Supplementary Information for: Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9

Christian R. Schwantes[†] and Vijay S. Pande[*,†,‡,¶]

Department of Chemistry, Stanford University, Stanford, CA, USA, Biophysics Program, Stanford University, Stanford, CA, USA, and Department of Computer Science, Stanford University, Stanford, CA, USA

E-mail: pande@stanford.edu

---

[*] To whom correspondence should be addressed

[†] Department of Chemistry, Stanford University, Stanford, CA, USA

[‡] Biophysics Program, Stanford University, Stanford, CA, USA

[¶] Department of Computer Science, Stanford University, Stanford, CA, USA

Detailed Proof of the Solutions to the tICA Problem

Recall that the first solution to the tICA problem is found by solving the maximization problem shown below.

$$\max_{|\alpha_0\rangle} f\big(|\alpha_0\rangle\big) = \max_{|\alpha_0\rangle} \langle\alpha_0|\mathbf{C}^{(\Delta t)}|\alpha_0\rangle \tag{1}$$

subject to:

$$\langle\alpha_0|\Sigma|\alpha_0\rangle = 1$$

We solve this using the same strategy used in a classical proof of the PCA problem, which is to use Lagrange multipliers.[1] First we set up the Lagrangian:

$$\Lambda = \langle\alpha_0|\mathbf{C}^{(\Delta t)}|\alpha_0\rangle - \lambda_0\Big(\langle\alpha_0|\Sigma|\alpha_0\rangle - 1\Big) \tag{2}$$

Next we differentiate with respect to $|\alpha_0\rangle$ and λ_0 yielding a system of two equations:

$$\frac{\partial\Lambda}{\partial|\alpha_0\rangle} = \mathbf{C}^{(\Delta t)}|\alpha_0\rangle - \lambda_0\Sigma|\alpha_0\rangle = 0 \tag{3}$$

$$\frac{\partial\Lambda}{\partial\lambda_0} = \langle\alpha_0|\Sigma|\alpha_0\rangle - 1 = 0 \tag{4}$$

The solution to this system is an eigenvector of the generalized eigenvalue problem in Eq. (5) such that it has unit variance.

$$\mathbf{C}^{(\Delta t)}|\alpha_0\rangle = \lambda_0\Sigma|\alpha_0\rangle \tag{5}$$

Now that we have the slowest component, we wish to find the next slowest component, $|\alpha_1\rangle$, such that it is uncorrelated with the first. This can be written as another maximization problem but with an additional constraint:

$$\max_{|\alpha_1\rangle} f\big(|\alpha_1\rangle\big) = \max_{|\alpha_1\rangle} \langle\alpha_1|\mathbf{C}^{(\Delta t)}|\alpha_1\rangle \tag{6}$$

subject to:

$$\langle \alpha_1 | \Sigma | \alpha_1 \rangle = 1$$

$$\langle \alpha_1 | \Sigma | \alpha_0 \rangle = 0$$

We can again solve this with the method of Lagrange multipliers. First we set up the Lagrangian:

$$\Lambda = \langle \alpha_1 | \mathbf{C}^{(\Delta t)} | \alpha_1 \rangle - \lambda_1 \Big(\langle \alpha_1 | \Sigma | \alpha_1 \rangle - 1 \Big) - \phi_1 \Big(\langle \alpha_0 | \Sigma | \alpha_1 \rangle \Big) \tag{7}$$

Taking derivatives with respect to $|\alpha_1\rangle$, λ_1, and ϕ_1 produces a system of three equations:

$$\frac{\partial \Lambda}{\partial |\alpha_1\rangle} = \mathbf{C}^{(\Delta t)} | \alpha_1 \rangle - \lambda_1 \Sigma | \alpha_1 \rangle - \phi_1 \Sigma | \alpha_0 \rangle = 0 \tag{8}$$

$$\frac{\partial \Lambda}{\partial \lambda_1} = \langle \alpha_1 | \Sigma | \alpha_1 \rangle - 1 = 0 \tag{9}$$

$$\frac{\partial \Lambda}{\partial \phi_1} = \langle \alpha_0 | \Sigma | \alpha_1 \rangle = 0 \tag{10}$$

Now we multiply Eq. (8) by $\langle \alpha_0 |$ on the left to get:

$$\langle \alpha_0 | \mathbf{C}^{(\Delta t)} | \alpha_1 \rangle - \lambda_1 \langle \alpha_0 | \Sigma | \alpha_1 \rangle - \phi_1 \langle \alpha_0 | \Sigma | \alpha_0 \rangle = 0 \tag{11}$$

