Supporting information for: A Surface-Accelerated String Method for Locating Minimum Free Energy Paths

Timothy J. Giese, Şölen Ekesan, Erika McCarthy, Yujun Tao, and Darrin M. York*

Laboratory for Biomolecular Simulation Research, Institute for Quantitative Biomedicine and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA

E-mail: Darrin.York@rutgers.edu

1 Smoothing of the parametric curve control points

A parametric spline constructed from a series of N_{img}, N_{dim}-dimensional control points will often contain unwanted oscillations if the control points are tainted with numerical noise. One approach for smoothing a sequence of control points is to use a "simple moving average", which is a type of low pass filter. The filtering performs a centered uniform average of the data sequence using an odd-valued integer width, $w = 2i + 1$. The disadvantage of a simple moving average is that the smoothed data will "cut corners" in areas where the values rapidly change. To reduce the amount of corner cutting, we use an iterative procedure that applies a series of corrections. Let $q_{\text{c},nd}^{(k)}$ denote the control point of image n in dimension d at string iteration k. The sequence of control points may contain noise which we seek to reduce. For this purpose, let $q_{\text{c},nd}^{(k,i)}$ denote an estimate of a "smoothed control point" at iteration i of the smoothing procedure. To initiate the

*To whom correspondence should be addressed iterative procedure, set $q_{\mathrm{c},nd}^{(k,0)} = 0$ for all $n \in [1, N_{\mathrm{img}}]$ and $d \in [1, N_{\mathrm{dim}}]$. The estimate of the smoothed control points at iteration $i > 0$ is then given by eq. 1.

$$q_{\mathrm{c},nd}^{(k,i)} = q_{\mathrm{c},nd}^{(k,i-1)} + \frac{s_i}{2i+1} \sum_{m=n-i}^{n+i} f_{md}(\boldsymbol{q}_{\mathrm{c}}^{(k)}, \boldsymbol{q}_{\mathrm{c}}^{(k,i-1)}) \tag{1}$$

$$f_{md}(\boldsymbol{q}_{\mathrm{c}}^{(k)}, \boldsymbol{q}_{\mathrm{c}}^{(k,i-1)}) = \begin{cases} 2q_{\mathrm{c},1,d}^{(k)} - q_{\mathrm{c},|m|+2,d}^{(k,i-1)}, & \text{if } m < 1 \\ 2q_{\mathrm{c},N_{\mathrm{img}},d}^{(k)} - q_{\mathrm{c},2N_{\mathrm{img}}-m,d}^{(k,i-1)}, & \text{if } m > N_{\mathrm{img}} \\ q_{\mathrm{c},md}^{(k)} - q_{\mathrm{c},md}^{(k,i-1)}, & \text{otherwise} \end{cases} \tag{2}$$

$$s_i = \begin{cases} 1, & \text{if } i < 3 \\ 1/2^{i-3}, & \text{otherwise} \end{cases} \tag{3}$$

When the averaging window is within the range $m \in [1, N_{\mathrm{img}}]$, f_{md} returns the difference $q_{\mathrm{c},md}^{(k)} - q_{\mathrm{c},md}^{(k,i-1)}$. Special care is taken when the averaging window extends outside the range of available data. As shown in eq. 2, f_{md} extends the data by treating the endpoints as inversion centers. For example, the point at $N_{\mathrm{img}} + 1$ is a reflection of the point $N_{\mathrm{img}} - 1$ such that: $q_{\mathrm{c},N_{\mathrm{img}}+1,d} - q_{\mathrm{c},N_{\mathrm{img}},d} = -(q_{\mathrm{c},N_{\mathrm{img}}-1,d} - q_{\mathrm{c},N_{\mathrm{img}},d})$.

The net effect of the symmetry operation is to increase the weight of the nearest endpoint while removing the contribution of other points in the window. Furthermore, the symmetry operations enforce the conditions: $q_{\mathrm{c},1,d}^{(k)} = q_{\mathrm{c},1,d}^{(k,i)}$ and $q_{\mathrm{c},N_{\mathrm{img}},d}^{(k)} = q_{\mathrm{c},N_{\mathrm{img}},d}^{(k,i)}$ for all $i > 0$. The s_i quantity dampens the magnitude of the correction to quickly force convergence as the iterations increase. Our experience is that corrections beyond widths of 11 ($i_{\mathrm{max}} = 5$) have a negligble impact on the resulting curve.

