Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary information for: "*Improving the analysis of biological ensembles through extended similarity measures*"

Liwei Chang,[a] Alberto Perez [*ab] and Ramón Alain Miranda-Quintana [*ab]

a. Department of Chemistry, University of Florida, Gainesville, FL 32611, USA.
b. Quantum Theory Project, University of Florida, Gainesville, FL 32611, USA.
Emails: quintana@chem.ufl.edu, perez@chem.ufl.edu

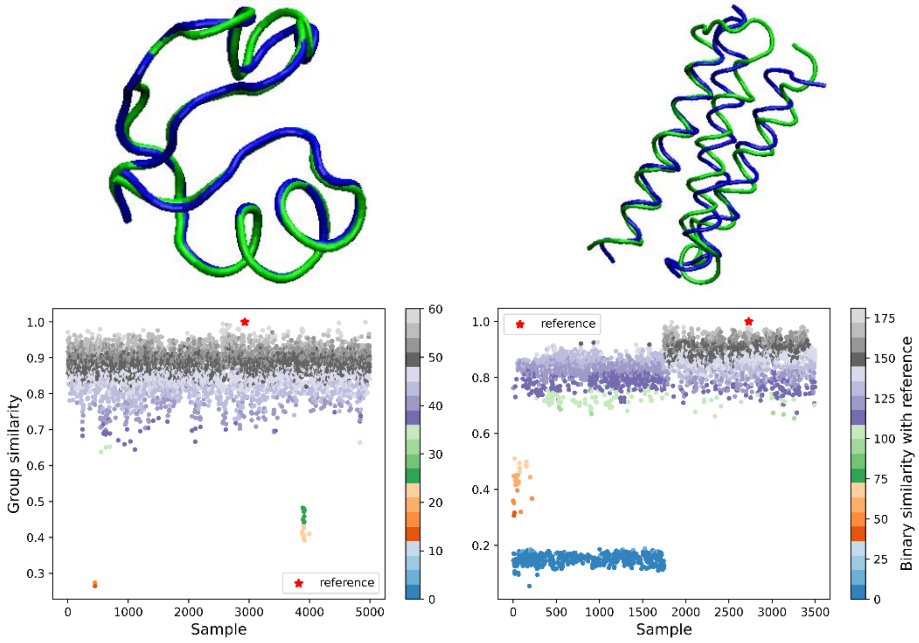

Figure S1. Top plots show the conformation with maximum group similarity for NTL9 (left) and A3D (right) in blue aligned to experimental structure in green. Bottom plots show the group similarity of all samples for ASTP-7041 (left) and 1NS1 (right).

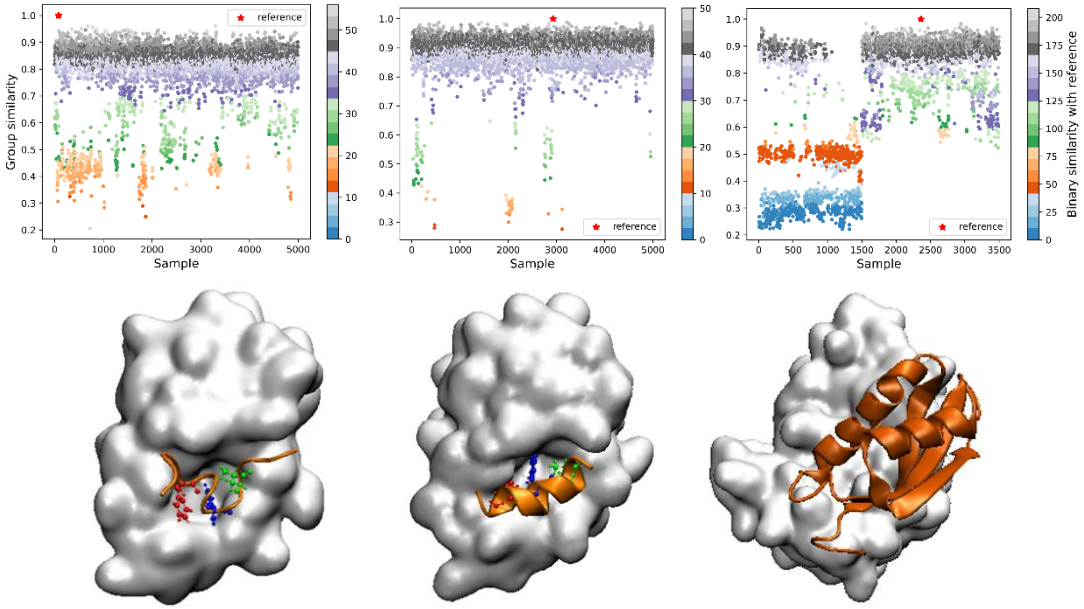

Figure S2. Additional tests for p53 (left), pdiq (middle) and 2MMA (right). Group similarity plots are shown on top and the bottom indicates the representative structures for each data set.

