A Simplified Confinement Method (SCM) for Calculating Absolute Free Energies and Free Energy and Entropy Differences

V. Ovchinnikov,[1] M. Cecchini[2] and M. Karplus[1,2,a]

[1] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, 02138
[2] Laboratoire de Chimie Biophysique, ISIS, Université de Strasbourg, 67000 Strasbourg, France
[a] marci@tammy.harvard.edu

Supporting Information

S.1 Probability distributions of the potential energy for larger proteins

In the main text, we derived the following approximation for the entropy of a macrostate Ω (c.f. Eq. (14))

$$TS_\Omega = 2(\pi\nu)^2\mathcal{M}\int_0^1 \langle\rho_m^2(\boldsymbol{X},\boldsymbol{X}_0)\rangle_{H^2_\lambda(\lambda)}d\lambda - \frac{3N}{\beta}\left[\log\beta h\nu - \frac{1}{2}\right] + \frac{\beta}{2}\sigma^2_\Omega(E), \tag{1}$$

which becomes exact in the limit of the frequency $\nu \to \infty$ and the probability distribution of the potential energy (*pdf(E)*) approaching a Gaussian. It has been noted before that *pdf(E)* for proteins approaches a Gaussian distribution as the protein size increases,[1] which can be explained heuristically by the fact that the total potential energy is a sum of a large number of interatomic interactions involving only a small number of distinct residue types (*i.e.* there are only twenty essential amino acids). In this section, we show *pdf(E)* computed from all-atom MD simulation of the protein myosin VI (MVI) in complex with MgADP in the "pre-powerstroke" conformation,[2,3] which has 12633 atoms. We also compute the *pdf(E)* from an MD simulation of an 1442-atom fragment of MVI (the converter domain[2] which is stable by itself in the simulation. Both simulations are performed using the FACTS implicit solvent model.[4] The full MVI simulation is performed for 16.2ns, and the converter fragment simulation is performed for 30ns. To quantify the degree of similarity between the computed *pdf(E)* and the Gaussian distribution, we apply the Lilliefors version of the Kolmogorov-Smirnov (LKS) test[5] implemented in the Matlab program.[6] The null hypothesis of the LKS test is that *the data in the series is normally distributed*. For the β-sheet converter distributions (Fig. S1a,b), the null hypothesis can be rejected with p-values of 0.01 and 0.03, respectively, indicating that the LKS test concludes that the corresponding *pdf(E)*s are not Gaussian. For the full MVI simulation, the corresponding p-value is 0.28. As a control test, we sampled several time series from the normal distribution (with the same size as the number of MD trajectory snapshots in the MVI simulation), and applied the LKS test to the test series. The resulting p-values were in the range 0.14 – 0.5 (maximum), indicating that the LKS test cannot reproducibly distinguish between *pdf(E)* from the MVI simulation and a Gaussian. We therefore believe that the entropy approximation in Eq. (1) will be justified for larger proteins (of the order of 10K atoms or more). However, the accuracy of Eq. (1) for arbitrary systems will be system-dependent, and may not be justified in some cases.

S.2 Removing anharmonicity by switching off the force field

In this section we compare the SCM and the confinement method with switching off the force field[7,8] (annihilation). To compare the accuracy of the two methods, we calculated the free energy difference

