Supplementary Material

Estimation of protein folding probability from equilibrium simulations

Francesco Rao, Giovanni Settanni, Enrico Guarnera, and Amedeo Caflisch

Department of Biochemistry,
University of Zurich,
Winterthurerstrasse 190,
CH-8057 Zurich, Switzerland
tel: +41 44 635 55 21,
fax: +41 44 635 68 62,
e-mail: caflisch@bioc.unizh.ch

I. SECONDARY STRUCTURE CLUSTERIZATION

Recently, the secondary structure has been used to cluster the conformation space of peptides (F. Rao et al, JMB 342, 299, 2004). Secondary structure along an MD simulation trajectory can be easily calculated using known algorithms (C.A.F. Andersen et al, Structure 10, 174, 2002). A cluster is a single string of secondary structure, e.g., the most populated conformation for beta3s is `-EEEESSEEEEEEESSEEEE-` where "`E`", "`S`", and "`-`" stand for extended, turn, and unstructured, respectively. There are 8 possible "letters" in the secondary structure "alphabet": "`H`", "`G`", "`I`", "`E`", "`B`", "`T`", "`S`", and "`-`", standing for α helix, 3/10 helix, π helix, extended, isolated β-bridge, hydrogen bonded turn, bend, and unstructured, respectively. Since the N- and C-terminal residues are always assigned an "`-`" a 20-residue peptide can in principle assume $8^{18} \simeq 10^{16}$ conformations.

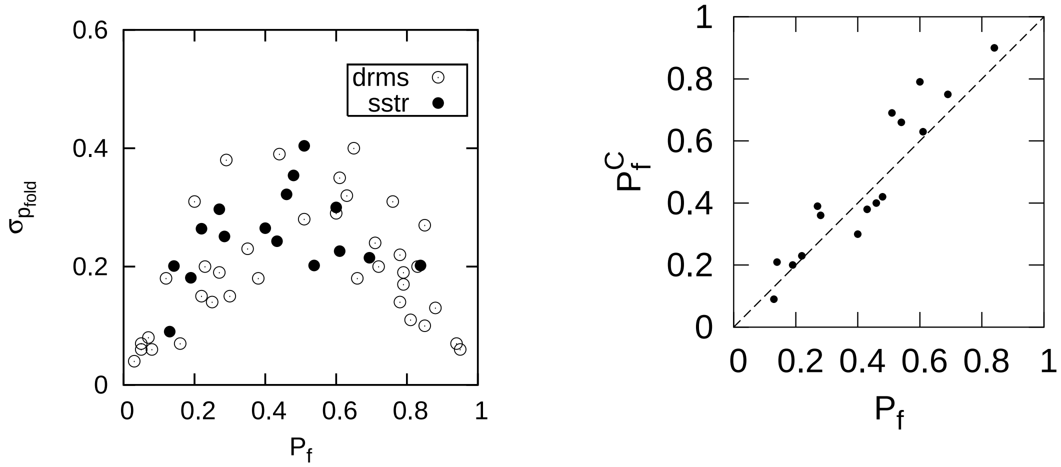

FIG. S1: **(left)** p_{fold} standard deviation inside a cluster for 16 secondary structure (*sstr*) and 37 DRMS 1.2 Å clusters. Both *sstr* and DRMS 1.2 Å clusterizations are defined by similar fluctuations. **(right)** Scatter plot of P_f^C versus P_f for *sstr* clusterization. In this case the folding criteria used is based on the native contacts Q (Settanni et al., PNAS 102, 628, 2005). A folding (unfolding) event is realized when $Q > 0.85$ ($Q < 0.15$).

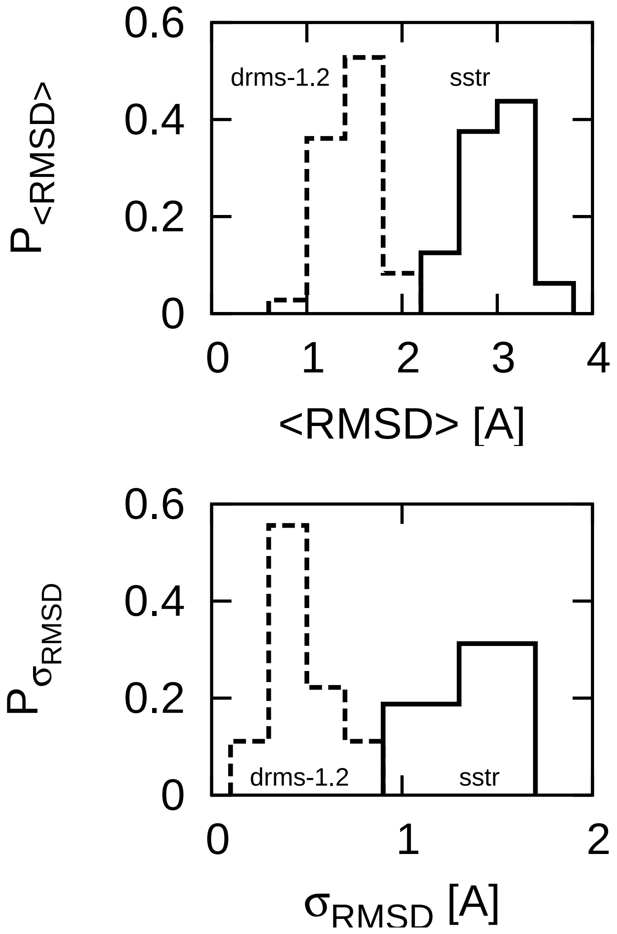

FIG. S2: **(top)** Probability to have a given pairwise root mean square deviation (RMSD) inside a cluster for the secondary structure (*sstr*) and DRMS 1.2 Å clusterizations. **(bottom)** Probability to have a given variance for the RMSD inside a cluster. Both plots show that secondary structure clusters are less structurally homogeneous than DRMS 1.2 Å clusters.

