CONSTRUCTING MULTI-RESOLUTION MARKOV STATE MODELS (MSMS) TO ELUCIDATE RNA HAIRPIN FOLDING MECHANISMS

XUHUI HUANG†
Department of Chemistry, The Hong Kong University of Science & Technology
Kowloon, Hong Kong, China
Department of Bioengineering, Stanford University
Stanford, CA, 94305, U.S.A.

YUAN YAO
Department of Mathematics, Stanford University
Stanford, CA, 94305, U.S.A.

GREG R. BOWMAN
Biophysics Program, Stanford University
Stanford, CA, 94305, U.S.A.

JIAN SUN
Department of Computer Science, Stanford University
Stanford, CA, 94305, U.S.A.

LEONIDAS J. GUIBAS
Department of Computer Science, Stanford University
Stanford, CA, 94305, U.S.A.

GUNNAR CARLSSON
Department of Mathematics, Stanford University
Stanford, CA, 94305, U.S.A.

VIJAY S. PANDE
Department of Chemistry, Stanford University
Stanford, CA, 94305, U.S.A.

Simulating biologically relevant timescales at atomic resolution is a challenging task since typical atomistic simulations are at least two orders of magnitude shorter. Markov State Model (MSMs) provide one means of overcoming this gap without sacrificing atomic resolution by extracting long time dynamics from short simulations. MSMs coarse grain space by dividing conformational space into long-lived, or metastable states. This is equivalent to coarse graining time by integrating out fast motions within metastable states. Therefore, MSMs are inherently multi-resolution. Here we introduce a new algorithm Super-level-set Hierarchical Clustering (SHC), to our knowledge, the first algorithm capable of constructing MSMs at multi-resolutions. The key insight of this algorithm is to generate a set of super levels covering different density regions of the phase space, then cluster each super level separately, and finally recombine this information into a single MSM. SHC is able to produce MSMs at different resolutions using different super level sets. To demonstrate the power of this algorithm we apply it to a small RNA hairpin, and generate MSMs at four different resolutions. These MSMs are successfully validated by their ability to reproduce the original simulation data. Furthermore, long time folding dynamics can be extracted from these models. The results show that there are no metastable on-pathway intermediate states, while the folded state serves as a hub by directly connecting to multiple unfolded/misfolded states separated by large free energy barriers.

†To whom correspondence should be addressed. E-mail: xuhuihuang@gmail.com
1. Introduction

Conformational changes are crucial for a wide range of biological processes including protein folding, RNA folding and the operation of key cellular machinery. Extensive genetic, biochemical, biophysical and structural experiments can be performed to understand these conformational changes. However, probing the mechanisms of conformational changes at atomic resolution is very difficult experimentally and without these details it is impossible to understand the fundamental chemistry they perform. Computer simulations may complement such experiments by providing dynamic information at an atomic level. However, there is a gap between the timescales where interesting biologically relevant conformational changes occur (typically microseconds and up) and those we can simulate at atomic resolution (typically only tens of nanoseconds). The length of atomistic simulations is limited by the need to take small timesteps (1 or 2 fs), which is determined by high frequency motions such as chemical bond stretching. One natural way to bridge this timescale gap is to use coarse grained models where the smallest unit of the system represents a group of atoms. In these models, much longer timesteps are allowed since the high frequency motions are not explicitly simulated. Coarse grained simulations work well for a variety of problems; however, these models sacrifice accuracy for speed, making them less than ideal for investigating the detailed mechanisms of conformational changes.

An alternative approach to overcome the timescale gaps is to build discrete-time Markov State Models (MSMs). These models may be built from many short (nanosecond timescale) simulations and then propagated to give long timescale dynamics, such as processes occurring on microsecond timescales or even longer. MSMs partition phase space into a number of distinct states, called metastable states, such that intra-state transitions are fast but inter-state transitions are slow. Such separation of timescales ensures that the model is Markovian, in that the probability of being in a given state at time \( t + \Delta t \) depends only on the state at time \( t \). In an MSM, the time evolution of a vector representing the population of each state may be calculated by repeatedly left-multiplying by the transition probability matrix.