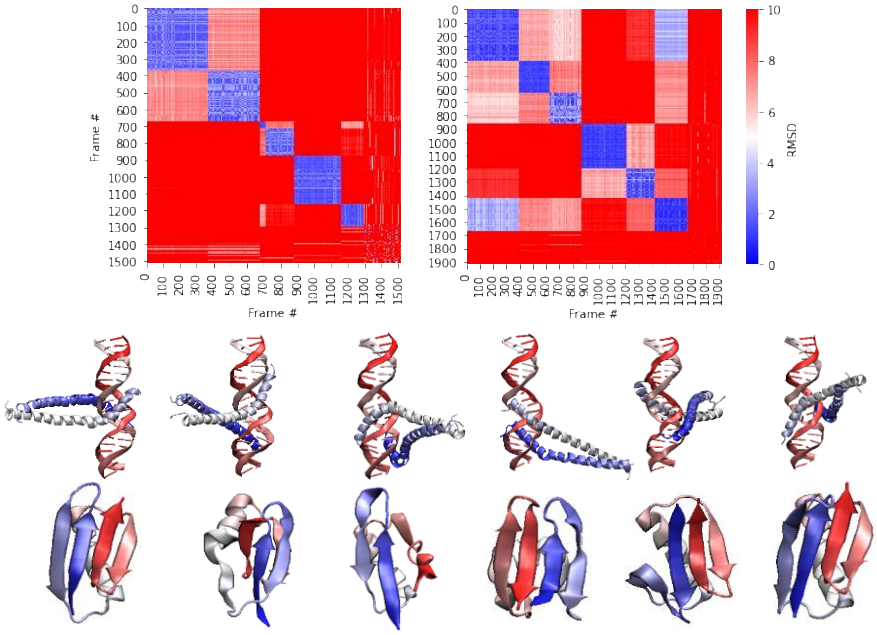

Figure S3**.** Pairwise RMSD plot of generated samples are shown on top for protein-DNA system (left) and protein G (right). Representatives of each cluster are shown at bottom. For each system, we picked six distinct seeds from MELD ensembles generated by selectively enforcing either distance restraints between protein and DNA or secondary structure restraints on alpha helix and two hairpins of protein G, then we choose the most similar conformations against each seed based on RMSD from the rest of ensemble. In addition, we added one noisy set to evaluate the robustness of each linkage criterion.