2 Cardinal B-Spline Representation of the Free Energy Surface

Consider a uniform discretization of the N_{dim}-dimensional space of reaction coordinates. Given a target bin width in each dimension, w_d, the observed samples are fully contained within a hyperrectangle whose opposing corners are the vertices $\mathbf{q}_{\text{min}} - \mathbf{w}/2$ and $\mathbf{q}_{\text{max}} + \mathbf{w}/2$, where $\mathbf{q}_{\text{min}}$ and $\mathbf{q}_{\text{max}}$ are the minimum and maximum values of the bin centers within the hyperrectangle. Because the histogram uniformly divides each dimension, one can express $\mathbf{q}_{\text{max}}$ in terms of the bin width and the number of bins in each dimension $q_{\text{max},d} = (N_d - 1)w_d + q_{\text{min},d}$.

A N_{dim}-dimensional Cardinal B-spline of order n, θ_n (see eq. 5), is a product of 1−dimensional weights, M_n (see eq. 6), used to perform a weighted average of nearby *bin parameters*, $\mathbf{p}$. Given the parameters, the free energy at a point $\mathbf{q}$ is given by eq. 4, where n is the B-spline order and $\mathbf{q}_b$ is the position of bin center b.

$$F(\mathbf{q};\mathbf{p}) = \sum_{b=1}^{N'_{\text{bin}}} \theta_n(\mathbf{q} - \mathbf{q}_b) p_b \tag{4}$$

$$\theta_n(\mathbf{q} - \mathbf{q}_b) = \prod_{d=1}^{N_{\text{dim}}} M_n\left(\frac{q_d - q_{bd}}{w_d} + \frac{n}{2}\right) \tag{5}$$

$$M_n(u) = \frac{1}{(n-1)!}\sum_{k=0}^{n}(-1)^k\binom{n}{k}[\max(u-k,0)]^{n-1} \tag{6}$$

The B-splines have compact support, such that only the $2\lceil n/2\rceil$ bins nearest to the evaluation point in each dimension can contribute a nonzero weight. If we limit the free energy evaluations (eq. 4) to those values of $\mathbf{q}$ that lie within an occupied bin, then we can pad the occupied regions with additional "auxiliary bins" that serve only to provide the parameters required to fully define the B-spline weights in the occupied regions. In other words, the auxiliary bins ensure that the mesh of $(2\lceil n/2\rceil)^{N_{\text{dim}}}$ bins have associated B-spline parameters. The N'_{bin} in eq. 4 is the total number of occupied and auxiliary bins.

If a bin is occupied, then we set the bin's B-spline parameter to the free energy of the bin produced by MBAR analysis; $p_b = F(\mathbf{q}_b)$. To pad the occupied regions, we successively add layers of auxiliary bins until $\lceil n/2 \rceil$ layers have been introduced. To create a single layer of new bins, we visit each existing bin and find all adjacent bins that do not currently exist. The unique set of bins in this search is the new layer. To assign B-spline parameters to the new layer, we visit each bin in the layer, find the list of its existing neighbors, and select the maximum B-spline parameter from this list. The auxiliary bin B-spline parameter is chosen to be the maximum bin parameter from its existing neighbors shifted by an additional 0.5 kcal/mol. This shift helps to ensure that the free energy increases as one approaches the edge of sampled region.

The B-spline evaluation does not exactly reproduce the MBAR-computed bin free energy values; they average the nearby parameters while giving the nearest bin parameter the largest weight. This is the desired effect because one often seeks to smooth the noise in the data when optimizing a minimum free energy path. As the B-spline order increases, so too does the discrepancy between the MBAR bin free energy values and eq. 4. This discrepancy can be reduced by modifying the B-spline parameters in an iterative procedure. Let $p_b^{(i)}$ denote the B-spline parameter of bin b at iteration i of the procedure. The procedure initializes the occupied bin parameters from the MBAR bin free energy values: $p_b^{(0)} = F(\mathbf{q}_b)$. At this point, the auxiliary bins and auxiliary bin parameters are constructed. The auxiliary bin parameters are held fixed during the iterative procedure; only the B-spline parameters of the occupied bins are modified. The updated B-spline parameters of the occupied bins is then given by eq. 7.

$$p_b^{(i)} = p_b^{(i-1)} + [F(\mathbf{q}_b) - F(\mathbf{q}_b; \mathbf{p}^{(i-1)})] \tag{7}$$

If one updates the parameters by iterating eq. 7 many times, the resulting B-spline evaluations will become indistinguishable from the MBAR bin free energy values, but it will consequently reproduce the noise in the data. The free energies shown in the present work use fourth order B-splines ($n = 4$) and a single iterative correction (eq. 7).

3 Comparison of SMCV, MSMCV, and SASM Optimizations of the MTR1 Path with Reduced Sampling

Figure S1 compares the SMCV, MSMCV, and SASM optimizations of the MTR1 minimum free energy path. Whereas Figure 1 in the main document performed the comparison with 32 images sampled for 4 ps/image, Figure S1 uses 32 images sampled for 500 fs/image.