\[
P(n\Delta t) = [T(\Delta t)]^n P(0)
\]

where \( P(n\Delta t) \) is a vector of state populations at time \( n\Delta t \) and \( T \) is the column-stochastic transition probability matrix. Any model is Markovian for a sufficiently long lag time (\( \tau = \Delta t \)), because the system is able to relax to an equilibrium distribution from any arbitrary starting distribution after one lag time. The key point is to build a model with a lag time that is shorter than the timescale of the process of interest with a reasonable number of states. This requires a very good state decomposition, which is difficult. A few different approaches have been developed to address this issue. (cite)

MSMs are also multi-resolution in nature. In order to achieve a Markovian model at a certain lag time, the states must be defined such that large internal free energy barriers are avoided and conformations within the same metastable state can interconvert within one lag time. Thus, the number of states needed in an MSM depends on the desired lag time. The smaller the lag time is, the more states the MSM needs to ensure that dynamics within each state are memoryless after one lag time. Requiring a short lag time would result in high resolution MSM having many metastable states, capturing numerous free energy minima separated by small barriers. Requiring a longer lag time results in a low resolution MSMs with only a few states, each of which contains multiple local free energy minima. We introduce a new algorithm, Super-density-level Hierarchical Clustering (SHC), to construct MSMs at different resolutions for conformational dynamics. To our knowledge, SHC is the first algorithm focusing on generating MSMs at multiple resolutions.

The key insight of the SHC algorithm is to cluster conformations hierarchically using super density level sets in a bottom-up fashion starting with the densest regions of phase space, which correspond to the bottoms of free energy minima. This algorithm can generate multi-resolution models by tuning the super density level sets, and each level of resolution constitutes a discrete-state MSM with a particular partitioning of the phase space. At low resolution, it generates a coarse state decomposition with a small number of metastable states while at high resolution it generates a finer state decomposition with more metastable states. This leaves one the flexibility to select an MSM at the best
resolution to study their biological problem. In addition, we show that SHC also helps improve the meatastability of MSMs by more appropriately assigning fuzzy states on transition regions to metastable states.

The procedure to build MSMs using SHC is as follows. (1) Partition the conformations into a large number of states, called microstates, according to their structural similarity. An approximate K-centers clustering algorithm is used here as it gives states with approximately uniform size, resulting in a correlation between the population of each state and its density. (2) Split the microstates into $n$ density levels ordered from high to low density ($L_1, ..., L_n)$ such that each level contains approximately the same number of conformations. Then construct super density level sets $S_i$, where $S_i = L_1 \cup L_2 \cup ... \cup L_{i-1} \cup L_i$. Thus each super density level contains all previous levels $S_1 \subseteq S_2 \subseteq ... \subseteq S_i$. (3) Within each super density level ($S_i$), perform spectral clustering to group kinetically related microstates. Metastable regions are better separated at high density super levels, since most of the fuzzy microstates in the transition region are excluded at these levels. Now, build a graph representing the connectivity of the states across super density levels. Then generate gradient flows along the edges of the graph from low to high density levels. Each attraction node where the gradient flow ends is assigned to a new metastable state. (4) Assign every microstate not belonging to an attraction node to the metastable state it has the largest transition probability to. Thus we have a complete state decomposition for an MSM. Furthermore, this procedure may be repeated with different super density level sets to construct MSMs at different resolutions. The larger the number of super density levels, the finer the resolution and the larger the number of metastable states in the final MSM.

In order to test SHC, we apply it to a small RNA hairpin with microsecond time scale dynamics: an eight nucleotide RNA GCAA tetraloop with the sequence 5'-GCGGCAGC-3'. It has 4 bases in the loop and two stem base pairs as shown in Figure 1. RNA hairpins are a ubiquitous secondary structure motif often involved in tertiary contacts. Much experimental work has been done on these systems as a step towards understanding larger RNA molecules but knowledge of their folding is still incomplete. Despite their small size, even eight nucleotide hairpins fold on a microsecond timescale, about two orders of magnitude longer than typical atomic simulations. However, using SHC, we are able to construct multi-resolution MSMs from many short 45 ns atomic simulations. These models are able to predict microsecond dynamics. We compare MSMs at different resolutions and also validate them by confirming their ability to reproduce the original simulation trajectories. Furthermore, we extract the kinetics between the most populated metastable states from our MSMs. The results suggest that the folded state is a hub connected to many non-native metastable states that are mostly uncoupled from one another. No metastable intermediate states are identified, while there are a few misfolded states such as states with shifted base pairing or an unfolded loop. This indicates that folding of an eight nucleotide RNA hairpin with only two stem base pairs might be different from RNA hairpins with longer stems where stable thermodynamic intermediate states were seen in previous simulations.