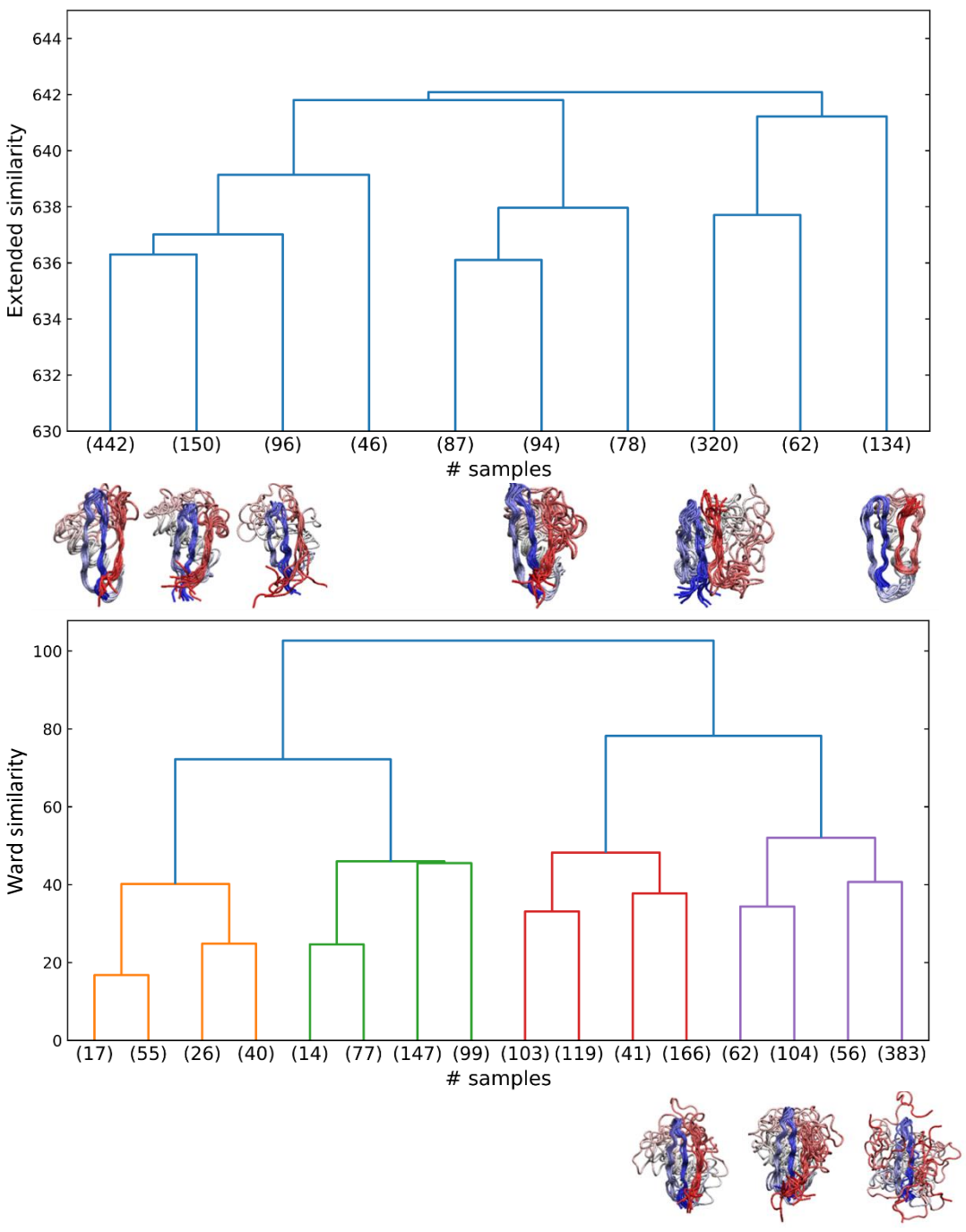

Figure S4. Hierarchical clustering result on intermediate state and representative samples for high population states by using extended similarity (top) and Ward linkage (bottom). The extended similarity value shown is subtracted by the maximum of all similarity values. The bottom plot shows the highest populated cluster using Ward linkage has high variance with many dissimilar conformations.

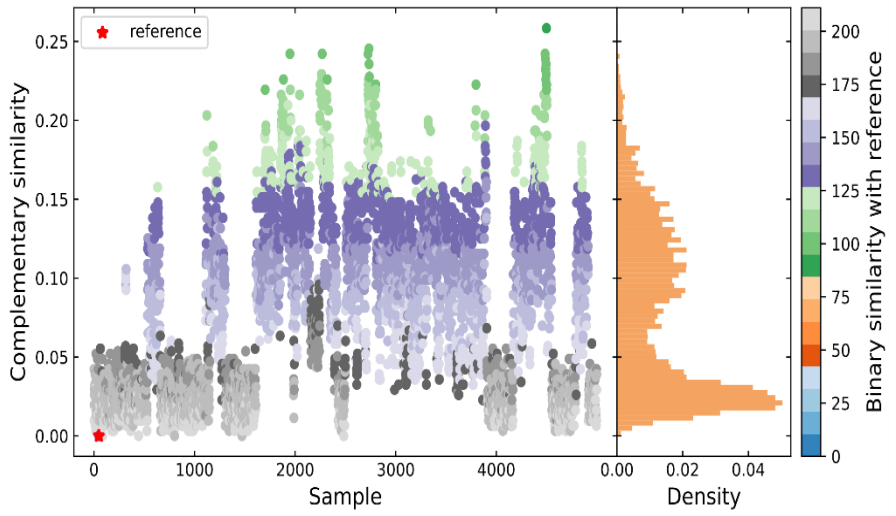

Figure S5. Selection strategy for NuG2 folding ensemble analysis. The color in the scatter plot represents binary similarity (the coincidence value of 1 in both fingerprints) of each sample with the reference conformation, which corresponds to the minimum of the complementary similarity (e.g., the medoid).

