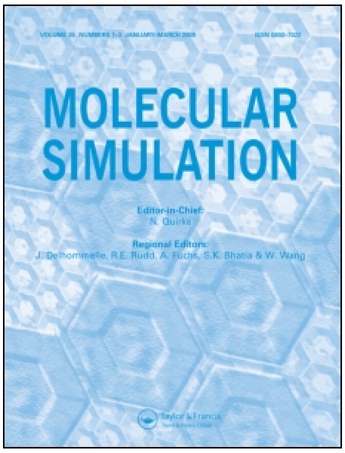

**Molecular Simulation**
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

## A replica-exchange approach to computing peptide conformational free energies

M. Scott Shell[a]
[a] Department of Chemical Engineering, University of California, Santa Barbara, CA, USA

Online publication date: 03 August 2010

**To cite this Article** Shell, M. Scott(2010) 'A replica-exchange approach to computing peptide conformational free energies', Molecular Simulation, 36: 7, 505 — 515

**To link to this Article: DOI:** 10.1080/08927021003720546

**URL:** http://dx.doi.org/10.1080/08927021003720546

*Molecular Simulation*
Vol. 36, Nos. 7–8, June–July 2010, 505–515

Taylor & Francis
Taylor & Francis Group

# A replica-exchange approach to computing peptide conformational free energies

M. Scott Shell*

*Department of Chemical Engineering, University of California, Santa Barbara, CA 93106-5080, USA*

(*Received 17 December 2009; final version received 22 February 2010*)

We describe a methodology for computing relative free energies of peptide conformational states. The method uses replica-exchange molecular dynamics (REMD) simulations with backbone restraints that hold a peptide within the conformational space basin surrounding a given structure of interest. Restraints are scaled from full strength to zero strength as one moves up the replica cascade, such that the highest temperature becomes a conformation-unspecific reference state. Then, weighted histogram analysis is used to compute the free energy of the structure of interest relative to the reference state. We test our method on a 14-residue sequence that adopts a variety of alpha and beta structures at equilibrium. We show that computed free energies bear close theoretical agreement with structure populations and have the expected relationship with structure unfolding temperatures evaluated in non-equilibrium heating simulations. By comparison, purely potential energy measures are not able to correctly rank populations, even with harmonic free energy corrections. These results suggest that the methodology can compute free energies underlying structural stability and populations. They further suggest that entropic contributions to free energies and populations are significant, and that force-field potential energies alone cannot be used as structure prediction scoring functions, at least for small peptides.

**Keywords:** replica exchange; free energies; peptides; conformations

## 1. Introduction

Proteins and peptides in solution experience conformational fluctuations that play critical roles in their function in living systems [1,2]. Allosteric changes are key to regulating enzyme activity, binding and recognition, and self-assembly, and can contribute to misfolding and aggregation-related diseases [3,4]. In small peptides, conformational fluctuations can be significant, with some peptides even adopting both helical and beta states, and these fluctuations can greatly affect binding propensity and self-assembly behaviour into larger structures and aggregates [5].

The free energies associated with different conformational states are the determinants of equilibrium polypeptide structural ensembles. The ability to assess such free energies, therefore, is a fundamental prerequisite in understanding the structural changes underlying protein function. However, in many cases, it remains challenging to compute quantitative conformational free energies, owing to practical difficulties in achieving extensive sampling of proteins' overwhelmingly large conformational spaces. It is particularly challenging to elucidate the free energies of structures that are higher in free energy than the native by more than a few $k_BT$, as these are normally visited with exponentially lower probability by the protein at equilibrium.

Thus, a computational method capable of determining accurate conformational free energies would be of great value. It would enable assessment of the relative stabilities of different structures of a protein. Given a putative sequence of structural events, such a method would also enable the calculation of free energy pathways and the driving forces associated with structural transitions. Practically, the calculation of conformational free energies could be coupled with experimentally determined structures to shed light on the quality of protein force fields, or could be applied to protein structure predictions, in principle, to characterise the quality of putative structures.

