Supporting Information

Fenley et al. 10.1073/pnas.1213180109

SI Text

Thermodynamics of Clusters Selected by Control Variables. Information is shown in Table S1.

Thermodynamic Decomposition Results. The potential energy decomposition of the 1-ms bovine pancreatic trypsin inhibitor (BPTI) simulation was computed using the software utilities mdrun and g_energy of the GROMACS 4.5.5 simulation package (1) using simulation parameters matched to the original Anton simulation, including accounting for the ff99SB force-field parameters with isoleucine corrections (2, 3). Three potential energy calculations were performed for each molecular dynamics snapshot: all atoms, only solvent, and only protein. The solvent–protein potential energy of each snapshot was calculated as the all-atom potential energy less the solvent-only and protein-only energies. The potential energy values per snapshot were also used to calculate the heat capacity of each cluster using equation 9 in the work by Prabhu and Sharp (4). All thermodynamic quantities are presented in Table S2, and convergence plots of these quantities are shown in Figs. S1 and S2.

Convergence of Potential Energies. Information is shown in Fig. S1.

Convergence of Relative Heat Capacities. Information is shown in Fig. S2.

Estimation of Configurational Entropy. In preparation for investigating the configurational entropy, we established an internal coordinate system for BPTI based on bond lengths, bond angles, and torsions (5–8). Because the torsions account for the soft degrees of freedom that are responsible for almost all of the significant differences in motion between the clusters (bonds and angles showed little difference between the clusters; see below), we focused on their contribution to the configurational entropy. Torsions associated with a symmetry rotation (e.g., phenylalanine and tyrosine $\chi 2$ angles) were corrected to account for the symmetry. There are 889 torsions comprising all of the backbone, side chain, and phase angles of the protein. We used cpptraj, a program included with AmberTools 12 (9), to compute the value of each torsion for all available 4.1 million snapshots of the 1-ms BPTI simulation (10). The results for each snapshot were then partitioned according to the conformational cluster to which it belongs. 1D and 2D probability distribution functions (pdfs) and the resulting mutual information between the distributions were generated using the ACCENT software package (11). We also improved the configurational entropy estimates by adding the maximum information spanning tree (MIST) algorithm (12, 13) to ACCENT. The MIST algorithm, when assuming sufficient sampling, yields a rigorous upper bound to the true entropy, and at the level of pairwise correlation, it is found to yield a better estimate of the true entropy than the mutual information expansion (MIE), because MIE tends to overcompensate and subtract too much mutual information (13).

Block permutation of data. Limited sampling can cause two variables (e.g., i and j) that are statistically independent to seem correlated and hence, have nonzero computed mutual information. We corrected for such errors as follows. We applied a series of cyclic permutations to groups of consecutive measurements (or blocks) for variable i relative to variable j. The mutual information between the permuted variable and unpermuted variable was then calculated for each permutation and averaged over the permutations before being subtracted from the original mutual information of variables i and j. The blocks used for the cyclic permutations correspond to all of the measurements in a set of frames that starts at the transition into a cluster and ends at the transition out of the cluster. Therefore, the frames inside each block represent a subset of the simulation, where the system remained in a given cluster without undergoing a transition into a different cluster. These blocks can consist of tens of microseconds or more of simulation data. We omitted the first and last 500 ns worth of frames around each cluster–cluster transition to minimize memory of the previous cluster at the transition.

The mutual information values from the permutations described above were averaged, and the average was subtracted from the mutual information associated with the original (physical) ordering of the data. Because the permutations disrupt the correlation between the two torsions without disrupting the independent time evolution of the torsions, the resulting mutual information is a representation of the spurious correlation that appears between sets of numerical measurements for the available sample set. The final computed configurational entropy estimates are, hence, conservative with respect to the total amount of correlation entropy compared with similar approaches, where the data are simply scrambled to estimate the level of spurious correlation rather than cyclically permuted in coarse blocks as done here.