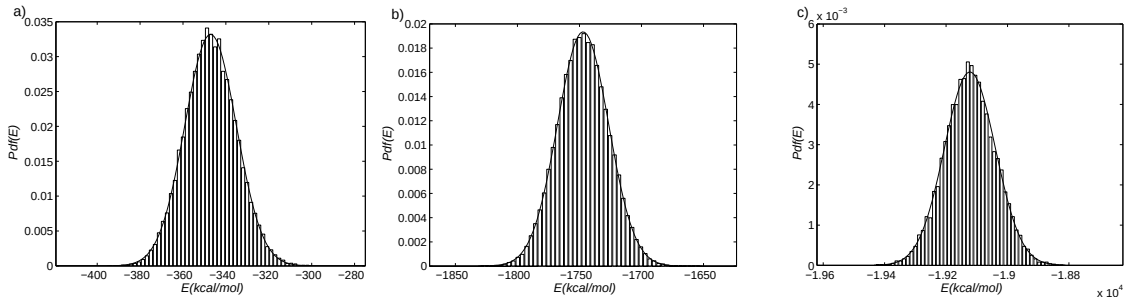

Figure S1: Normalized histograms of the potential energy for (a) β-hairpin (247 atoms), (b) Myosin VI Converter domain (1442 atoms), (c) Myosin VI motor domain (12633 atoms) The solid lines are Gaussian probability densities with the mean and variance computed from the corresponding histograms.

between the **c7eq** and **c7ax** conformers of the Alanine Dipeptide (AD) using the confinement method with annihilation (included in Table II of the main text). In this section, we provide the technical details of the calculations, and a discussion of the relative efficiencies of the two methods.

In analogy with Eq. (1) in the main text, the free energy of macrostate Ω can be written as

$$\begin{aligned} G_\Omega =& \frac{3N}{\beta}\log\beta h\nu - \Delta G_{\Omega\to HO^\nu_{int}} \\ & -\Delta G_{HO^\nu_{int}\to HO^\nu} \end{aligned} \tag{2}$$

in which the term $\Delta G_{\Omega\to HO^\nu_{int}}$ corresponds to the free energy change of transforming the macrostate Ω to the (interacting) HO^ν_{int} state according to Eq. (4) in the main text, and the term $\Delta G_{HO^\nu_{int}\to HO^\nu}$ is the free energy change associated with transforming the HO^ν_{int} state to the corresponding non-interacting HO^ν state. In analogy with Eq. (4) in the main text, this term can be computed by thermodynamic integration using the Hamiltonian

$$H(\boldsymbol{X};\lambda) = \lambda E(\boldsymbol{X}) + \boldsymbol{P}^T M^{-1}\boldsymbol{P}/2 + \sum_i^N (2\pi\nu)^2 m_i \|x^i - x_0^i\|^2/2, \tag{3}$$

which gives

$$\Delta G_{HO^\nu_{int}\to HO^\nu} = -\int_0^1 \langle E\rangle_{H(\lambda)}. \tag{4}$$

First, for each AD conformation, the system was confined to a harmonic oscillator state state with frequency ν=12.541 ps^{-1} (entry 17 in Table I of the main text). This was performed using the same confining procedure as that described in the main text, except that the highest simulated frequency was ν. At this stage, the first two terms on the right hand side of Eq. (2) were evaluated. They are 9.86$\pm$0.03 and 12.37$\pm$0.02 (units of kcal/mol) for the **c7eq** and **c7ax** states, respectively. To evaluate the final term, we performed twenty simulations, corresponding to the values $\lambda_i = [(i-1)/19]^2,\ i = 1,\ldots,20$. The above sequence of λ-values provides a fine distribution of windows in the region of small λ's, corresponding to near-annihilation of the force field. A 20ns MD simulation using the Hamiltonian in Eq. (3) was performed for each window, and the integral in Eq. (4) was evaluated using the trapezoidal rule. The computed values for $\Delta G_{HO^\nu_{int}\to HO^\nu}$ were 34.05$\pm$0.04 and 33.71$\pm$0.04 (units of kcal/mol) for the **c7eq** and **c7ax** states, respectively. Combining the terms gives the values $G_{\Omega_{c7eq}}$=-24.19$\pm$0.05 and $G_{\Omega_{c7ax}}$=-21.34$\pm$0.045, and a free energy difference ΔG=-2.85$\pm$0.07 (units of kcal/mol), as reported in Table II of the main text.