In this work, we present an algorithm for computing relative conformational free energies that is based on the replica-exchange molecular dynamics (REMD) method [6]. Our method applies flat-bottom harmonic torsional restraints to a given peptide, constraining it to a particular backbone configuration. The strength of these restraints is adjusted from 100% in the lowest temperature replica (at the temperature of interest) to 0% in the highest one. After running the REMD simulation to equilibrium, a histogram-reweighting approach [7] is used to determine the free energy difference between the low- and high-temperature replicas, taking into account differences in restraint energies. In this way, the free energy difference of a particular conformation, relative to a structure-independent high-temperature unfolded state, can be determined.

We apply this method to a 14-residue 'chameleon' sequence from the MATα2/MCM1/DNA complex that was earlier studied by Ikeda and Higo [8]. We show that the computed free energies correlate well with equilibrium

*Email: shell@engineering.ucsb.edu

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927021003720546
http://www.informaworld.com

conformational populations of a variety of alpha and beta configurations, whereas minimised energies and average energies do not, even when a normal-model harmonic vibrational free energy correction is applied. Moreover, we find good correspondence between the computed free energies and structure unfolding temperatures estimated from non-equilibrium heating simulations.

## 2. Theoretical background

The total configurational partition function $Z$ for a system composed of a protein and aqueous solvent is given by:

$$Z = \int\int \mathrm{e}^{-\beta U_{\mathrm{PP}}(\mathbf{r}) - \beta U_{\mathrm{PW}}(\mathbf{r},\mathbf{R}) - \beta U_{\mathrm{WW}}(\mathbf{R})}\,\mathrm{d}\mathbf{R}\,\mathrm{d}\mathbf{r}, \tag{1}$$

where $U_{\mathrm{PP}}$, $U_{\mathrm{PW}}$ and $U_{\mathrm{WW}}$ are the protein intramolecular, protein–water and water–water interaction potentials, $\mathbf{r}$ denotes the collective degrees of freedom of the protein atoms and $\mathbf{R}$ gives those of the water atoms. As per the usual notation, $\beta = 1/k_{\mathrm{B}}T$. The partition function above can be formally written as a set of integrals over the protein conformations alone:

$$Z = \int \mathrm{e}^{-\beta U_{\mathrm{PP}}(\mathbf{r}) - \beta F_{\mathrm{W}}(\mathbf{r};\beta)}\mathrm{d}\mathbf{r} = \int \mathrm{e}^{-\beta U(\mathbf{r})}\,\mathrm{d}\mathbf{r}. \tag{2}$$

In these expressions, the solvation free energy $F_{\mathrm{W}}(\mathbf{r};\beta)$ is defined by an integral over solvent degrees of freedom:

$$F_{\mathrm{W}}(\mathbf{r};\beta) = -k_{\mathrm{B}}T \ln \int \mathrm{e}^{-\beta U_{\mathrm{PW}}(\mathbf{r},\mathbf{R}) - \beta U_{\mathrm{WW}}(\mathbf{R})}\,\mathrm{d}\mathbf{R}. \tag{3}$$

Here, we also define the effective protein intramolecular energy function $U(\mathbf{r}) \equiv U_{\mathrm{PP}}(\mathbf{r}) + F_{\mathrm{W}}(\mathbf{r})$, where the $\beta$-dependence of $U$ due to the solvation free energy has been omitted to simplify the notation. It is important to note that $U(\mathbf{r})$ is the basis of many of the force fields commonly used in implicit-solvation simulations, where $F_{\mathrm{W}}(\mathbf{r})$ is typically formed as the sum of an approximate solution to the continuum electrostatics and an interfacial term proportional to the protein solvent-accessible surface area [9,10].