Number of histogram bins. In a preliminary study, we examined the sensitivity of the configurational entropy estimates on the number of histogram bins used in the ACCENT program. This analysis used all nonphase torsion angles. The raw second-order MIST configurational entropy estimates for each cluster varied over only ~2.5 kcal/mol as the number of bins varied between 60 and 180, and the estimates (after applying the cyclic permutation corrections) varied over only ~1.2 kcal/mol. Furthermore, the correlation entropies (defined as the second-order MIST estimate less the first-order MIST estimate) varied over only 0.2 kcal/mol as the number of bins varied. Therefore, the dependency of the configurational entropy estimates on the number of bins is essentially isolated to the first-order contribution. The differences between the permutation-corrected second-order entropies of the three clusters varied by only 0.3 kcal/mol as the number of bins varied. Subsequently, all reported configurational entropy estimates were done with the number of histogram bins set to 120.

Convergence of permutation-corrected second-order MIST entropy estimates. Information is shown in Fig. S3.

Entropy Contribution of Bond Stretches and Angle Bends. The bond angle–torsion coordinate system set up for the configurational entropy calculations included 453 bond lengths not involving hydrogens and 890 angle bends not involving hydrogens. We used second-order MIST to estimate the additional configurational entropy associated with these degrees of freedom without using the cyclic permutation correction. Adding the entropy contributions of these additional variables produced, at most, a 3.0-kcal/mol difference between cluster configurational entropies. Cyclic permutation corrections are expected to further reduce the range, making the bond and angle configurational entropies essentially the same for all three clusters.

Equilibrium Conditional pdf as a Response Function. For a molecular system at thermal equilibrium, let the spatial variables (atomic coordinates) be partitioned into one subset, x_c, termed the control variables, and a second subset, x_t, comprising all of the other variables. According to probability theory, the pdf over x_t conditioned on the control variables x_c, $p(x_t|x_c)$, connects the marginal

pdf of the control variables, $p(x_c)$, to the joint pdf over all variables, $p(x_t,x_c)$, as follows:

$$p(x_t|x_c) = \frac{p(x_t,x_c)}{p(x_c)}. \quad \textbf{[S1]}$$

Statistical thermodynamics allow these joint, marginal, and conditional pdfs, respectively, to be expressed in terms of configuration integrals

$$p(x_t,x_c) = \frac{e^{-\beta E(x_t,x_c)}}{\int e^{-\beta E(x_t,x_c)} dx_t dx_c}, \quad \textbf{[S2]}$$

where β is $1/RT$,

$$p(x_c) = \frac{\int e^{-\beta E(x_t,x_c)} dx_t}{\int e^{-\beta E(x_t,x_c)} dx_t dx_c}, \quad \textbf{[S3]}$$

and

$$p(x_t|x_c) = \frac{e^{-\beta E(x_t,x_c)}}{\int e^{-\beta E(x_t,x_c)} dx_t}. \quad \textbf{[S4]}$$

Here, $E(x_t,x_c)$ signifies the potential function over all variables. If the control variables are now manipulated by the addition of a perturbing potential function that depends on them alone, $E'(x_c)$, then the new equilibrium joint pdf will be given by

$$p'(x_t,x_c) = p(x_t|x_c)p'(x_c), \quad \textbf{[S5]}$$

where $p'(x_c)$ is the new pdf over the control variables. (Note that the conditional pdf is unchanged by the perturbation.) This result is derived by writing out the right-hand side of the latter expression in terms of configuration integrals:

$$\begin{aligned} p(x_t|x_c)p'(x_c) &= \frac{e^{-\beta E(x_t,x_c)}}{\int e^{-\beta E(x_t,x_c)} dx_t} \frac{\int e^{-\beta\left[E(x_t,x_c)+E'(x_c)\right]} dx_t}{\int e^{-\beta[E(x_t,x_c)+E'(x_c)]} dx_t dx_c} \\ &= \frac{e^{-\beta E(x_t,x_c)}}{\int e^{-\beta E(x_t,x_c)} dx_t} \frac{e^{-\beta E'(x_c)} \int e^{-\beta E(x_t,x_c)} dx_t}{\int e^{-\beta[E(x_t,x_c)+E'(x_c)]} dx_t dx_c} \\ &= \frac{e^{-\beta\left[E(x_t,x_c)+E'(x_c)\right]}}{\int e^{-\beta[E(x_t,x_c)+E'(x_c)]} dx_t dx_c} = p'(x_t,x_c) \end{aligned} \quad \textbf{[S6]}$$