Now using Eq. (5), we can solve this equation for ϕ_1. First notice that:

$$\begin{aligned}
\langle \alpha_0 | \mathbf{C}^{(\Delta t)} | \alpha_1 \rangle &= \langle \alpha_1 | \mathbf{C}^{(\Delta t)} | \alpha_0 \rangle \\
&= \langle \alpha_1 | \lambda_0 \Sigma | \alpha_0 \rangle \\
&= \lambda_0 \langle \alpha_1 | \Sigma | \alpha_0 \rangle \\
&= \lambda_0 (0) \\
&= 0
\end{aligned}$$

Plugging this result into Eq. (11) we find that:

$$\phi_1 = 0 \tag{12}$$

And then we can see that Eq. (8) becomes:

$$\mathbf{C}^{(\Delta t)}|\alpha_1\rangle = \lambda_1 \Sigma |\alpha_1\rangle \tag{13}$$

Now, we finally consider the case that we have k solutions $\{|\alpha_i\rangle\}_{i=0}^{k-1}$, and wish to find the next solution $|\alpha_k\rangle$. Again we set up a maximization problem but with more constraints:

$$\max_{|\alpha_k\rangle} f\big(|\alpha_k\rangle\big) = \max_{|\alpha_k\rangle} \langle\alpha_k|\mathbf{C}^{(\Delta t)}|\alpha_k\rangle \tag{14}$$

subject to:

$$\langle\alpha_k|\Sigma|\alpha_k\rangle = 1$$

$$\langle\alpha_k|\Sigma|\alpha_i\rangle = 0 \text{ for } i = 0\ldots k-1$$

We can setup the Lagrangian as before:

$$\Lambda = \langle\alpha_k|\mathbf{C}^{(\Delta t)}|\alpha_k\rangle - \lambda_k\Big(\langle\alpha_k|\Sigma|\alpha_k\rangle - 1\Big) - \sum_{i=0}^{k-1} \phi_i \langle\alpha_i|\Sigma|\alpha_k\rangle \tag{15}$$

Again we can differentiate with respect to α_k, λ_k, and all ϕ_i's. This results in a system of $k+2$ equations:

$$\frac{\partial\Lambda}{\partial|\alpha_k\rangle} = \mathbf{C}^{(\Delta t)}|\alpha_k\rangle - \lambda_1\Sigma|\alpha_k\rangle - \sum_{i=0}^{k-1}\phi_i\Sigma|\alpha_i\rangle = 0 \tag{16}$$

$$\frac{\partial\Lambda}{\partial\lambda_k} = \langle\alpha_k|\Sigma|\alpha_k\rangle - 1 = 0 \tag{17}$$

$$\frac{\partial\Lambda}{\partial\phi_i} = \langle\alpha_i|\Sigma|\alpha_k\rangle = 0 \tag{18}$$

Now for each previous solution, $\{|\alpha_i\rangle\}_{i=0}^{k-1}$, we first multiply Eq. (16) by $|\alpha_i\rangle$ on the left to get:

$$\langle\alpha_i|\mathbf{C}^{(\Delta t)}|\alpha_k\rangle - \lambda_k\langle\alpha_i|\Sigma|\alpha_k\rangle - \sum_{i=0}^{k-1}\phi_i\langle\alpha_i|\Sigma|\alpha_i\rangle = 0 \tag{19}$$

Now it is obvious to see that Eq. (19) can be simplified and become:

$$\lambda_i(0) - \lambda_k(0) - \sum_{i=0}^{k-1}\phi_i\delta_{ij} = \phi_i = 0 \tag{20}$$

Again, we see that Eq. (16) can be re-written and we find the next solution $|\alpha_k\rangle$ satisfies:

$$\mathbf{C}^{(\Delta t)}|\alpha_k\rangle = \lambda_k\Sigma|\alpha_k\rangle \tag{21}$$

Finally, we see that all solutions to the tICA problem satisfy the same generalized eigenvalue problem:

$$\mathbf{C}^{(\Delta t)}|\alpha\rangle = \lambda\Sigma|\alpha\rangle \tag{22}$$

Selection of Tunable Parameters for tICA

The tICA metric requires the use of two tunable parameters. The first is the correlation lag time, Δt, used in calculating the cross-correlation matrix, $\mathbf{C}^{(\Delta t)}$. We have used a value of 200 ns above, but we believe the analysis is fairly robust to the choice of Δt. We found that using 100 or 500 ns also led to similar models (Fig. 1). We chose $\Delta t = 200$ ns because the folding timescale remained as slow as the 100 ns model, but the faster timescales were flatter and slower than the 100 ns model. In contrast to the correlation lag time, we found that the tICA method was less robust to choices of the number of tICs to project onto. We found that using too few tICs produced poor models that could only capture the folding timescale, while using too many tICs produced models that were simply too fast. This behavior can be explained by realizing that adding extra, faster tICs includes more degrees of freedom into the distance calculation, however if these degrees of freedom do not decorrelate slowly, then we are essentially adding noise into our state decomposition, which can

