Supporting Information: Interplay between secondary and tertiary structure formation in protein folding cooperativity

Tristan Bereau,[†] Michael Bachmann,[‡] and Markus Deserno[†*]
[†] *Dept. of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA, and*
[‡] *Institut für Festkörperforschung, Theorie II, Forschungszentrum Jülich, 52425, Jülich, Germany*

I. COARSE-GRAINED PEPTIDE MODEL

Coarse-grained (CG) simulations were performed using an intermediate resolution, implicit solvent peptide model [1]. This model is aimed at reproducing the general behavior of various amino acid sequences without biasing the force field towards any particular structure. An amino acid is represented by three or four beads (Fig. S1). These beads represent the amide group N, central carbon C_α, carbonyl group C', and (for non-glycine residues) a side chain C_β. The first three beads belong to the backbone, whereas the last one represents the side chain and is responsible for amino acid specificity. The high level of resolution devoted to the backbone allows for accurate representations of secondary structure.

The long-range behavior is governed by nonbonded interactions: (1) Excluded volume for all beads is modeled via a Weeks-Chandler-Andersen (WCA) potential [2]; (2) Side chain-side chain interactions need to compensate for the lack of explicit solvent. This is achieved by a phenomenological cohesion that will depend on the involved amino acid propensities. This effective attraction (of standard Lennard-Jones 12-6 type) maps its strength to the Miyazawa-Jernigan matrix [3], a statistical analysis of residue-residue contacts. Thereby, hydrophobic side chains attract more than hydrophilic ones; (3) An important contribution to protein stabilization comes from hydrogen bonds. As the model does not contain any electrostatics, it is necessary to model hydrogen bonds effectively. A radial 12-10 Lennard-Jones potential combined with an angular term provides a simple representation of this interaction between amide and carbonyl groups. Note that the level of resolution of the model does not allow for side chain hydrogen bonds; (4) A nearest-neighbor dipole-dipole interaction permits the stabilization of β-conformations. After parametrization (see original paper for further details [1]), the model was shown capable of folding several simple helical proteins and reproduce an oligopeptide aggregation scenario using the same force field without secondary structure bias.

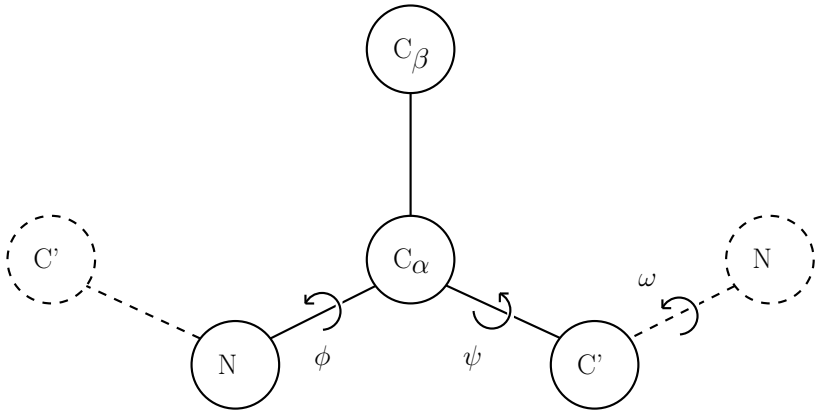

FIG. S1: Schematic figure of the local geometry of the peptide chain. The solid beads represent one amino acid. Neighboring amino acid beads are shown in dashed lines. The backbone dihedrals ϕ, ψ, and ω involve the beads C'NC_αC', NC_αC'N, and C_αC'NC_α, respectively. Reprinted with permission from Bereau, T. and Deserno, M., *J. Chem. Phys.*, **130**, 235106 (2009). Copyright 2009, American Institute of Physics.

*Electronic address: deserno@andrew.cmu.edu

Peptide	Sequence
short helix	$(\texttt{AAQAA})_3$
long helix	$(\texttt{AAQAA})_{15}$
bundle α3D	MGSWA EFKQR LAAIK TRLQA LGGSE AELAA FEKEI AAFES ELQAY KGKGN PEVEA LRKEA AAIRD ELQAY RHN

TABLE S1: Amino acid sequences used in this work. The three helical regions of the native state (from NMR structure, PDB 2A3D) of the helix bundle α3D [15] are underlined (as predicted by STRIDE [16]).