![Figure 1](image_url)

Figure 1. (A) Structure of the 8 nucleotide RNA GCAA tetraloop, generated by truncating two terminal base pairs from the NMR structure of a 12 nucleotide tetraloop (PDB ID 1zih). (B) The cartoon representation of the same structure using sticks to represent orientation of the bases. The same representation will be used in Figure 6 again to illustrate representative structures from different metastable states.
2. Methods

In the previous section, we have introduced a few key steps in building MSMs using SHC. Here, we will use the RNA GCAA tetraloop as an example to go through the detailed procedure. The dataset we examine here contains 9,963 45ns explicit solvent molecular dynamics simulations with an aggregated simulation time of 448 microseconds. Conformations are saved every 0.2 ns, therefore the total number of conformations is about 2.3 million. These simulations are initiated from different metastable regions of phase space identified by initial Simulated Tempering simulations according to the Adaptive Seeding Method (ASM) procedure. More simulation details are available in the Appendix A.

2.1. Partitioning conformations into microstates

Modern computer simulations can easily generate massive data sets with millions of conformations, so that it is very difficult to directly work on each individual conformation. Thus the first step is to group conformations into a large number (a few thousand or tens of thousands) of small clusters or microstates, according to their structural similarly such as Root Mean Square Deviation (RMSD). Each microstate has to be small enough such that conformations within each of them can interconvert quickly. An approximate K-centers clustering algorithm was used here to generate the microstates by minimizing the maximum cluster size, where the cluster radius is defined as the maximum heavy atom RMSD distance between the cluster center and any other conformation within the cluster. The detailed implementation of the algorithm is discussed elsewhere (cite), and the code for the approximate K-centers clustering is available through MSMBuilder package (cite). This algorithm is fast with the computational complexity O(kN), where k is the number of clusters and N is the number of conformations to be clustered. Moreover, it gives states with approximately equal size. This makes the population of each microstate approximately correlates to the density of conformations, which allows us to define density levels in the subsequent steps.

We have clustered 2.3 million conformations into 10,000 microstates, and the same microstate decomposition is used to build all MSMs. The cluster radius distribution has a sharp peak around 4 Å, confirming that clusters have approximately equal size (data not shown). Thus, the population of each microstate is a reasonable indicator of the density of the conformation space. However, we note that even small difference in the size of microstates may imply relatively large variations in their volumes in the high dimensional conformation space. We empirically found that assuming clusters with approximately equal size is useful. In the future, we can improve the density estimation by working on the low dimensional sub-manifolds where density estimation is consistent and accurate. These low dimensional sub-manifolds can be constructed by nonlinear dimensionality reduction techniques (cite).

2.2. Super density level set formation

In this step, we first split the microstates into n density levels \( L = \{L_1, \ldots, L_n\} \). As discussed above, density of microstates \( \{d_1, \ldots, d_n\} \) can be estimated by their populations. We order microstates according to values of \( d_i \) and classify the microstates into n consecutive levels. Each level contains about the same number of conformations. Density levels are ordered from high to low density regions, and labeled as 1 to n. For example, from our RNA dataset, we have generated a density level set with three levels \( L = \{L_1, L_2, L_3\} \). \( L_1, L_2, \) and \( L_3 \) contain 146, 615, and 1810 microstates respectively, and approximately equal number of conformations each (about 25% of total conformation at each level). Thus, level \( L_1 \) with least number of microstates should locate at the highest density region. From the density level set, we can easily construct the super density level set \( S = \{S_1, \ldots, S_n\} \) by defining \( S_i = L_1 \cup L_2 \ldots \cup L_{i-1} \cup L_i \). Each super density level contains all previous levels \( S_i \subseteq S_{i+1} \subseteq \ldots \subseteq S_n \). In our example, three super density levels \( S_1, S_2, \) and \( S_3 \) contain 25%, 50% and 75% of total conformations respectively.