The above expressions for the partition function correspond to the total configurational Helmholtz free energy of the system, $A = -k_{\mathrm{B}}T \ln Z$. If one is interested in computing the free energy for a particular conformation and fluctuations about it, it is necessary to restrict the configurational integral to the subset of configuration space in the vicinity of the structure of interest. Denoting this subset as $\Gamma^*$ for a configuration $\mathbf{r}^*$, the configurational free energy $A^*$ is given by:

$$\begin{aligned} A^* &= -\,k_{\mathrm{B}}T \ln Z^* \\ &= -k_{\mathrm{B}}T \ln \int_{\Gamma^*} \mathrm{e}^{-\beta U(\mathbf{r})}\,\mathrm{d}\mathbf{r}. \end{aligned} \tag{4}$$

Similarly, the equilibrium fractional population of the configuration (including fluctuations) is

$$\begin{aligned} \wp(\mathbf{r}^*) &= \frac{Z^*}{Z} \\ &\propto \mathrm{e}^{-\beta A^*}. \end{aligned} \tag{5}$$

An immediate consideration is how to determine the subset of configuration space $\Gamma^*$ for use in these expressions. For configurational free energies to be meaningful, it is essential that they are insensitive to small errors in determining $\Gamma^*$. Indeed, the very significance of such free energies relies upon the natural emergence of distinct regimes of configuration space that are populated by the system at equilibrium, and this task is intimately tied to the determination of metastable states in the system [11]. One typically envisions an energy landscape characterised by a number of 'metabasins' surrounding each metastable configuration that are mutually separated by mostly high, infrequently traversed energy barriers and a very small configuration-space volume of low-energy pathways connecting them [12]. Here, as per Equation (13), such energy landscapes include interatomic potential energies and a conformation-dependent solvation free energy. If one then defines the subset $\Gamma^*$ using a particular metabasin, the contribution of states to $Z^*$ at the borders of $\Gamma^*$ is negligible due to exponential damping in the partition function. Then, variations in $\Gamma^*$ at its borders have negligible effects on $A^*$ as the primary contributions to the integral emerge from states surrounding the basin minimum.

These ideas are illustrated with the well-known normal-mode approximation to the free energy [13]. Assume that a metabasin can be approximated by a harmonic expression of the form $U = U_0 + (1/2)\sum \lambda_i q_i^2$, where $q_i$ give the displacement about the minimum along the eigenvectors of the mass-weighted Hessian (normal modes) and $\lambda_i$ give the corresponding eigenvalues [14]. Then, the free energy is given by:

$$\begin{aligned} A^* &= U_0 - k_{\mathrm{B}}T \sum_i \ln \int_{-L}^{L} \mathrm{e}^{-\beta\lambda_i q^2/2}\,\mathrm{d}q \\ &= U_0 + k_{\mathrm{B}}T \sum_i \left[ \ln \sqrt{\frac{\beta\lambda_i}{2\pi}} - \ln \operatorname{erf} \sqrt{\frac{\beta\lambda_{\mathrm{i}}L^2}{2}} \right]. \end{aligned} \tag{6}$$

Here, we have defined the volume of the metabasin in configuration space as that which spans a distance $\pm L$ along each eigenvector surrounding the minimum. The last of the two terms inside the sum is the only one involving the location of the borders of $\Gamma^*$, by virtue of $L$. The contribution of this term to the free energy grows to less than $0.005\,k_{\mathrm{B}}T$ when $\lambda L^2/2 > 2\,k_{\mathrm{B}}T$ and decays sharply thereafter. Therefore, provided the basin volume characterised by $L$ is sufficient to include the low-energy portions of the landscape (roughly those within $2k_{\mathrm{B}}T$ of the minimum), the exact location of its borders has little effect on the free energy.