The peptides' amino acid sequences studied in this work are listed in Tab. S1. Besides several other examples already discussed in Ref. 1, we find that the identified lowest energy conformations of $(\texttt{AAQAA})_3$ and α3D agree well with experimentally determined native states: Scholtz *et al.* have shown that $(\texttt{AAQAA})_3$ indeed forms a water-soluble α-helix using circular dichroism [12]; and the CG model was parametrized to reproduce the NMR structure of α3D (PDB: 2A3D) [15], leading to a root mean square deviation of 4 Å between the CG ground state and the experimentally determined structure [1]. However, we are not aware of any detailed structural study of $(\texttt{AAQAA})_{15}$ and thus an experimental verification for our prediction of a stable long helix would be desirable.

The results presented in this Communication ultimately depend on the accuracy of the force field, the most important limitation being here the treatment of side chains. A single C_β bead, calibrated energetically with a statistical potential, represents the sterics and interactions of an entire side chain. This limits the ability of the model to reproduce hydrophobic cores correctly. Still, certain sequences do fold because side chain packing is only of limited importance in simple topologies; instead, hydrogen-bonding and backbone rearrangement prevail. Using statistical potentials—such as the Miyazawa-Jernigan matrix [3]—offers a valuable approach to imprint amino acid specificity. It was shown that different, though correlated, statistical potentials can widely affect the stability of ground state (lattice) conformations [13]. On the other hand, *optimized* sequences, which are especially stable, can tolerate relatively large errors in the interaction potentials [14], and are therefore more robust. The helix bundle α3D can be thought of as such an optimized sequence as it is a *de novo* peptide, *i. e.,* the sequence was *designed* to fold into a 3-helix bundle. While the short and long helices were not optimized, their sequences are simple and rather featureless as they avoid strong side chain interactions (*e. g.,* charged or strongly hydrophobic residues). We therefore expect that the combination of a finely resolved, generic coarse-grained model with such sequences is likely to provide robust results, at least qualitatively.

In finite-size systems, the barrier and critical temperature associated with a transition depend on the reaction coordinate used [9]. The results presented here as a function of energy thus cannot be trivially extended to any other order parameters. We choose to emphasize the connection between entropy and energy, represented by $S(E)$, as it is particularly suited to obtain information on finite-size thermodynamic transitions. The strength of a microcanonical analysis is its ability to unambiguously distinguish between discontinuous and continuous transitions using a quantitative measure: the latent heat ΔQ, which is nonzero for a discontinuous and zero for a continuous transition.

Sequence effects are striking when comparing $(\texttt{AAQAA})_{15}$ with α3D. While the former folds into a single α-helix, the latter forms a three-helix bundle. The difference between the two folds can be explained by the energetics of the system: 1) alanine, which covers much of $(\texttt{AAQAA})_{15}$, is only weakly hydrophobic (normalized hydrophobicity of 26%, whereas leucine—the most hydrophobic residue according to the parametrization [1]—is set to 100%); 2) the creation of a turn is sterically hindered for most amino acids, except glycine and proline [1]; 3) hydrogen-bond formation tends to favor helical conformations over turns or coils because it lowers the system's energy. Overall, the rupture of a helix into fragments is not favorable for the (ground state) long helix because the cohesive energy gained by alanine side chain interactions does not compensate for the rupture of hydrogen bonds and the steric constraints around the turn regions. On the other hand, the presence of glycine residues at the turn regions and various strongly hydrophobic amino acids drive α3D into a three-helix bundle. This can be observed best by looking at the helix bundle's sequence, as shown on Tab. S1, where the helix regions (underlined residues) contain several phenylalanines, leucines, and isoleucines—three of the most hydrophobic amino acids—and the turn regions between the helices are composed of several glycines and one proline.