In this sense, then, the conditional pdf $p(x_t|x_c)$ of the unperturbed system represents a response function enabling one to predict how the rest of the system will respond to a change in the pdf of any chosen control variables. Intuitively, this finding results from the fact that the perturbing potential depends, by assumption, only on the control variables, and therefore, the perturbing potential does not affect the energy landscape of the transducing variables for any given values of the control variables. Instead, the perturbing potential changes the distribution of the control variables, and this change in distribution indirectly changes the distribution of the transducing variables. This indirect effect is captured by the conditional pdf. Although no real perturbation is perfectly local, as assumed here, the approximation of locality is expected to be good in the common situation where a small ligand binds a large protein; therefore, many atoms of the protein are too far from the ligand to interact with it significantly. Note that, when the conditional pdf is computed from simulation data, it can predict the system's response only if the modified probability density of the control variables, $p'(x_c)$, falls within the range of x_c, where the original pdf of the control variables, $p(x_c)$, is significantly populated. The conformational selection scenario considered in the case of BPTI meets this criterion.

General Formalism for Entropy–Enthalpy Transduction. The standard free energy of binding is given by $\Delta G^o = \mu_q^o - \mu_p^o - \mu_l^o$, where μ_X^o (X = q, p, l) represents the standard chemical potential of the complex, free protein, and free ligand, respectively, and each chemical potential can be divided into a partial molar energy and entropy: $\mu_X^o = E_X - TS_X^o$. (The superscript o is associated with the entropy, because this term is affected by the standard concentration, C_0.) We wish to partition the binding energy and entropy and hence, the binding free energy into intrinsic and transduced contributions. This partitioning will be accomplished by corresponding partitionings of the energies and entropies of both the free protein and complex. No partitioning is needed for the free ligand, because its whole energy and entropy will be incorporated into the intrinsic binding thermodynamics. Below, we derive the partitioning for the free protein; the partitioning for the complex is not derived here, because it follows by straightforward analogy. For simplicity, solvent degrees of freedom are not explicitly considered. Note that the energy changes presented here essentially equal enthalpy changes because of the low compressibility of aqueous solutions and the small energy density of 1 atm ambient pressure.

The standard chemical potential of the free protein is

$$\mu_p^o = \langle E_p \rangle - TS_p^o, \quad \textbf{[S7]}$$

where angle brackets indicate a Boltzmann average. The spatial coordinates, x, are separated into control variables associated with atoms local to the binding site, x_c, and transducing variables, x_t, associated with more remote atoms. The potential energy function of the whole system $E_p(x_c, x_t)$ is expressed as the sum of a binding site part that depends only on the control variables and a transducing part that depends on the transducing variables as well:

$$E_p(x_c,x_t) = E_c(x_c) + E_t(x_c,x_t). \quad \textbf{[S8]}$$

Therefore,

$$\langle E_p \rangle = \langle E_c \rangle_p + \langle E_t \rangle_p, \quad \textbf{[S9]}$$

where p subscripts are included on the angle brackets to clarify that these averages are taken in the ensemble of the free protein. The first average, which will contribute to the intrinsic part of the binding energy, is left simply as

$$\langle E_c \rangle_p = \int p_{p,c}(x_c) E_c(x_c) dx_c, \quad \textbf{[S10]}$$

where $p_{p,c}(x_c)$ is the pdf over the control variables in the ensemble of the free protein. The second average, which will contribute to the transduced part of the binding energy, is rewritten in terms of the conditional probability of the transducing variables on the control variables:

$$\langle E_t \rangle_p = \int p_p(x_c,x_t) E_t(x_c,x_t) dx_c dx_t. \quad \textbf{[S11]}$$

Now $p_p(x_c, x_t)$, the pdf over all variables in the ensemble of the free protein, is rewritten as $p_p(x_c)\, p(x_t|x_c)$ to yield

$$\langle E_t \rangle_p = \int p_{p,c}(x_c) \left[\int p(x_t|x_c) E_t(x_c, x_t) dx_t \right] dx_c = \int p_{p,c}(x_c) \overline{E}_t(x_c) dx_c. \quad \text{[S12]}$$

Here, the energy response function of the transducing part of the protein, $\overline{E}_t(x_c)$, is given by the bracketed integral in the first line and matches the corresponding expression in the text.