We return now to the definition of the configurational free energy. The restricted configuration space integral can be rewritten using an energetic restraint that penalises conformations outside of $\Gamma^*$,

$$Z^* = \lim_{\alpha\to\infty} \int \mathrm{e}^{-\beta U(\mathbf{r}) - \alpha\beta U_{\mathrm{rest}}(\mathbf{r};\mathbf{r}^*)}\, \mathrm{d}\mathbf{r}, \tag{7}$$

where $\alpha$ measures the strength of the restraint and has a role similar to that of a coupling parameter (with respect to the restraint energy). The function $U_{\mathrm{rest}}(\mathbf{r};\mathbf{r}^*)$ satisfies the following properties:

$$U_{\mathrm{rest}} = 0 \;\; \text{for } \mathbf{r} \in \Gamma^*, \quad U_{\mathrm{rest}} > 0 \;\; \text{otherwise}. \tag{8}$$

While formally exact, such an expression is not convenient for MD simulations in the light of the limit, which implies an infinite restraint energy for configurations outside of $\Gamma^*$. We may instead consider the case in which the restraint energy grows to large (several $k_{\mathrm{B}}T$) but finite values in the vicinity of the borders of $\Gamma^*$ and continues to grow larger outside. As the border contributions to the configurational free energy are negligible as per the discussion above, the exact form of the restrained energy should not affect its final value so long as they maintain the system within the metabasin surrounding $\mathbf{r}^*$ and are only non-zero at and outside of the borders. With these considerations, we finally express the free energy as

$$A^* \approx -k_{\mathrm{B}}T \ln \int \mathrm{e}^{-\beta U(\mathbf{r}) - \beta U_{\mathrm{rest}}(\mathbf{r};\mathbf{r}^*)}\, \mathrm{d}\mathbf{r}. \tag{9}$$

## 3. Methods

### 3.1 Basic models and methods

We study the 14-residue chameleon sequence from the MATα2/MCM1/DNA complex as was examined by Ikeda and Higo [8]. In our runs, we modify the sequence with N- and C-terminal capping groups. The AMBER 9 program [15,16] is used for all of our calculations, with custom Python wrappers. Interatomic interactions are modelled using the AMBER ff96 force field [17] with the implicit solvation model, denoted ‘igb = 5’, of Onufriev et al. [18]. In previous efforts, we and our co-workers found that this force field well stabilised the folds of a number of small peptides [19,20] and, when combined with a mechanism-based conformational search algorithm, could fold several small proteins [21,22]. Efforts by other researchers have found in short simulations that the ff96 force field did not adequately stabilise helices [23–26], although other groups and newer work using longer simulations have found it to exhibit alpha/beta balance with implicit GBSA models [27–29]. Our recent work shows that this force field performs reasonably given adequate simulation time, of the order of 40–60 ns [20]. We emphasise that this study is not a test of force-field performance, but rather one of self-consistency between equilibrium structural populations and calculated free energies. In addition to the ff96 force field, we also apply the fix proposed by Simmerling and co-workers [30] that diminishes the intrinsic hydrogen radii on basic residues, in order to bring implicit solvation ion-pairing stabilities in better line with those found in atomistic simulations.

We simulate the systems using REMD [6]. In usual REMD, a number of copies (‘replicas’) of a system are evolved in parallel at different temperatures spanning physiological conditions to a heated state more conducive to traversing free energy barriers. Periodic swaps of neighbouring replicas enable conformations to heat up and cool down, and form a Markov chain that maintains an overall Boltzmann-weighted ensemble at each temperature, provided swaps are accepted with probability:

$$\wp_{\mathrm{acc}} = \min[1, \mathrm{e}^{-\beta_1 U_1(\mathbf{r}_2) - \beta_2 U_2(\mathbf{r}_1) + \beta_1 U_1(\mathbf{r}_1) + \beta_2 U_2(\mathbf{r}_2)}]. \tag{10}$$

Here $\beta = 1/k_{\mathrm{B}}T$ and the subscripts 1 and 2 indicate the replicas of interest, with $\mathbf{r}_1$ and $\mathbf{r}_2$ their corresponding configurations prior to the swap. In the usual case that the potential energy function remains constant across all replicas, this acceptance probability simplifies to

$$\wp_{\mathrm{acc}} = \min[1, \mathrm{e}^{\Delta\beta\Delta U}], \tag{11}$$

where $\Delta\beta = \beta_2 - \beta_1$ and $\Delta U = U(\mathbf{r}_2) - U(\mathbf{r}_1)$.