2.3. Spectral Clustering at super density levels

Spectral clustering algorithm (cite) is performed at each super density level (\( S_i \)) to lump kinetically related microstates into larger metastable states. Metastable regions are better separated at high density super levels, since
most of the fuzzy microstates in the transition region are excluded at these levels. For example, in the RNA dataset, multiple disconnected blocks are found in the transition probability matrix at the level $S_1$, indicating good separations of metastable regions. When we move up to levels containing lower density microstates, less and less disconnected blocks are found in the transition probability matrix, and eventually every microstate will be connected. An example is shown in Figure 2. The first level $S_1$ contains 35 metastable states, it decrease to 25 for $S_2$, and only 6 states are identified at $S_3$. In order to identify nearly disconnected blocks in the transition matrix, we choose eigenvalues very close to 1 in spectral clustering. In particular, a constant spectra gap with $\Delta \lambda = 0.0001$ is used for this example.

Figure 2. A graph describing the connectivity of metastable states generated by SHC. Each node in the graph denotes to one metastable state. Each row corresponds to one super density level: states belonging to $S_1$(in red), $S_2$(in blue), and $S_3$(in green) contain 25%, 50%, and 75% of all the conformations respectively. Two nodes are connected if they share microstates, and the arrows represent gradient flows from low density to high density region, i.e. from $S_3$ to $S_1$. Arrows representing self transitions are plotted at attraction nodes where the flow ends. The radius of nodes is scaled linearly by their populations within each super level.

At last, we build a graph representing the connectivity of the metastable states across super density levels. Figure 2 is an example of such graph with three levels. Each node in the graph represents one metastable state. As discussed above, number of nodes in each level decrease from $S_1$ to $S_3$. In $S_1$, there is a large node (node 1) containing 64% of all conformations in that level. Similar nodes can also be found in other levels such as node 2 (83%) in $S_2$ and node 3 (99%) in $S_3$. These results suggest that there exists a large metastable state, which actually corresponds to the folded region. More discussions can be found in the Results section. In the next step, gradient flows are generated along the edge of the graph from low to high in density level. Each attraction node where the gradient flow ends is assigned to a new metastable state. For example, node 1 is an attraction node, while node 2 and 3 are not. As shown in Figure 2, there are 46 attraction nodes in this model (35 in $S_1$ and 11 in $S_2$). Thus the model contains 46 metastable states.

2.4. Assigning microstates not in attraction nodes

In the previous step, all the attraction nodes have been assigned to metastable states. Here, we will assign all other microstates not belonging to any attraction node to a particular metastable state. This is achieved by computing the transition probabilities from this microstate to all metastable states, and assigning it to the metastable state it has the largest transition probability to. If a particular microstate is not connected to any of the metastable states, we raise the order of the transition probability matrix iteratively until we see transition between this microstate and some metastable state.

Following the above steps, we have a complete state decomposition for an MSM. In the example shown in Figure 2, a 46-state MSM is generated. In order to construct MSMS at different resolutions, the same procedure may be repeated by varying number of super density levels.
3. Results and Discussions

3.1. Constructing MSMs at different resolutions

Using SHC, we have constructed four different MSMs by varying the number of super density levels (L) all with a lag time of 0.2 ns. Specifically, we used 3, 6, 9, and 15 super density levels, yielding MSMs referred to as L3 MSM, L5 MSM, L9 MSM and L15 MSM respectively. In addition, we also built a model (L1 MSM) with L = 1 as a control. Some properties of these models are listed in Table 1.

<table>
<thead>
<tr>
<th>L</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>6</td>
<td>46</td>
<td>57</td>
<td>63</td>
<td>68</td>
</tr>
<tr>
<td>Q</td>
<td>5.95</td>
<td>44.3</td>
<td>54.2</td>
<td>59.3</td>
<td>63.4</td>
</tr>
<tr>
<td>&lt;T&gt;</td>
<td>99.1%</td>
<td>96.3%</td>
<td>95.1%</td>
<td>94.1%</td>
<td>93.2%</td>
</tr>
</tbody>
</table>

The first property in the table is the number of macrostates in each MSM. This number increases with L, and L15 MSM contains more than ten times more states than L1 MSM. With many more states, L15 MSM is a higher resolution model than L1 MSM. Thus SHC is able to generating multi-resolution MSMs by changing number of levels L. Metastability is another important property for a MSM. A good MSM should contain a state decomposition which maximizes the separation of timescales. The self-transition probability, indicating the stability of each macrostate, is a simple and straightforward way to check if there is a good separation of timescales. The metastability (Q) listed in the table is defined as the sum of the self-transition probabilities (T_{ii}) of each macrostate. The average self transition probability: <T_{ii}> = Q/N, where N is the number of microstates. <T_{ii}> decrease with L, indicating higher resolution models have smaller average self transition probabilities. This is consistent with the fact that higher resolution models will capture smaller free energy minima, which are separated by smaller free energy barriers and therefore less metastable.