Now the entropy of the free protein may be written in terms of the Gibbs–Shannon formula as

$$S_p^o = R \ln\left(\frac{8\pi^2}{C^o}\right) - R \int p_p(x_c, x_t) \ln\, p_p(x_c, x_t) dx_c dx_t. \quad \text{[S13]}$$

The initial term on the right comes from the overall orientation and translation of the protein in solution at standard concentration (14). Substituting $p_p(x_c)\, p(x_t|x_c)$ for $p_p(x_c, x_t)$, splitting the log, and gathering terms on the right, we obtain

$$S_p^o = R \ln\left(\frac{8\pi^2}{C^o}\right) - R \int p_{p,c}(x_c) p(x_t|x_c) \ln\, p_{p,c}(x_c) dx_c dx_t - R \int p_{p,c}(x_c) \left[\int p(x_t|x_c) \ln\, p(x_t|x_c) dx_t \right] dx_c. \quad \text{[S14]}$$

Recognizing that $\int p(x_t|x_c) dx_t = 1$ for any x_c and identifying the quantity in brackets in the third term on the right as the entropy response, we write

$$S_p^o = R \ln\left(\frac{8\pi^2}{C^o}\right) + S_{p,c} + S_{p,t}$$
$$S_{p,c} = -R \int p_{p,c}(x_c) \ln\, p_{p,c}(x_c) dx_c \quad \text{[S15]}$$
$$S_{p,t} = -R \int p_{p,c}(x_c) \overline{S}_t dx_c.$$

The first two terms on the right of the first line contribute to the intrinsic binding entropy, and the final term contributes to the transduced binding entropy.

For the ligand–protein complex, in addition to the existing set of control variables, the potential function also depends on the conformation, position, and orientation of the bound ligand, (x_l, r, ω), and the ligand's long-ranged energetic interactions with atoms beyond the binding site may be approximated as a small constant E_{lt}:

$$E_q(x_l, r, \omega, x_c, x_t) = E_{lc}(x_l, r, \omega, x_c) + E_{lt} + E_t(x_c, x_t). \quad \text{[S16]}$$

We decompose the energy and entropy of the complex by exact analogy with the free protein, obtaining

$$\langle E_{lc} \rangle_q = \int p_{lc}(x_l, r, \omega, x_c) E_{lc}(x_l, r, \omega, x_c) dx_l dr d\omega dx_c$$
$$\langle E_t \rangle_q = \int p_{q,c}(x_c) \overline{E}_t(x_c) dx_c$$
$$S_{q,lc} = -R \int p_{lc}(x_l, r, \omega, x_c) \ln p_{lc}(x_l, r, \omega, x_c) dx_l dr d\omega dx_c dx_t \quad \text{[S17]}$$
$$S_{q,t} = -R \int p_{q,c}(x_c) \overline{S}_t dx_c,$$

where $p_{lc}(x_l, r, \omega, x_c)$ is the pdf over the ligand and control variables for the bound complex and $p_{q,c}(x_c)$ is the pdf over the control variables for the bound complex.

The thermodynamics of binding may now be partitioned into intrinsic and transduced parts:

$$\Delta G^o = \Delta G^o_{int} + \Delta G_t$$
$$\Delta E = \Delta E_{int} + \Delta E_t \quad \text{[S18]}$$
$$\Delta S^o = \Delta S^o_{int} + \Delta S_t.$$

The intrinsic components of the binding thermodynamics are given by

$$\Delta E_{int} = \langle E_{lc} \rangle_q + E_{lt} - \langle E_c \rangle_p - \langle E_l \rangle_l$$
$$\Delta S^o_{int} = -R \ln \frac{8\pi^2}{C^o} + S_{q,lc} - S_{p,c} - S_l \quad \text{[S19]}$$
$$\Delta G^o_{int} = \Delta E_{int} - T \Delta S^o_{int},$$

where S_l and E_{ll} represent, respectively, the entropy and mean energy of the free ligand. The transduced components of the binding thermodynamics, in turn, are given by

$$\Delta E_t = \langle E_t \rangle_q - \langle E_t \rangle_p = \int \left[p_{q,c}(x_c) - p_{p,c}(x_c) \right] \overline{E}_t(x_c) dx_c$$
$$\Delta S_t = S_{q,t} - S_{p,t} = \int \left[p_{q,c}(x_c) - p_{p,c}(x_c) \right] \overline{S}_t(x_c) dx_c \quad \text{[S20]}$$
$$\Delta G_t = \Delta E_t - T \Delta S_t.$$

1. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4(3):435–447.
2. Hornak V, et al. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65(3):712–725.
3. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78(8):1950–1958.
4. Prabhu NV, Sharp KA (2005) Heat capacity in proteins. *Annu Rev Phys Chem* 56: 521–548.
5. Pitzer KS (1946) Energy levels and thermodynamic functions for molecules: II. Unsymmetrical tops attached to a rigid frame. *J Chem Phys* 14(4):239–243.
6. Herschbach DR, Johnston HS, Rapp D (1959) Molecular partition functions in terms of local properties. *J Chem Phys* 31(6):1652–1661.
7. Gō N, Scheraga HA (1976) On the use of classical statistical mechanics in the treatment of polymer chain conformation. *Macromolecules* 9(4):535–542.
8. Chang C-E, Potter MJ, Gilson MK (2003) Calculation of molecular configuration integrals. *J Phys Chem B* 107(4):1048–1055.
9. Case DA, et al. (2012) *AMBER 12* (University of California, San Francisco).
10. Shaw DE, et al. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* 330(6002):341–346.
11. Killian BJ, Yundenfreund Kravitz J, Gilson MK (2007) Extraction of configurational entropy from molecular simulations via an expansion approximation. *J Chem Phys* 127 (2):024107.
12. King BM, Tidor B (2009) MIST: Maximum Information Spanning Trees for dimension reduction of biological data sets. *Bioinformatics* 25(9):1165–1172.
13. King BM, Silver NW, Tidor B (2012) Efficient calculation of molecular configurational entropies using an information theoretic approximation. *J Phys Chem B* 116(9): 2891–2904.
14. Gilson MK, Given JA, Bush BL, McCammon JA (1997) The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys J* 72(3):1047–1069.

