Biophysical Journal, Volume 97

Supporting Material

How does a simplified-sequence protein fold?

Enrico Guarnera, Riccardo Pellarin, and Amedeo Caflisch

How does a simplified-sequence protein fold? SUPPLEMENTARY MATERIAL

Enrico Guarnera, Riccardo Pellarin, and Amedeo Caflisch*

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
FAX 0041 44 635 68 62

(Dated: July 3, 2009)

I. MOLECULAR DYNAMICS SIMULATIONS

All simulations and most of the analysis of the trajectories were performed with the program CHARMM (1); the rest of the analysis was done with the program WORDOM (2), which is particularly efficient in handling large sets of trajectories. Protein ssG was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field (3) with the default cutoff of 7.5 Å for the nonbonding interactions). A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent (4). More explicitely, the screening of the electrostatic interactions is approximated by the distance-dependent dielectric function $\epsilon(r) = 2r$, while the remaining solvation effects are approximated by replacement of the monopole moment of charged groups by strong dipole moments and a linear function of atomic SAS values. The latter requires only two surface-tension like parameters and takes into account both polar and apolar solvation effects by a negative (i.e., favorable) value of the surface-tension parameter for nitrogen and oxygen atoms, and a positive (unfavorable) value for carbon and sulfur atoms. The coarsest approximation in the SAS model is the neutralization of charged groups but these groups are not present in the ssG variant of protein G which has one polyAla and two polyThr-Gly-Gly-polyThr segments spanning the central α-helix and terminal β-hairpins, respectively. In other words, if one neglects the two termini there are only four different functional groups in protein ssG: Secondary amide and methylene in backbone, and methyl and hydroxyl in the side chains.

Despite the neglect of collisions with water molecules (frictional effects) in the simulations with the implicit solvent model, the *relative* rates of folding for different secondary structural elements are comparable with the values observed experimentally; i.e., helices fold in about 1 ns (5), β-hairpins in about 10 ns (5) and triple-stranded β-sheets in about 100 ns (6), while the experimental values are $\sim$0.1 μs, $\sim$1 μs and $\sim$10 μs , respectively (7; 8). A 15-μs molecular dynamics simulation of protein ssG was performed at 330 K which is a temperature at which the unfolded and molten-globule state are significantly populated (see Results). The temperature was kept constant by means of the Berendsen thermostat with time constant of 5 ps. A time step of 2 fs was used and the coordinates were saved every 20 ps for a total of 750000 snapshots. The simulation required about 18 weeks of a 2.8 Ghz Athlon CPU.

II. COARSE-GRAINING OF CONFORMATIONAL SPACE

A molecular dynamics trajectory is a long series of microscopic configurations each of which is visited only once. For this reason the analysis of the system needs a preliminary coarse-graining

*corresponding author: caflisch@bioc.uzh.ch

of the trajectory that allows the grouping of similar structures. There are several meaningful approaches that are known to efficiently achieve coarse-graining. For a system like protein ssG, root mean square deviation (RMSD) clustering and secondary structural symbolization are reasonable choices (9–12). In this work both approaches were used for different type of analysis. For the C_α-RMSD clustering we adopted the quality-threshold algorithm (13) in the version implemented in the program WORDOM (2). The choice of a clustering cutoff of 5 Å was particularly effective in capturing the salient structural motifs of the stable states of the ssG protein due to the high flexibility of the main chain. The number of clusters with two or more snapshots are 3124 and include 77% of the total sampling while the remaining 23% are unassigned structures. Not all the clusters can be considered statistically significant due to the finite size of the sampling. The statistical significance can be evaluated from the distribution of cluster sizes, namely the number of clusters $k(n)$ with n members (the cluster size distribution is shown in Fig. 1 of this document). The profile of this distribution follows a lognormal dependence, whose mean value gives the order of magnitude of the cluster sizes cutoff above which the clusters are statistically significant, which is $\gtrsim 100$.

As an alternative to RMSD clustering, strings of secondary structure can be employed to "symbolize" the trajectory of the protein ssG. According to the DSSP (14) code each residue can have one of eight symbols `-` (coil), `E` (extended strand in a β ladder), `S` (bend), `T` (hydrogen bonded turn), `B` (residue in isolated β-bridge), `G` (3_{10} helix), `H` (α helix), `I` (π helix). Hence each protein conformation is identified by an octal string (string of secondary structure, SSS[8]). The maximum number of strings for a polypeptide chain of 56 amino acid is $8^{54} \sim 10^{48}$, as the N-terminal and C-terminal residues have no assigned secondary structure.