In our simulations, our replica temperatures span 270–600 K and swaps are attempted every 20 ps of the MD simulation. At each round of replica swaps, five attempts between neighbour pairs are made in order to facilitate convergence; all swaps in each round are randomised in order. The number of replicas is 20, adjusted so that swap attempts are accepted on average with ~30–50% probability. Other details of the REMD procedure can be found in our previous publications [19,20,22,31]. For initial runs, simulations are initiated with peptides in extended conformations, whereas for conformational free energy calculations, simulations are initiated from the reference structure. Trajectory frames are recorded every 1 ps and used in subsequent analysis.

Clustering is performed using a modified $k$-means algorithm and a cluster member cut-off of 2 Å RMSD [19,20,22,31]. Populations are determined from the fraction of total trajectory frames that are found to be a member of each cluster, and the structure representing a cluster is taken from the member closest to the cluster centroid.

### *3.2 Conformational free energies*

To compute conformational free energies, we modify the REMD protocol to include backbone torsional restraints that hold the system in a target reference configuration. These restraints are not constant across all replicas, but gradually taper from full strength in the lowest temperature replica to zero in the highest. In this way, the cascade of coupled replicas ultimately provides a way to compute the free energy difference of the conformation-restrained replica at the low temperature of interest and a conformation-unspecific high-temperature reference state free of any restraints.

We express the total force field in each replica $i$ as the sum of the usual interatomic potential and a scaled restraint term:

$$U'_i(\mathbf{r}) = U(\mathbf{r}) + \lambda_i U_{\text{rest}}(\mathbf{r}).$$

The choice of this simple linear scaling in the restraint energy is not unique; alternatives include changes to the shape of the restraint itself, although we did not explore such modifications. With this form of the energy function, the acceptance probability for swaps to maintain a canonical, restraint-weighted stationary distribution in each replica becomes

$$\wp_{\text{acc}} = \min\left[1, \mathrm{e}^{\Delta\beta\Delta U + \Delta(\beta\lambda)\Delta U_{\text{rest}}}\right], \tag{12}$$

with $\Delta(\beta\lambda) = \beta_2\lambda_2 - \beta_1\lambda_1$ and $\Delta U_{\text{rest}} = U_{\text{rest}}(\mathbf{r}_2) - U_{\text{rest}}(\mathbf{r}_1)$.

For the backbone torsional restraints, we use a flat-bottom harmonic well for each phi and psi angle (of the form implemented in AMBER [15]):

$$u_{\text{rest}}(\theta) = \begin{cases} \frac{k}{2}(\theta - \theta^* + \delta\theta)^2 & \theta < \theta^* - \delta\theta, \\ 0 & \theta^* - \delta\theta \le \theta < \theta^* + \delta\theta, \\ \frac{k}{2}(\theta - \theta^* - \delta\theta)^2 & \theta \ge \theta^* + \delta\theta. \end{cases} \tag{13}$$

Here $\theta^*$ is the corresponding reference angle from the conformation whose free energy is of interest, $\delta\theta$ is half of the flat-bottom width and $k$ is the force constant of the harmonic restraint. In practice, the value of $\theta$ for a given dihedral angle is translated by $2\pi$ so as to be closest to $\theta^*$. Our use of the flat-bottom restraint, rather than a simple parabolic form, is motivated by the desire to leave unchanged the energetics of conformational fluctuations about the reference structure in the vicinity of the metabasin minimum. The values of the force constant is chosen as $k = 1.25\,\text{kcal/mol}\,\text{Å}^2$, based on initial test runs in which we explored values of these parameters that maintained helical and hairpin configurations slightly at higher temperatures. For the width of the flat-bottom well, we explore two values for $\delta\theta$, 15° and 45°. The scaling factor $\lambda_i$ is chosen to decay approximately exponentially with temperature from 1 to 0 as one moves from the lowest to the highest replica temperature.