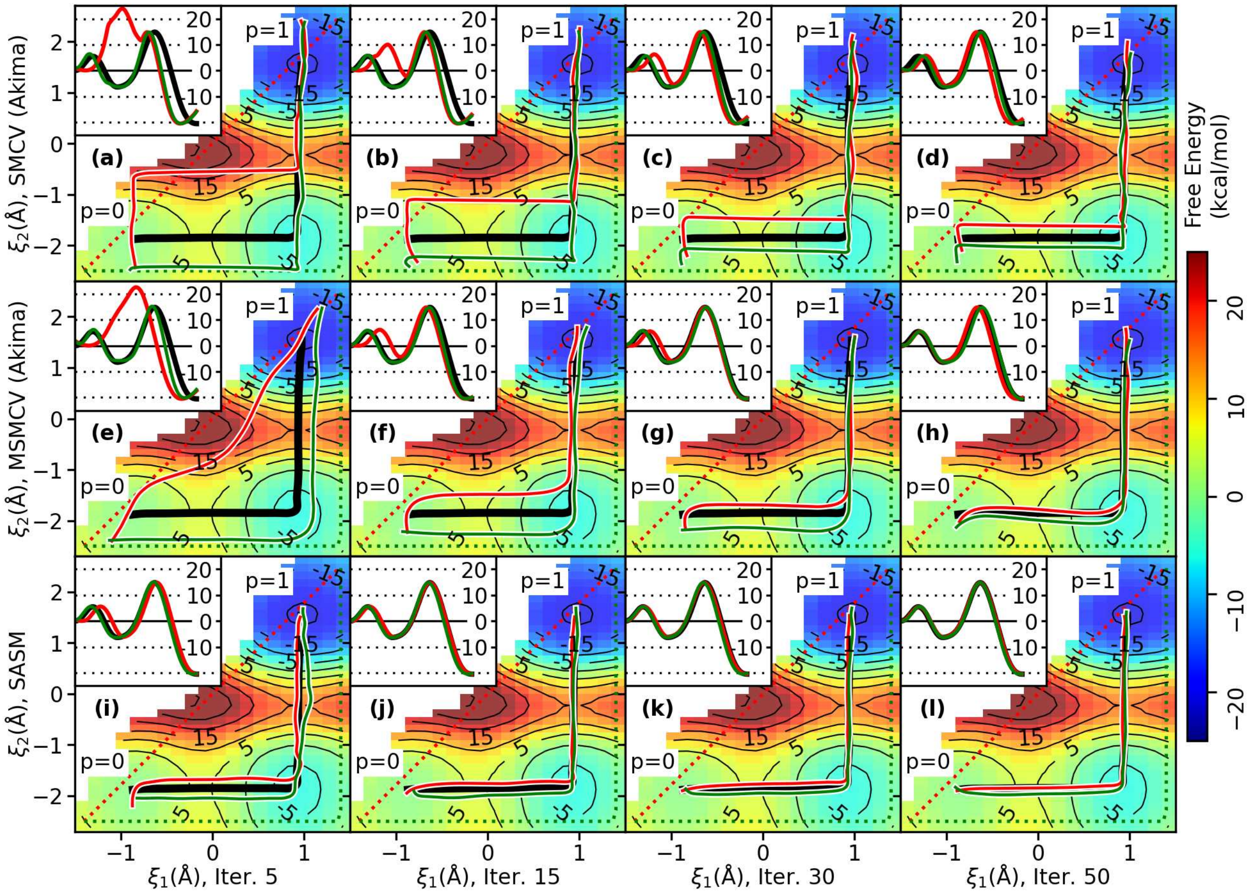

Figure S1: Comparison of the SMCV (a-d), MSMCV (e-h), and SASM (i-l) optimiztions for the minimum free energy path of the MTR1 reaction. The black line is a reference path optimized on the free energy surface calculated from the aggregate sampling from all simulations. The red and green lines are the paths starting from concerted and stepwise initial guesses (the dotted lines), respectively. The insets display the free energy (kcal/mol) along the paths.

The colored areas of the image are the free energy values of the occupied bins, calculated from the aggregate sampling of the 50 string iterations from all 3 methods. In other words, the free energy surface was calculated from 2.4 ns of aggregate sampling.

$$(3\ \text{Methods})\left(\frac{50\ \text{Iter.}}{\text{Method}}\right)\left(\frac{32\ \text{Img.}}{\text{Iter.}}\right)\left(\frac{0.5\ \text{ps}}{\text{Img.}}\right) = 2400\ \text{ps} \tag{8}$$

The free energy values and the optimization progress are nearly indistinguishable from the results shown in Figure 1 of the main document.