II. FIRST PASSAGE TIMES

The first passage time (fpt) to a given cluster α is computed as the time along the MD trajectory that any given snapshot takes to the first subsequent snapshot belonging to α. In fig. S3 the fpt distribution to the folded state is shown for two different clusterizations of the conformation space. The double peak shape of the distribution provides evidence of the different time scales between *intra*-basin and *inter*-basin transitions. The wider shape of the *intra*-basin peak for the secondary structure clusterization is consistent with the higher degree of structural diversity with respect to the DRMS 1.2 Å clusterization (see previous section).

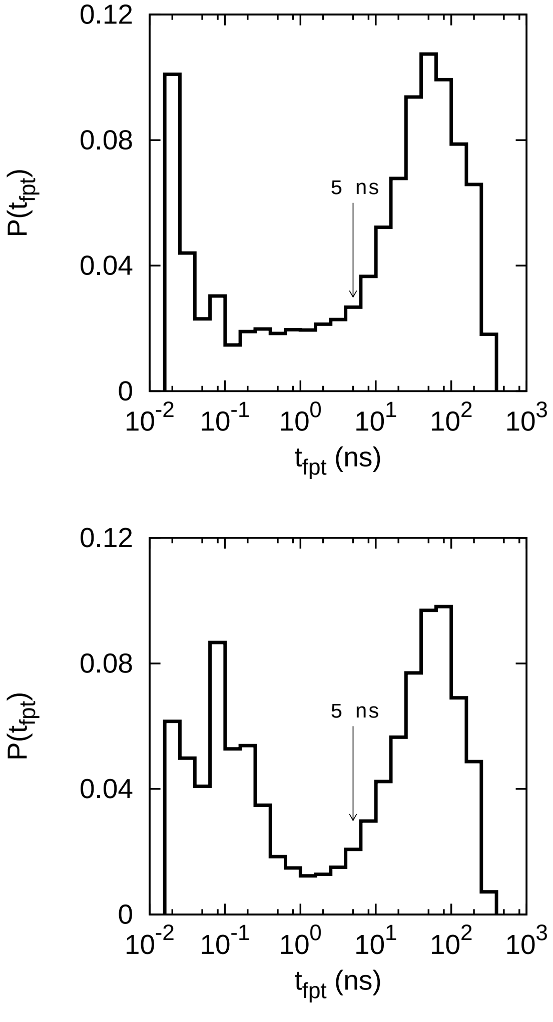

FIG. S3: Probability distribution for the first passage times (fpt) to the most populated cluster (*folded state*). **(top)** DRMS 1.2 Å clusterization. **(bottom)** Secondary structure clusterization.

III. RANDOM CLUSTERIZATION

The results of this section were obtained using the DRMS 1.2 Å clusterization. In the text evidence was provided that the standard deviation of p_{fold}

$$\sigma_{p_{fold}} = \sqrt{\langle (p_{fold}(i) - P_f[\alpha])^2 \rangle_{i \in \alpha}}$$

is not compatible with the one of a Bernoulli distribution. This means that snapshots in a cluster have similar values of p_{fold} and are kinetically homogeneous. This is not the case for a random clusterization of the snapshots. Since it is not feasible to compute the p_{fold} for every snapshot of a simulation, the assumption that p_{fold} of snapshot i is equal to the cluster folding probability P_f^C of its cluster (as computed in the text) is made. Then, snapshots are reshuffled in 50000 random clusters. The folding probability for a random cluster α_R is computed as $P_f = \langle p_{fold} \rangle_{\alpha_R}$. Most of the snapshots will have p_{fold} close to 1 or 0 (see Fig. 3B in the text) and because of the random grouping, i.e., no kinetic homogeneity, the above standard deviation $\sigma_{p_{fold}}$ resembles the one of a Bernoulli distribution as shown in Fig. S4. Data obtained from a DRMS 1.2 Å clusterization deviates from this behavior (compare Fig. 2A and Fig. S4). Moreover this deviation becomes bigger as the number of trials n_t, in this case 10, increases (see Fig. 1B).

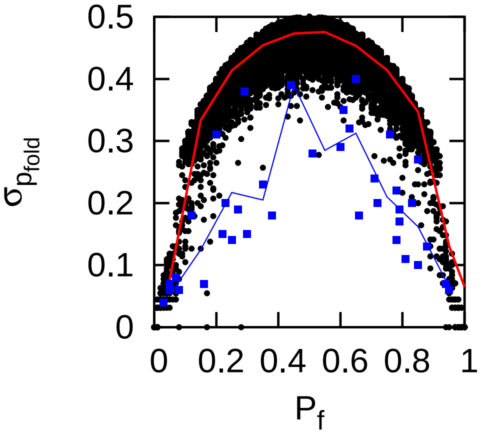

FIG. S4: Standard deviation $\sigma_{p_{fold}}$ for a random clusterization. Black dots, red curve, blue squares, blue curve show $\sigma_{p_{fold}}$ for the random clusters, its histogram, $\sigma_{p_{fold}}$ for 37 non-random clusters (see text), and its histogram, respectively.