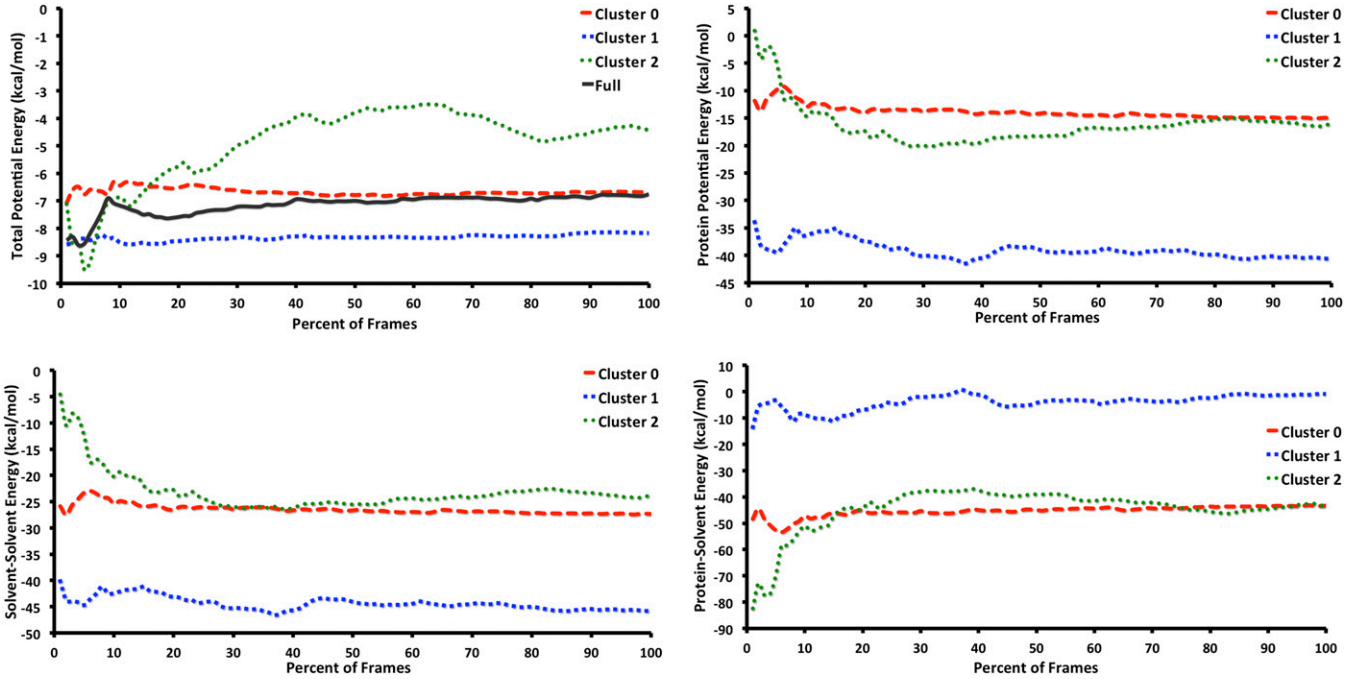

Fig. S1. Convergence plots of the total potential energy and its components (kilocalories per mole) plotted as a function of the percentage of available trajectory frames. Note that much of the apparent difference in convergence stems from the fact that each cluster has a different number of frames (cluster 2 has the fewest). (*Upper Left*) Total potential energy of the three highest population clusters and the full trajectory. Potential energies are offset by 48,820 kcal/mol for readability. (*Upper Right*) Potential energy of protein considered in isolation offset by 850 kcal/mol. (*Lower Left*) Potential energy of solvent–solvent interactions considered in isolation offset by 45,760 kcal/mol. (*Lower Right*) Protein–solvent interaction energy offset by 2,119 kcal/mol for readability.

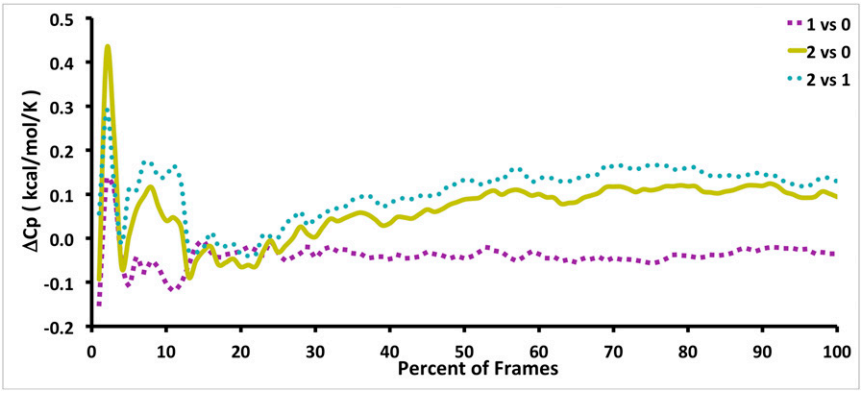

Fig. S2. Convergence plots of the relative heat capacities computed from the BPTI trajectory as $Cp = \langle dU^2 \rangle / RT^2$ between the conformational clusters (kilocalories per mole per Kelvin) plotted as a function of the percentage of available trajectory frames.