At the end of the simulation, we use the weighted histogram analysis method (WHAM) to compute free energies at each replica [7]. We use the so-called binless implementation described by Kumar et al. [7] and also recently addressed by Shirts and Chodera [32]. The iterative equation determining the dimensionless free energy $\beta_j A_j$ for each replica $j$ is:

$$-\beta_i A_i = \ln\left[\sum_{j=1}^{J}\sum_{k=1}^{n} \mathrm{e}^{-\beta_i U(\mathbf{r}_{jk}) - \lambda_i\beta_i U_{\text{rest}}(\mathbf{r}_{jk})} \times \left(n\sum_{m=1}^{J} \mathrm{e}^{-\beta_m U(\mathbf{r}_{jk}) - \lambda_m\beta_m U_{\text{rest}}(\mathbf{r}_{jk}) + \beta_m A_m}\right)^{-1}\right], \tag{14}$$

where $n$ is the number of trajectory snapshots at each temperature, $J$ is the number of replicas and $\mathbf{r}_{jk}$ gives configuration $k$ (of $n$) for temperature $j$ (of $J$). The above equation must be iterated until the set of free energies $A_i$ converges, demanding that one of them has a fixed zero value. After convergence, we obtain the free energy difference between the low-temperature restrained replica and the high-temperature unrestrained reference state:

$$\Delta(\beta A) = \beta_1 A_1 - \beta_J A_J. \tag{15}$$

As the high-temperature state is identical between REMD simulations for different target configurations, the relative (temperature-normalised) free energies of different target configurations at the low temperature of interest are given by the values $\Delta(\beta A)$ from each respective restrained REMD run.

### *3.3 Temperature schedule optimisation*

One challenge that emerges in the application of varying torsional restraints is that the acceptance probabilities for neighbouring replicas in some parts of the cascade can become extremely small. That is, a bottleneck in replica space emerges due to changes in the scaling of the restraint energy as one moves between temperatures. To alleviate this problem, we modify the distribution of temperatures in the replicas to increase the average acceptance probabilities and abilities of replicas to traverse the entire temperature range. In recent years, a number of sophisticated techniques have been introduced to optimise the temperature intervals [33–37]. However, these can often require lengthy iterative simulations to reach an optimal temperature schedule. Here, instead, we pursue a simpler, more approximate approach that is sufficient to ensure good acceptance probabilities for the purposes of our free energy calculations.

Our approach to temperature optimisation is based on an analysis of average transition probabilities between pairs of replicas. Let $T_{ij}$ be the average probability that a system in temperature $i$ will swap to temperature $j$, with the normalisation condition $\Sigma_j T_{ij} = 1$. As replicas can only swap between neighbouring temperatures, this condition takes the form $T_{i,i-1} + T_{i,i} + T_{i,i+1} = 1$ for all but the low- and high-$T$ replicas. The set of transition probabilities takes the form of a matrix $\mathbf{T}$ that can be estimated after several rounds of swap attempts. The estimate follows

$$T_{ij} \propto \sum_k \wp^k_{\mathrm{acc},ij}, \quad T_{ii} \propto \sum_k \sum_j \left(1 - \wp^k_{\mathrm{acc},ij}\right), \qquad (16)$$

where the index $k$ tabulates all swap attempts for a replica pair $ij$ and $\wp_{\mathrm{acc}}$ is the computed acceptance probability (Equation (12)). The normalisation condition is used to determine the constant of proportionality.

The matrix $\mathbf{T}$ can be used to estimate the equilibrium flux of replicas between the low- and high-temperature bounds. Here, we follow the ideas of Katzgraber et al. [33] and compute the fractional population $f_i$, for each temperature $i$, of replicas that have most recently visited the low-$T$ vs. high-$T$ state. Unlike [33], we do not measure the actual flux of replicas, but use a ‘macroscopic’ description based on the average transition probabilities. While this approach certainly neglects correlations of replicas travelling between neighbouring temperatures, it provides a way to estimate a pseudo-optimal temperature schedule that avoids long computations of actual fluxes.