Another interesting property, which is not listed in the table, is the population of each macrostate. For the control model L1 MSM, the populations of the six states ordered from high to low are: 98.0%, 1.6%, 0.2%, 0.05%, 0.05%, and 0.05%. Only two states have populations greater than 1%, and the rest have negligible populations. A closer look at the data shows that these four states each only contains one microstate, and they are almost disconnected from the rest of phase space. Thus these four states might not be significant metastable regions, but just noise due to insufficient sampling. This is one issue with the straightforward spectral clustering algorithms such as PCCA and PCCA+, which tend to first separate the most disconnected blocks from the transition probability matrix. This makes it difficult to choose a proper number of metastable states in order to identify all the significant metastable regions. SHC is able to overcome this issue by clustering from the highest density super level, which guarantees that the most populated metastable regions are identified first. L3 MSM, L5 MSM, L9 MSM, and L15 MSM contain 8, 15, 12, and 10 states with populations larger than 1% respectively.

3.2. Validating MSMs

In this section, we will validate the MSMs discussed above in two ways: implied timescales and Chapman-Kolmogorov equation.

**Implied timescales.** Examining the behaviors of the implied timescales is one way to check if the model is Markovian as first suggested by Swope et al. Implied timescales (τ_k) can be computed from the eigenvalues of the transition matrix T as shown below:

\[
\tau_k = -\frac{\tau}{\ln \mu_k(\tau)}
\]
where $\mu_k$ is an eigenvalue of the transition matrix with the lag time $\tau$. Each implied timescale describes an aggregate transition between subsets of macrostates. If the model is Markovian and Equation (1) holds, the exponentiation of $T$ should be identical to an MSM constructed with a longer lag time, and the implied timescales will be independent of the lag time. This requires that lag times are sufficiently long. The shortest lag time for this condition to hold is defined as the Markovian time, which is correlated with the longest internal equilibrium time of any state. Figure 3 displays implied timescales plots as a function of the lag time for L3 MSM. As shown in Figure 3 (a), the implied timescales level off around a lag time of 20ns. This implies that the model is Markovian with long enough lag times. However, big fluctuations are observed for the three slowest timescales. A further investigation shows that these slow timescales are due to low-population states which are nearly disconnected from the other states. If we exclude three states (with populations 0.1%, 0.09%, and 0.04%), containing very few non-self transition counts, from our analysis, these slowest timescales disappear (see Figure 3 (b)). The implied timescale plots for other resolution MSMs also level off as shown in Figure 4. These results suggest that MSMs generated from SHC are Markovian with sufficiently long lag times. Higher resolution MSMs with a finer discretization of phase space should have shorter Markovian times, since the intra-state equilibrium times are shorter. Looking at Figure 4, the implied timescales of L15 MSM seem to level off slightly faster than those of L6 MSM. However, it is hard to tell by eye whether there is any large difference in the Markovian times for these models. Thus, the implied timescales check has some drawbacks. It is difficult to determine by eye if and where the implied timescales level off. In addition, small uncertainties in the eigenvalues can induce large uncertainties in the implied time scales.

![Figure 3](image1.png)

**Figure 3.** Top twenty implied timescales as a function of lag time for the L3 MSM (L3 denotes the super density level set containing 3 levels). The plots are generated by using (a) the transition probability matrix with all 46 states. (b) the transition probability matrix with only 43 states with three nearly uncoupled states excluded (These three states have very few transition counts to other states).

![Figure 4](image2.png)

**Figure 4.** Top twenty implied timescales as function of lag time for the (a) L6 MSM, (b) L9 MSM, and (c) L15 MSM. L6, L9, L15 represents that 6, 9 and 15 super density levels are used to generate these MSMs respectively. The insert in (b) is the same as the main figure except that y axis goes up to 7 microseconds in order to show one very long implied timescale.

**Chapman-Kolmogorov Check.** An alternative way to validate MSMs is to directly check if Equation (1), a form of the Chapman-Kolmogorov equation, holds. Figure 4 shows the time evolution of the populations of the top eight most populated states in L3 MSM. Populations extracted from the raw data are compared with those generated by the MSM starting from the same initial populations (see Equation (1)). As shown in Figure 4, these populations agree well within statistical error. Similar agreement was found for the other MSMs as well (data not shown). These
results suggest that MSMs generated by SHC are consistent with the original dataset from which they were constructed.