II. SIMULATION METHOD

Molecular dynamics simulations were performed using the ESPResSo package [4]. All simulations were run in the canonical ensemble (NVT) using a Langevin thermostat with friction constant $\Gamma = \tau^{-1}$, where τ is the intrinsic unit of time of the CG model. The CG unit of energy $\mathcal{E}$ relates to thermal energy, where we define $\mathcal{E} = k_\mathrm{B}T_\mathrm{room} = 1.38\times10^{-23}$ J K^{-1} 300 K $\simeq 4.1 \times 10^{-21}$J $\simeq 0.6$ kcal mol^{-1}. In parallel, parameters of the force field were tuned to reproduce features of proteins observed at room temperature [1]. The temperature T was expressed in terms of the intrinsic

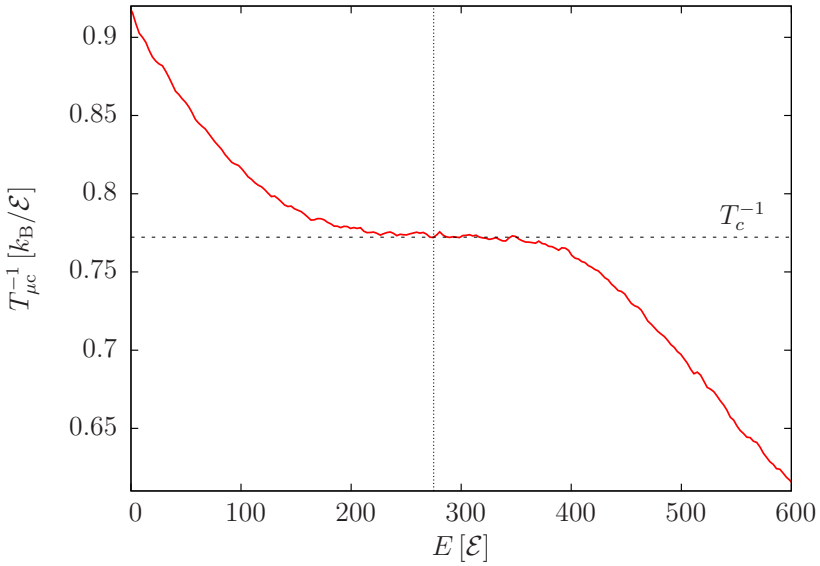

FIG. S2: Microcanonical inverse temperature $T_{\mu c}^{-1} = \partial S/\partial E$ for $(\texttt{AAQAA})_{15}$. The absence of a backbending loop indicates a continuous (*i. e.,* downhill) transition.

unit of energy: $T = \mathcal{E}/k_B$. The integration time step used for all simulations is $\delta t = 0.01\tau$.

III. ANALYSIS METHOD

Density of states calculations were achieved by simulating a given peptide at different temperatures, combined with the Weighted Histogram Analysis Method (WHAM) [5]. WHAM is a minimum variance estimator of the density of states $\Omega(E)$. It relies on the energy histograms that are sampled in various canonical simulations and merged to reconstruct a given energy interval of the density of states. From $\Omega(E)$, one can calculate the entropy of the system $S(E) = k_B \ln \Omega(E)$. In order to improve sampling, the abovementioned canonical simulations were coupled in a parallel tempering scheme, where coordinates are swapped according to a Metropolis criterion (this helps decorrelate the data points). For each peptide studied a total of 36 replicas were simulated, most of them were set close to the transition temperature in order to improve the estimate of the density of states around the transition region/point. For each replica, the simulation was run between 5×10^6 and $10^7\,\tau$ (depending on the peptide), and the potential energy was recorded regularly (every $100\,\tau$ on average to avoid correlations).

The canonical inverse temperature plot for the short helix was also obtained by using WHAM. Here, the average energy, denoted $\langle E \rangle_{\rm can}$ (canonical average), was calculated for a set of temperatures comprised within the temperature interval at which the system was simulated.