The above free energy values are consistent with those obtained using SCM (Table II in main text), although the uncertainty is slightly larger (*i.e.* 0.07 kcal/mol *vs.* 0.05 kcal/mol for the free energy difference). As noted in the main text, the computational effort involved in computing the absolute free energies using SCM is slightly larger than that with the confinement method followed

by force field annihilation. The annihilation above required 20ns simulation × 20 windows = 400ns of simulation for each state (a somewhat larger time would be required to reduce the overall uncertainty to 0.05 kcal/mol). From Fig. 3b in the main text, computing the free energies with SCM requires reaching frequencies of $\simeq$ 310 ps^{-1}. In addition, the simulations corresponding to $\nu \simeq$86, 163, 310 ps^{-1} require a reduced timestep. With the timestep set to 0.1fs, the 20ns SCM simulations for the three frequencies above are equivalent to 600ns of simulations with a time step of 1fs. An additional 40ns would be needed to reach the frequency $\nu \simeq$86 ps^{-1} from $\nu \simeq$12 ps^{-1} (according to the SCM simulation details given in the main text).

If only the free energy differences are desired, it is sufficient to perform SCM at the highest frequency of $\nu \simeq$86 ps^{-1} using Δt=1fs, which would require 3 × 20ns of simulation, starting from $\nu \simeq$12 ps^{-1}. The total computer time required to compute free energy differences between the two states in AD using SCM is 14 runs × 20ns / run × 2 states = 560ns. The total time required for the confinement method with annihilation is 11 runs × 20ns / run × 2 states = 440ns for the confinement part, in addition to the 400ns for the annihilation part computed above, totaling 840 ns. In the case that only the free energy differences are desired, the SCM is more efficient.

An important part of the comparison concerns the complexity of implementation of the confinement method with annihilation compared to that of SCM. Some MD software such as CHARMM[9] or GROMACS[10] have built-in options for scaling the potential energy at the level of the input script. For other programs, turning off the force field will involve code modifications, or force field modifications. The energy terms to be annihilated include both bonded and nonbonded interactions, and the latter also include energies from *e.g.* implicit solvation models. Depending on the software, this can be done at the level of the time integration code, at the level of the potential energy and force computation (which usually involve modifying multiple source-code files), or at the level of a force-field parameter file (*e.g.* one could create a series of force field files with scaled parameters and use them for different λ-windows). All of the approaches above are error-prone to different extents, and require testing. In addition, some MD software distributions do not include the source code. In our view, it is preferable not to have to modify the program source code, or the force field files. This is the case with SCM, which only requires that harmonic positional restraints be available in the distributed MD code. Such restraints are usually available because they are used for the initial heating and equilibration of simulation structures. We therefore believe that the implementation simplicity of SCM is another advantage that it has over the confinement method with annihilation.

References

[1] M. Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1909, 1996.

[2] J. Ménétrey, A. Bahloul, A.L. Wells, C.M. Yengo, C.A. Morris, H.L. Sweeney, and A. Houdusse. The structure of the myosin VI motor reveals the mechanism of directionality reversal. *Nature*, 435:779–785, 2005.

[3] J. Ménétrey, P. Llinas, M. Mukherjea, H.L. Sweeney, and A. Houdusse. The structural basis for the large powerstroke of myosin VI. *Cell*, 131:300–308, 2007.

[4] U. Haberthür and A. Caflisch. FACTS: Fast analytical continuum treatment of solvation. *J. Comput. Chem.*, 29:701–715, 2007.

[5] H. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Am. Statist. Assoc.*, 62:399–402, 1967.

[6] MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

[7] S. Park, A. Y. Lau, and B. Roux. Computing conformational free energy by deactivated morphing. *J. Chem. Phys.*, 129:134102, 2008.

[8] U. Hensen, H. Grubmüller, and O.F. Lange. Estimating absolute configurational entropies of macromolecules: The minimally coupled subspace approach. *PLoS ONE*, 5:e9179, 2010.

[9] B.R. Brooks, C.L. Brooks III, A.D. Mackerell Jr., L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, and et. al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30:1545–1614, 2009.

[10] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008. doi: 10.1021/ct700301q.