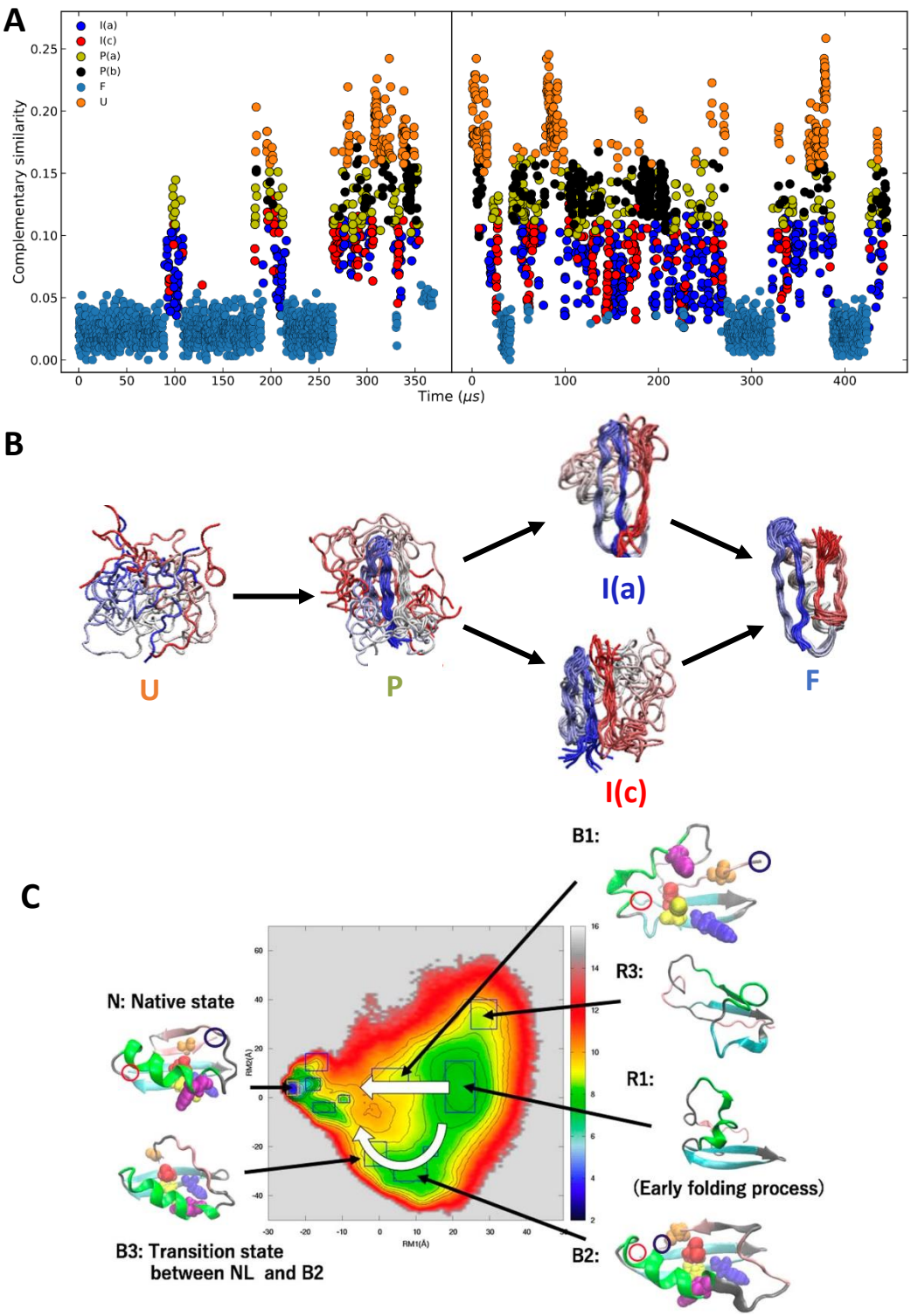

Figure S6. NuG2 folding pathways. (A) Time series of complementary similarity for samples in the most populated clusters, which clearly indicates two distinct pathways shown in (B), which is in agreement with (C) adapted from the recent work based on relaxation mode analysis by Mitsutake, A. and Takano, H. J. Chem Phys., 151, 044117, 2019 licensed under a Creative Commons Attribution (CC BY) license.