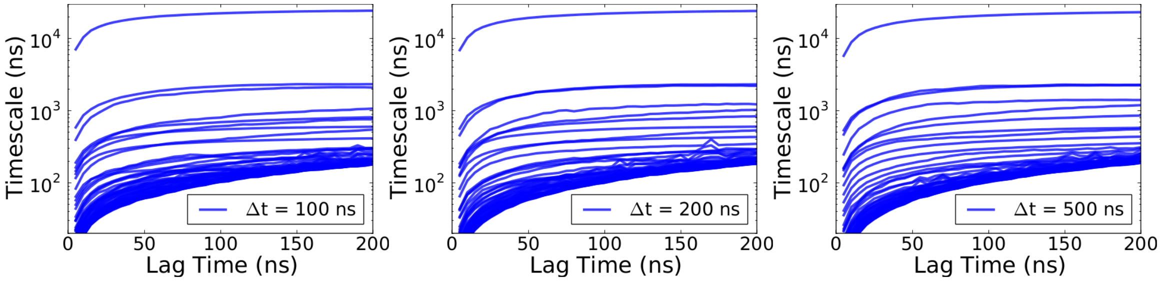

Figure 1: The tICA metric applied to MSM construction is fairly robust to choices of Δt. Models built using 100, 200, and 500 ns for Δt showed very similar timescales.

produce artificially fast processes. For this system, as well as several other, unpublished protein folding systems we found that using between five and seven degrees of freedom produced similar MSMs.

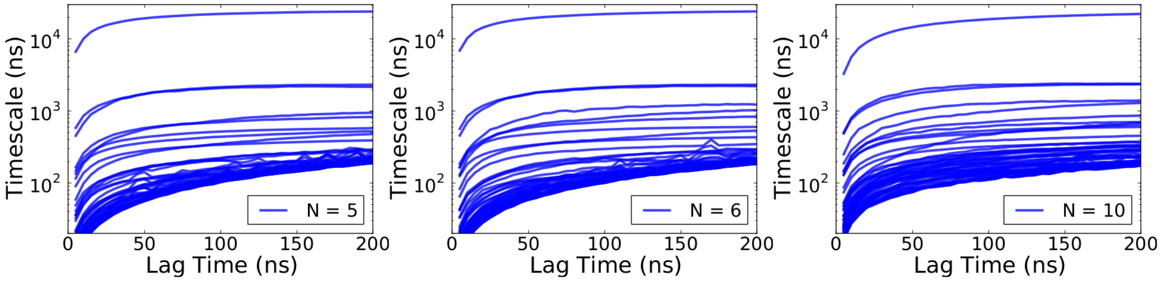

Figure 2: Models built with five or six tICs had very similar timescales, whereas including more tICs produced an MSM that was too fast. Using even fewer tICs (not depicted) resulted in a similar plot but with one faster μs timescale removed. Notice that the $N = 10$ model has slower timescales in the sub-microsecond regime, suggesting that picking a larger N makes your model better able to predict the faster motions, but at the expense of the slower eigenvectors. Although, it is important to note, that the timescales are just subtly different, suggesting that the method is at least somewhat robust to choices of N.

Folding Timescales in the MSM Comparison

The folding timescale in the tICA MSM is significantly slower than those produced by building MSMs with PCA or the contact map approach. As has been shown by Djurdjevac et al.,[2] slower timescales indicate a reduced discretization error. This is even more apparent when the timescales are plotted in a linear scale.

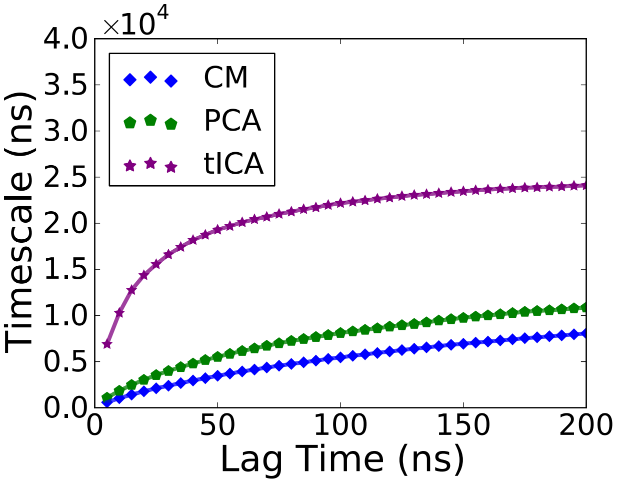

Figure 3: The folding timescale of the tICA MSM is significantly slower than MSMs built using the PCA and contact map (CM) approaches with the same number of states. The folding timescales increase linearly, which is a generally observed phenomenon in most MSMs built on “real” data.

References

(1) Jolliffe, I. T. *Principal Component Analysis*; Springer, 2002; pp 1–9.

(2) Djurdjevac, N.; Sarich, M.; Schütte, C. *Multiscale Model. Simul.* **2012**, *10*, 61–81.