Figure 5. Comparison between the time evolution of the population of eight top populated states (with the population bigger than 1%) in the L3 MSM (red) and the raw data (black). The error bars in black curves are standard deviations computed from one hundred bootstrapping runs by randomly select 8,000 from 9,963 trajectories with replacement. A 20ns lag time is used to build transition probability matrices using L3 MSM state decomposition.

3.3. RNA hairpin folding mechanisms

Despite the small size of RNA hairpins, there is some debate whether they fold in a two-state or multi-state manner. Thermodynamic measurements such as temperature melting (cite) support the two-state model, while kinetic experiments such as temperature jump suggest a multi-state model (cite). Using laser temperature jump technique, the Gruebele group(cite) observed two unfolding relaxation phases of the eight nucleotide gcUUCGgc hairpin at low temperatures: a fast phase of 1-2 microseconds, and a slow phase of 5-10 microseconds.(cite) They also developed a lattice model with four metastable states to accurately reproduce the experimental data. However, it is difficult to extract information at atomic resolution from this simple model.

MSMs are a useful tool to extract kinetics from atomistic simulations. From L3 MSM, we have computed the kinetic property, Mean First Passage Time (MFPT), between eight most populated metastable states. MFPT is defined as the average time taken to get from the initial state to the final state. It can be computed from the transition probability matrix (see Appendix B for details). The results of MFPTs are displayed in Figure 6 along with the representative structures from each state. State 1 with the largest population (77.1%) is the folded state, indicating the free energy surface is biased to the native state at 300K. Multiple non-native states, each directly connected to the folded state, are also identified: e.g. state 3 and 4 with coil structures, state 2 with shifted base pairing, and state 5 with an unfolded loop. MFPTs for folding (i.e. transitions from non-native states to the folded state) are a few hundred nanoseconds, while MFPTs for unfolding are at least an order of magnitude longer (from a few to tens of microseconds). This confirms that folded state is the most stable state at 300 K. MFPTs between non-native states are at least eight microseconds, much longer than those for folding. This suggests that these states are uncoupled from each other. Therefore, no metastable on-pathway intermediates states are indentified in this system. The transition from state 1 (folded) to 8 (shifted base pairing) has longest MFPT (45.7 microseconds) among unfolding transitions, indicating big energy barriers for breaking non-native base pairing/stacking followed by forming native ones. State 5 (unfolded loop) has shortest MFPT (0.16 microsecond) among folding transitions, which suggests a fast kinetics of loop rearrangements.

We have successfully extracted kinetic information between the most populated metastable states from our MSMs. The overall unfolding timescales are in a range of a few to tens of microseconds, which are qualitatively consistent with experimental observations. However, direct comparison between our simulations and laser T-jump experiments are not possible because our simulations at a single temperature is not able to capture effects due to temperature change. No stable thermodynamic intermediate states were found for folding of an 8 nucleotide RNA...
hairpin, which is different from folding of a 12 nucleotide hairpin. These results suggest that the increase of the number stem base pairs may complicate the folding mechanisms.

Figure 6: Mean First Passage Times (MFPT) between eight top populated states in the L3 MSM (generated by a super level set with three levels, see Table 1 for details). All the MFPTs are in the unit of microseconds. States are labeled in red from 1 to 8 according to their populations. Populations of each state are shown in black. Two representative conformations of each state, extracted from centers of the top populated microstate within each metastable state, are drawn using Pymol (cite) with a cartoon representation.

4. Conclusion and Future Plans

Markov State Model (MSM) provides a useful tool to bridge the gap between experimental and computational timescales. MSMs are inherently multi-resolution, however, algorithms focused on constructing MSMs at different resolutions are lacking. Here we have introduced a new algorithm Super-level-set Hierarchical Clustering (SHC), which is capable of constructing MSMs at multiple resolutions for conformational dynamics. The key insight of this algorithm is to perform spectral clustering hierarchically using super level sets starting from the highest density level, which guarantees to identify highly populated metastable regions first. This overcomes the issue with straightforward spectral clustering, where nearly disconnected components of the transition matrix, which are separated first, are often noise due to insufficient sampling. We apply SHC to an 8 nucleotide GCAA RNA tetraloop, and built four MSMs at different resolutions. All these models have been validated by both implied timescale and Chapman-Kolmogorov checks. The overall unfolding timescales predicted from our MSMs are between a few and tens of microseconds, which are qualitatively consistent with those observed by laser temperature jump experiments. Our results suggest that there are no metastable intermediate states, while the folded state is directly connected to multiple unfolded or misfolded states, which interconvert slowly from one another.