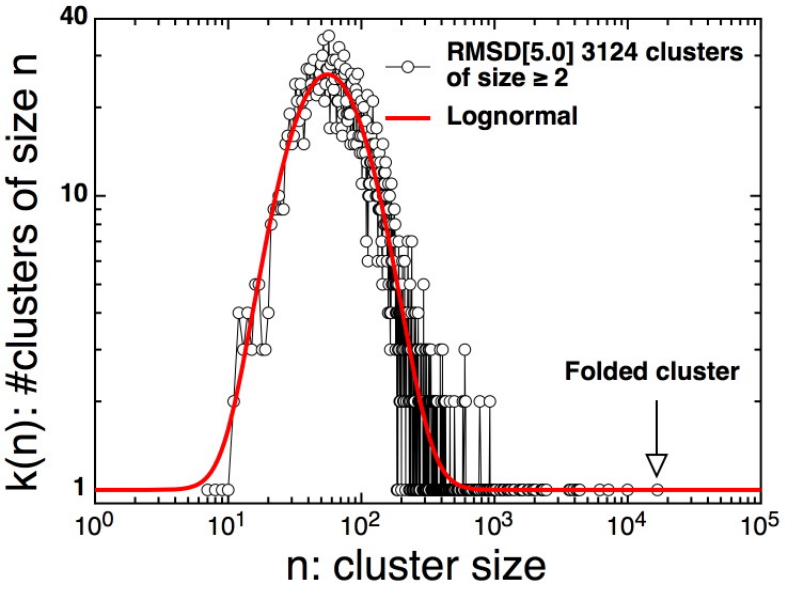

FIG. 1 Statistical significance of the clusters. The 5Å cutoff in C_α-RMSD and the quality-threshold algorithm used for clustering yielded 23% of unassigned conformers. A total of 3124 clusters were found with size ≥ 2. The distribution follows a lognormal profile. On the right tail for $n \gtrsim 100$ are the statistically significant clusters.

III. TESTS OF MARKOVIANITY

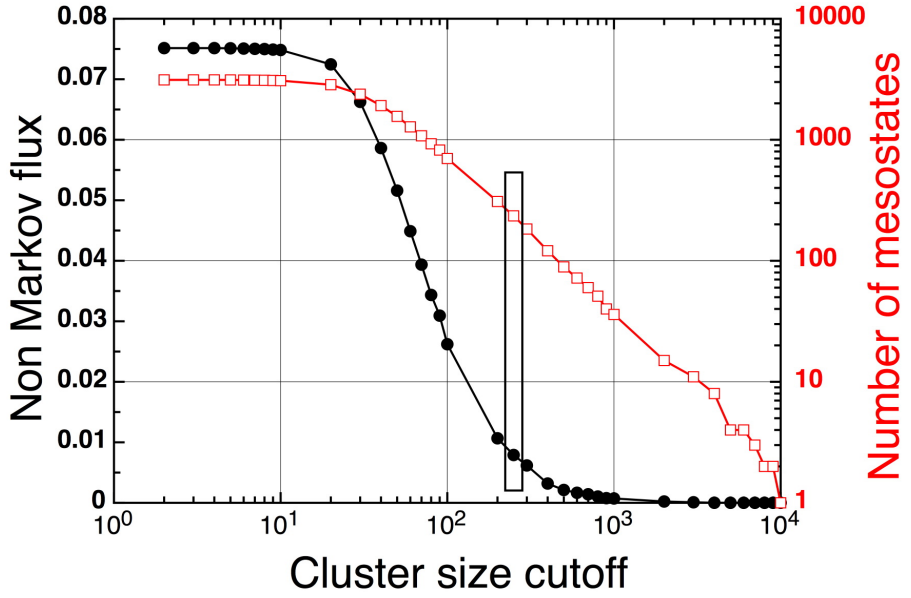

FIG. 2 Causal grouping of clusters. The causal grouped procedure, as explained in the main text of this paper, causally reassigns the conformers of the clusters whose size is less than a cutoff, to the clusters with size greater than the cutoff. The black curve in figure shows the values of the non-Markov flux on the new time series obtained with the causal grouping at a certain value of the cutoff. The curve has a sigmoidal shape with a midpoint corresponding to cluster size ~ 70 a flux 0.04. For cluster size ≥ 250 (rectangular box) the non-Markov flux is less than 0.01, which means that only about the 1% of the pathways in the new causal grouped time series are affected by long memory effects. There are 211 mesostates corresponding to a cluster size of 250.