The order parameters reported in this Communication (R_g, $E_{\rm Hb}$, $E_{\rm sc}$) were sampled from the abovementioned canonical simulations. In order to analyze them in the microcanonical ensemble, they were binned in histograms according to the corresponding energy at which they were sampled (irrespective of the sampled temperature), and the average and error of the mean were calculated for each energy bin.

A microcanonical analysis of finite-size systems has shown more informative than its corresponding canonical counterpart for various systems [6–9], especially around first-order like transitions. Finite-size thermodynamic transitions, characterized by features in the derivative of the entropy $\partial S/\partial E$, allow for the determination of the microcanonical transition temperature: while a downhill transition (continuous) exhibits a flat region where $T_{\mu c}^{-1} = \partial S/\partial E$ is virtually constant (corresponding to the transition temperature), a two-state transition (discontinuous) will show a so-called backbending effect in the same quantity [9], reminiscent of a van-der-Waals loop (though here the loop is physical, rather than an artifact from a mean-field assumption), with equal areas for both loops at the transition temperature. The microcanonical inverse temperature of $(\texttt{AAQAA})_{15}$ is shown in S2, where the absence of a backbending loop indicates a continuous transition. We found that the microcanonical transition temperature is virtually identical to the corresponding location of the canonical specific heat peak (canonical transition temperature). The microcanonical temperatures obtained for each peptide are: short helix: $T = 1.187\,\mathcal{E}/k_B$; long helix: $T = 1.295\,\mathcal{E}/k_B$; α3D: $T = 1.242\,\mathcal{E}/k_B$.

Error bars on $\Delta S(E)$ were obtained by bootstrapping each energy histogram, thereby recreating a new distribution for each replica. WHAM was then used to calculate the corresponding density of states. This scheme was applied 50 times for each peptide to estimate the standard deviation of each point in $\Delta S(E)$.

For each peptide, calorimetric ratios were obtained by first extracting the canonical specific heat curve $C_{\mathrm{V}}(T)$ from WHAM, and then apply the calorimetric ratio κ_2 as defined from Kaya and Chan [10]. No baseline subtraction was performed.

Snapshots of peptide conformations were rendered with VMD [11].

---

[1] Bereau, T.; Deserno, M. *J. Chem. Phys.* **2009**, *130*, 235106
[2] Weeks, J.; Chandler, D.; Andersen, H. *J. Chem. Phys.* **1971**, *54*, 5237
[3] Miyazawa, S.; Jernigan, R. *J. Mol. Biol.* **1996**, *256*, 623
[4] Limbach, H.; Arnold, A.; Mann, B.; Holm, C. *Comp. Phys. Comm.* **2006**, *174*, 704
[5] Kumar, S.; Rosenberg, J.; Bouzida, D.; Swendsen, R.; Kollman, P. *J. Comput. Chem.* **1995**, *16*, 1339
[6] Deserno, M. *Phys. Rev. E* **1997**, *56*, 5204
[7] Junghans, C.; Bachmann, M.; Janke, W. *Phys. Rev. Lett.* **2006**, *97*, 218103
[8] Taylor, M. P.; Paul, W.; Binder, K. *Phys. Rev. E* **2009**, *79*, 050801
[9] Gross, D. *Microcanonical Thermodynamics: Phase Transitions in 'Small' Systems*; World Scientific Publishing Company, 2001
[10] Kaya, H.; Chan, H. *Proteins* **2000**, *40*, 637–661
[11] Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33
[12] Scholtz, J.; York, E.; Stewart, J.; Baldwin, R. *J. Am. Chem. Soc.* **1991**, *113*, 5102–5104
[13] Betancourt, M.; Thirumalai, D. *Protein Science* **1999**, *8*, 361–369
[14] Pande, V.; Grosberg, A.; Tanaka, T. *J. Chem. Phys.* **1995**, *103*, 9482
[15] Walsh, S.; Cheng, H.; Bryson, J.; Roder, H.; Degrado, W. *Proc. Natl. Acad. Sci.* **1999**, *96*, 5486–5491
[16] D. Frishman, P. Argos *Proteins*, **23**, 566-579 (1995)