In SHC, we use population of microstates from K-centers clustering to approximate the density of the conformation space. We are aware of the fact that the estimation of density at high dimensional space is difficult. Furthermore, K-centers algorithm only generates clusters with approximately equal radius, and the small variance in the cluster size may induce large difference in the volume. In the future, we plan to refine the density estimators by computing the kernel density around microstate centers or the average of the kernel density for a few randomly
selected conformations within the state. Furthermore, nonlinear dimensionality reduction techniques may also be used to reduce the dimensionality, and allow us to work on the low dimensional sub-manifolds where density estimation is easier and more accurate. We have demonstrated that SHC is able to generate a large number of MSMs at different resolutions. However, we haven’t discussed how to select the best model among them. A Bayesian approach to compare different MSMs by Bacallado et al. may be used for model selection in the future. Finally, in addition to the metastable states, the SHC models can also provide information about intermediate and transition states by investigating non-attraction nodes located in low density super levels. in a spirit similar to the topological approach recently exploited by Yao et al. In addition to being biologically relevant themselves, identification of these states could allow us to perform adaptive sampling by starting more simulations from transition states in order to rapidly sample transition events between metastable states.

Acknowledgments

XH is supported by NIH Roadmap U54 GM072970 and a startup fund from Hong Kong University of Science & Technology. XH would also like to acknowledge the support of Prof. Michael Levitt. YY, JS, LG, and GC are supported by DARPA HR0011-05-1-0007. YY and GC are also supported by NSF DMS-0354543, and LG by NSF FRG-0354543 and NIH GM-072970. GRB is supported by the NSF Graduate Research Fellowship Program. This work is also funded by NIH R01-GM062868 and NIH P01 GM066275. Computer resources are provided by NSF award CNS-0619926 and Folding@Home volunteers.

Appendix A: Simulation Details

This section describes in detail how we generate Molecular Dynamics simulation trajectories examined in this study. We generate them following the procedure described by Adaptive Seeding Method (ASM). At first, two sets of 1120 27ns Simulated Tempering (ST) simulations were run: one started from a folded state and the other from a random coil. An independent MSM with 10 states was then built using MSMBuilder (cite) for each dataset in order to identify the dominant metastable states. (2). One hundred random conformations were then chosen from each state and used as starting points (2,000 points in total). Five 45ns constant temperature 300K MD simulations were then launched from each point. This results in a dataset with 9,963 trajectories (some simulations are not finished). All the simulations are performed using Folding@Home (cite). We used nucleic acid parameters from the AMBER99 force field.[9, 10] The RNA molecule was solvated in a water box with 2,543 TIP3P[11] waters and 7 Na+ ions. The simulation system was minimized using a steepest descent algorithm, followed by a 100ps MD simulation applying a position restraint potential to the heavy atoms. All NVT simulations were coupled to a Nose-Hoover thermostat with a coupling constant of 0.02ps⁻¹.[12] A cutoff of 10 Å was used for both vdw and short range electrostatic interactions. Long-range electrostatic interactions were treated with the Particle-Mesh Ewald (PME) method.[13] Nonbonded pair-lists were updated every 10 steps with an integration step size of 2 fs in all simulations. All bonds were constrained using the LINCS algorithm.[14]

Appendix B: Mean First Passage Time (MFPT)

The mean first passage time (MFPT) from initial state $i$ to final state $f$ in a MSM can be defined as the average time taken to get from state $i$ to state $f$. The MFPT ($X_{if}$) given that the transition to state $j$ was made is the time it took to get from state $i$ to $j$ plus the mean first passage time from state $j$ to $f$. Thus the MFPT ($X_{if}$) can be defined as (cite),

$$X_{if} = \sum_{j} P_{ij} (t_{ij} + X_{jf})$$

(3)

where $t_{ij}$ equals the lag time of the transition matrix $T$. And the boundary condition is:

$$X_{ff} = 0$$

(4)
This set of linear equations in Equation (3) and (4) can be solved to obtain the MFPT $X_f$.

References