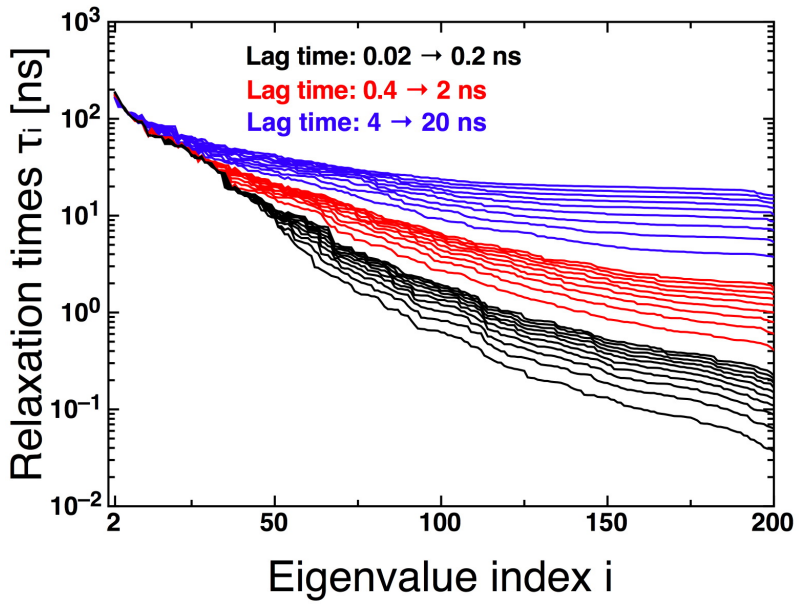

FIG. 3 Relaxation times of the decaying modes from the 200 causally-grouped mesostates. The slowest relaxations (i.e., indices 2-30) are robust with respect to changes in the lag time up to 20 ns. Note that a lag time of 20 ps was used for the Markov state model in the main text. To obtain the relaxation times from the transition matrices we calculated the reciprocal of the eigenvalues for the rate matrices $\mathbf{K}(\tau) = \mathbf{1} - \mathbf{T}(\tau)$, where $\mathbf{1}$ is the identity matrix.

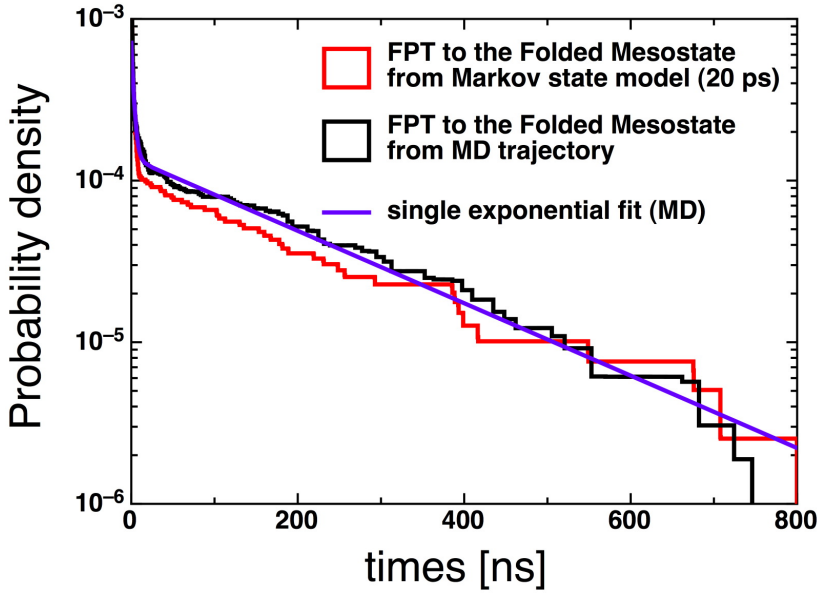

FIG. 4 Distribution of first passage time to the folded mesostate extracted directly from the MD trajectory, i.e., from the time series of causally grouped mesostates (black), and calculated by the Markov state model with a lag time of 20 ps (red). The solid line is a single exponential fit of the MD data. Note that folding is only slightly faster with the Markov state model than in the MD trajectory.

