solver (LINCS) algorithm [31]. The simulations used the PME technique for the long range electrostatics [32] with a 0.12 nm Fourier grid spacing and a 1.2 nm cut-off radius. Unfolding of the 8-residue peptide (DDATKTFT) of immunoglobulin G-binding protein G was studied at five different temperatures using Gromacs 3.3.1: 290, 300, 325, 430, and 500 K. We also performed simulations at 300 K with the 6 and 7 residue versions DDATKT and DDATKTF. The initial conformation of the peptide was taken from the protein data bank (PDB code: 2OED, residues 46 to 51, 52 or 53 respectively). Two crystal structures are available for the immunoglobulin G-binding protein G, 2OED and 2 GB1. The overall RMSD between the C_α atoms of these structures is only 0.111 nm such that they represent the same fold. The peptide was confined into a cubic box of water molecules. The energy of the system was initially minimized using steepest descent. After the minimization, the resulting system was heated to five different temperatures independently, keeping pressure and temperature fixed (*NPT* ensemble) using a Berendsen thermostat with a coupling time constant of 0.1 ps and an anisotropic Berendsen barostat [33] with a coupling time constant of 0.2 ps with reference pressure 1 bar in all directions and compressibility 1.12 10^{-6} bar^{-1}. The shift scheme was used to compute the short range van der Waals interactions with a 1 nm cut-off radius. Equations of motion were numerically integrated using the leap-frog algorithm [34]. The visual molecular dynamics (VMD) software [35] was used for displaying snapshots. The simulations were carried out for 9 ns at 290 K and for 20 ns at the other temperatures.

The same settings were applied for studying the folding and unfolding of chignolin starting from different initial conformations. The respective peptide input structures in terms of sequence and conformation were generated by tleap, a program of the Amber molecular simulation package, version 9 [36]. The entire simulation system was composed of the decapeptide immersed in 588–1110 water molecules (depending on the boxsize required to accommodate the diverse input conformations) and subjected to *NVT*-simulations before switching to *NPT*. Simulations were carried out for 20–50 ns. For analyzing the trajectories with respect to the evolution of energy terms, hydrogen-bonding network or backbone conformation, auxiliary programs of the Gromacs package as well as from the Amber package [37], were employed.

2.2. Principal component analysis

MD simulations produce a large amount of data on the dynamics of the protein molecule during folding or unfolding processes. Simple time series plots or 3D movies, while useful, are not very effective means to extract correlations among backbone atoms and identify key mechanistic patterns. Principal components analysis (PCA) or Karhunen–Loeve expansion (KLE) has been used extensively to explain the salient features of MD trajectories [24–26,38]. The goal of PCA is to reduce the dimensionality of large scale data sets and summarize the motion in a small number of modes. For this analysis, only the backbone C_α atoms are considered here. Before the data is analyzed the center of mass motion has to be removed to eliminate the trivial translation modes. Furthermore, we rotate all molecules into a common frame so that rotational degrees of freedom are removed and the x direction is the direction of the largest eigenvalue of the gyration tensor. To accomplish this, we determine at every simulation step the axes of the gyration tensor and align them with the Cartesian coordinates. The resulting matrix has the dimensions $M \times 3N$, where M represents the number of time steps and N is the number of residues in the peptide chain. Often, we find $M >> 3N$. The final form of the data matrix can be expressed as the following array:

$$R = \begin{bmatrix} x_1(t_1) & x_2(t_1) & \dots & y_1(t_1) & y_2(t_1) & \dots & z_N(t_1) \\ x_1(t_2) & \vdots & \dots & \vdots & \vdots & \vdots & \vdots \\ \dots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots \\ x_1(t_M) & x_2(t_M) & \dots & y_1(t_M) & y_2(t_M) & \dots & z_N(t_M) \end{bmatrix} \quad (1)$$

Then, the elements of the spatial covariance matrix Φ for Eq. (1) are computed from:

$$\Phi_{ij} = \frac{1}{M} \sum_{m=1}^{M} R_i(t_m) R_j(t_m) \quad (2)$$

An eigenvalue decomposition of Φ yields,

$$\Phi \phi_j = \lambda_j \phi_j \quad (3)$$

where λ_j is the j^{th} eigenvalue, and ϕ_j is the j^{th} eigenvector of Φ, respectively. ϕ_j is also referred to as the spatial eigenvector or spatial mode of the matrix. The matrix R can be expressed in terms of its finite expansion as

$$R(t_m) = \sum_{j=1}^{3N} c_j(t_m) \phi_j \quad m = 1, 2, \dots M \quad (4)$$

The temporal mode $c_j(t_m)$ represents the time varying amplitude of the spatial mode and is computed by projecting the data onto the spatial modes

$$c_j(t_m) = R^T(t_m) \phi_j \quad (5)$$

If this calculation is repeated for all sampling times, $m = 1, 2, \dots M$, one gets the time series $c_j = [c_j(t_1)\ c_j(t_2) \dots c_j(t_M)]$ for the j^{th} modal amplitude. The modal amplitudes c_j are zero mean, and orthogonal, i.e., $\langle c_j(t) \rangle = 0$ and $\langle c_i^T c_j \rangle = \lambda_i \delta_{ij}$.

The PCA, or KLE, will have K modes (eigenvector directions) and each eigenvalue measures the mean amplitude of projection of its corresponding mode and establishes the relative importance of the mode. Among the class of all linear expansions, KLE is optimal in the sense that, on a subspace of lower dimension $K << 3N$, it offers optimal fidelity. As each mode explains a certain amount of variance