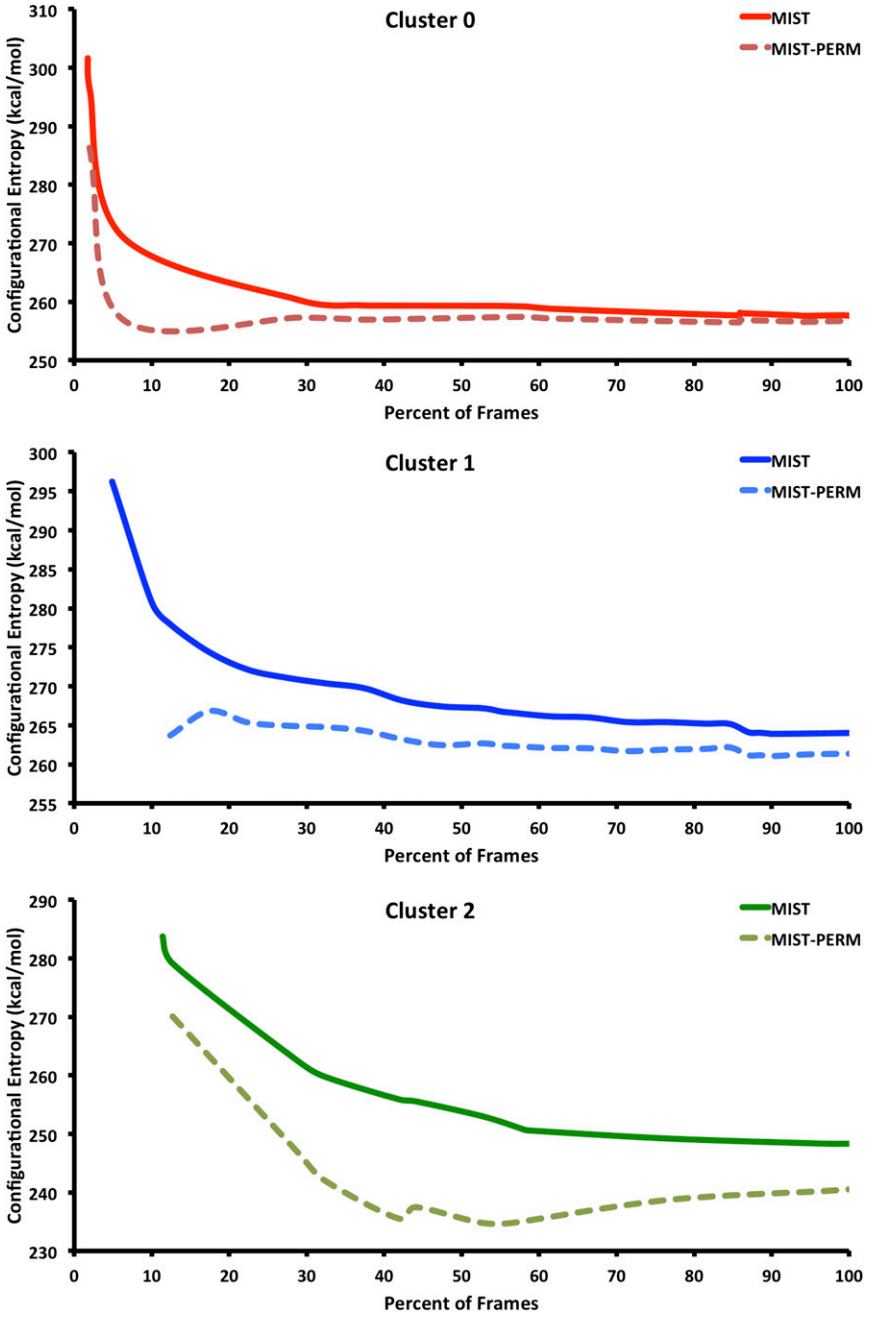

Fig. S3. Convergence plots of the configurational entropy (kilocalories per mole) with (MIST-PERM) and without (MIST) cyclic permutation corrections plotted as a function of the percentage of available trajectory frames. Note that much of the apparent difference in convergence stems from the fact that each cluster has a different number of frames (cluster 2 has the fewest). PERM, permutation.

Table S1. Numerical values from thermodynamic decomposition, where the cluster to which each trajectory frame belongs is determined based on the optimal control variable for each cluster (kilocalories per mole)

Cluster	ΔG	ΔE	−TΔS
0 vs. 1	−0.33	1.10	−1.43
2 vs. 1	0.36	3.70	−3.34

Table S2. Cross-cluster comparisons of energies (kilocalories per mole) and heat capacities (kilocalories per mole per Kelvin) from thermodynamic decomposition of the BPTI trajectory

Cluster comparison	ΔG	ΔE	−TΔS	$\Delta E_{protein}$	$\Delta E_{solvent\text{-}solvent}$	$\Delta E_{solvent\text{-}protein}$	$-T\Delta S_{config}$	$-T\Delta S_{solvent}$	ΔCp
0 vs. 1	−0.43	1.49	−1.92	25.5	18.4	−42.4	−2.75	0.83	0.04
2 vs. 1	0.52	3.74	−3.22	24.4	21.8	−42.5	−18.9	15.7	0.13