Defining a "folding" measurement for oligomers with arbitrary distributions

The purpose of this document is to introduce a measurement for the extent of "folding" in non-biological oligomers with arbitrary secondary structure.

As a review, we wish to define a measurement for the extent of "folding" so that we can: 1) quantify secondary structure formation, and 2) optimize model parameters (so that states corresponding to "folded" structures exhibit a lower potential energy than "unfolded" states). The standard approach for optimizing parameters to model protein folding is to minimize their Z-score:

$$Z_{score} = \frac{\left\langle U_{Non-Native} \right\rangle - U_{Native}}{\sigma_{Non-Native}}. \tag{1.1}$$

Here $\left\langle U_{Non-Native} \right\rangle$ is the mean potential energy for an ensemble of "non-native" configurations, and $\sigma_{Non-Native}$ is the standard deviation in the potential energy for the "non-native" ensemble.

Per our discussion this morning, 8/6/19, we are interested in identifying a Z-score-like measurement that avoids the need to define one "native" configuration as a reference (to account for models where there is not a well-funneled free energy surface), with a preference for using ensemble average structures and energies. Below is an attempt to define a folding metric which emulates the Z-score while avoiding the limitations of using a single reference configuration. More specifically, we define a "folding score" which can be used to evaluate the effectiveness of

a set of model parameters, where we define “effectiveness” as the ability of that set of parameters to distinguish “folded” structure from “unfolded” structures.

We first define a set of relevant thermodynamics states for our coarse grained model(s) by applying the MSMBuilder clustering machinery (K-centers) to our MD trajectories. Notably, one advantage of MSMBuilder is its ability to identify important states that are sampled infrequently, such as transition states and highly-funneled (folded) minimum energy structures. For these reasons we would expect MSMBuilder clustering approaches to provide a good approximation to the states in our coarse grained model simulations, including folded states. Furthermore, in contrast to other approaches (and like MBAR) MSMBuilder allows us to estimate the properties for states that are infrequently sampled.

We cluster configurations (assign them to thermodynamic states) based upon the RMSD fluctuations in their particle positions (RMSD < 0.2 nm) from the geometric center calculated for that state. The geometric center is also called a “centroid configuration.” We propose to classify individual thermodynamic states as “folded” or “unfolded” through further analysis of the statistical and energetic properties of their centroid configurations. More specifically, we define a thermodynamic state as “folded” if it satisfies either of the following criteria: 1) That thermodynamic state (cluster) has the lowest mean potential energy , or 2) the mean potential energy for that state is within a standard deviation of the energy for the lowest state. In order to evaluate the performance of a set of model parameters, with respect to the task of distinguishing “folded” and “unfolded” structures, we evaluate a linear combination of Z-scores for all “folded” structures:

$$Z_{fold} = \sum_{i=1}^{N_{folded}} \frac{\sum_{j=1}^{N_{unfolded}} W_j \left(U_j \left(\vec{x}_j \right) - U_i \left(\vec{x}_i \right) \right)}{\exp\left[-\left(U_i \left(\vec{x}_i \right) - U_{\min} \left(\vec{x}_{\min} \right) \right) \right] U_i \left(\vec{x}_i \right)}. \quad (1.2)$$

Here $U_i \left(\vec{x}_i \right)$ is the potential energy for thermodynamic state *i* at centroid configuration $\vec{x}_i$. W_j is the statistical probability (weight) for state *j* and can, in principle, be evaluated a variety of ways, including the number of counts for that state (simplest definition). In the denominator, the contributions of each "folded" state are modulated by the exponential of their energy difference from the lowest energy state. In the event that there is only one "folded" ("native") and one "unfolded" ("non-native") state, this definition simplifies to the Z-score. However, in the event that there are more than one of these states, this definition could enable more flexible optimization of our model parameters. For example, I could imagine an approach like this working well for a model that exhibits two degenerate and structurally distinct minima.