IV. THE MATRIX OF FLUXES BETWEEN MESOSTATES

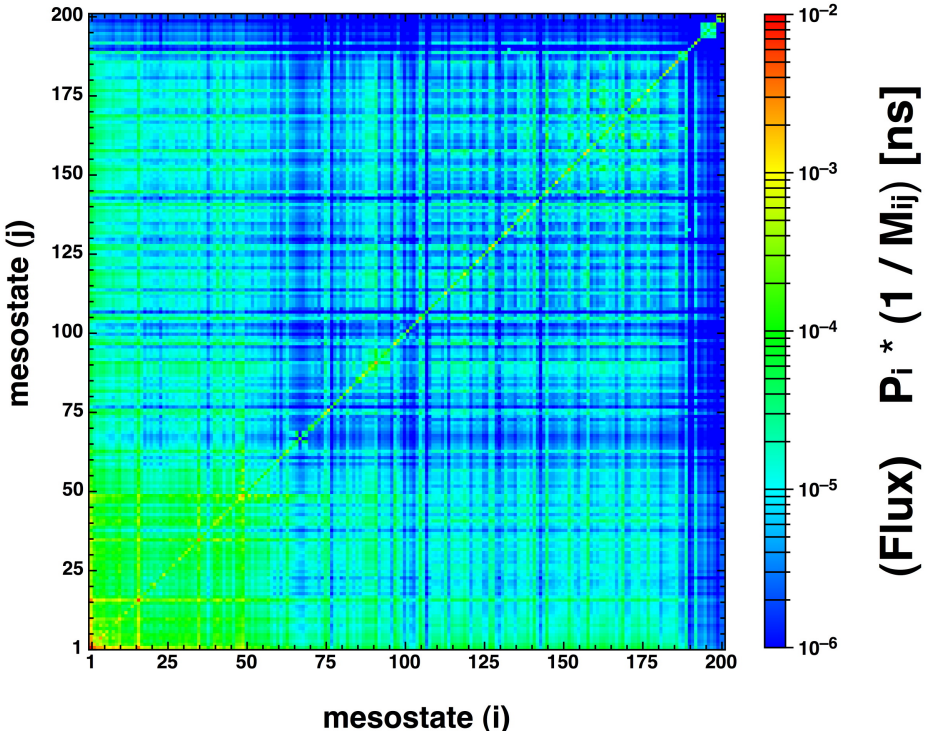

FIG. 5 The matrix of the probability fluxes for the 200 causally grouped mesostates. The matrix is obtained from the MFPT matrix shown in Figure 4 of the main text by defining the flux as $P_i^{eq} K_{ij}$ where the rate K_{ij} for the transition $i \to j$ is given by the reciprocal of the MFPT M_{ij} and P_i^{eq} is the steady state probability of the ith mesostate. The matrix is symmetric beacuse of detailed balance. It clearly shows high exchange fluxes for the mesostates kinetically close to the folded one, i.e., mesostates 1-60 (folded basin "■" and the basin with the two hairpins flipped "■" in Figure 3 of the main text). The β-rich kinetic traps (e.g., mesostates 198 and 200) have the lowest fluxes to folded and unfolded mesostates. The unfolded mesostates mainly stabilized by the entropy have id from 110 to about 180 and show interconversion fluxes in average lower than those within the folded state (e.g., unfolded basins "■" and "■" in Figure 3 of the main text).

References

[1] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, CHARMM - a program for macromolecular energy, minimization, and dynamics calculations, J. Comput. Chem. 4 (1983) 187–217.

[2] M. Seeber, M. Cecchini, F. Rao, G. Settanni, A. Caflisch, Wordom: a program for efficient analysis of molecular dynamics simulations, Bioinformatics 23 (2007) 2625–2627.

[3] E. Neria, S. Fischer, M. Karplus, Simulation of activation free energies in molecular systems, J. Chem. Phys. 105 (1996) 1902–1921.

[4] P. Ferrara, J. Apostolakis, A. Caflisch, Evaluation of a fast implicit solvent model for molecular dynamics simulations, Proteins: Structure, Function, and Bioinformatics 46 (2002) 24–33.

[5] P. Ferrara, J. Apostolakis, A. Caflisch, Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations, J. Phys. Chem. B 104 (20) (2000) 5000–5010.

[6] G. Settanni, F. Rao, A. Caflisch, Φ-Value analysis by molecular dynamics simulations of reversible folding, Proc. Natl. Acad. Sci. USA 102 (2005) 628–633.
[7] W. A. Eaton, V. Munoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, J. Hofrichter, Fast kinetics and mechanisms in protein folding., Ann. Rev. Biophys. Biomol. Struct. 29 (2000) 327–359.
[8] E. De Alba, J. Santoro, M. Rico, M. A. Jiménez, De novo design of a monomeric three-stranded antiparallel beta-sheet, Prot. Sci. 8 (4) (1999) 854–865.
[9] F. Rao, A. Caflisch, The protein folding network., J. Mol. Biol. 342 (1) (2004) 299–306.
[10] S. V. Krivov, M. Karplus, Hidden complexity of free energy surfaces for peptide (protein) folding, Proc. Natl. Acad. Sci. USA 101 (2004) 14766–14770.
[11] I. A. Hubner, E. J. Deeds, E. I. Shakhnovich, Understanding ensemble protein folding at atomic detail, Proc. Natl. Acad. Sci. USA 103 (2006) 17747–17752.
[12] J. Ihalainen, B. Paoli, S. Muff, E. Backus, J. Bredenbeck, G. Woolley, A. Caflisch, P. Hamm, Alpha-Helix folding in the presence of structural constraints., Proc. Natl. Acad. Sci. USA 105 (28) (2008) 9588–93.
[13] L. Heyer, S. Kruglyak, S. Yooseph, Exploring Expression Data: Identification and Analysis of Coexpressed Genes, Genome Research 9 (11) (1999) 1106.
[14] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.