To compute $f_i$, we assume that the low-$T$ replica is a probability source and the high-$T$ one a sink. That is, we demand $f_i = 1$ and $f_i = 0$ for the two cases, respectively. In addition, we demand that the flux of replicas coming from the low-$T$ side at the highest temperature is zero. With these assumptions, the row vector $\mathbf{f}$ with elements $f_i$ can be computed by

$$\mathbf{f} = \lim_{n\to\infty} \mathbf{f}_0 \cdot (\mathbf{T}')^n, \qquad (17)$$

where $\mathbf{f}_0$ is zero except for its first element, which is 1, and $\mathbf{T}'$ is a modified version of $\mathbf{T}$ that has its last row (corresponding to the highest temperature, replica $m$) as zero except for $T'_{mm} = 1$. This modification of the transition probability matrix ensures the zero-flux condition. In practice, the limit need not be taken beyond a small exponent on the transition probability matrix, as the equilibrium probabilities converge rather quickly. This approach is essentially a practical way to find the highest eigenvalue eigenvector of $\mathbf{T}'$.

Once the $f_i$ are computed, these can be used to determine a new temperature distribution that maximises the flux of replicas between the temperature extremes. Following the suggestion of Katzgraber et al. [33], we select new temperatures in an attempt to make the change in $f_i$ linear in replica number. To do so, we construct a monotonic, spline-interpolated curve for $f_i$ vs. $T_i$ and successively pick new temperatures such that $f_{i+1} - f_i = (m-1)^{-1}$, where $m$ is the number of replicas.

Practically, the procedure of running a short REMD simulation, estimating $\mathbf{T}$ from it, and adjusting the temperature schedule must be iterated several times because the fractional populations $f_i$ are also affected by the temperature distribution. Our runs begin with the usual exponential temperature distribution and the process of temperature optimisation entails a series of 10 REMD iterations of 1 ns (per replica) each in duration. After the 10 ns total temperature optimisation period, we hold the temperatures fixed for the remainder of the run. In practice, we find that the optimisation converges in typically five to seven iterations, with changes to the temperatures less than 1% for further iterations.

## 4. Results and discussion

### *4.1 Reference equilibrium conformational ensemble*

To compute the conformational distribution of the MATα2 peptide to high accuracy, we first performed five independent REMD folding simulations of 100 ns each, starting from an extended structure. The use of long simulations and multiple trials is designed to facilitate accurate *equilibrium* structure populations. As such, the last 10 ns of each run was concatenated into an ‘equilibrium trajectory’ and clustered. The top six clusters, as shown in Figure 1, contain a variety of helical and hairpin structures. The dominant structure for this peptide and force field appears to be a helix, and the next three subdominant clusters entail partially unfolded versions of these. The fifth and sixth largest clusters contain hairpin structures that represent 4–5% of the configurational ensemble. In general, these six structures account for ~97% of the ensemble in total. The high population of helix is in agreement with the earlier results of Ikeda and Higo [8].

We took the computed configurational populations as a reference for later comparison with free energy calculations. To estimate the errors in these, we repeated the clustering for the aggregate last 50 ns of each of the five runs, for a total of 250 ns, and found that the populations did not change by more than 2%. The centroid structures of each cluster were then energy-minimised to provide a reference structure, labelled A–F for the six, in order of decreasing population. In addition to cluster population, we also investigated a trajectory population. The latter was designated the fraction of configurations from the aggregate trajectory that lied within a 2 Å RMSD distance from each reference structure. Certainly, a high correlation between the cluster and trajectory populations is expected (we find $R^2 = 0.998$). However, subtle differences do underlie these two approaches. Notably, the trajectory

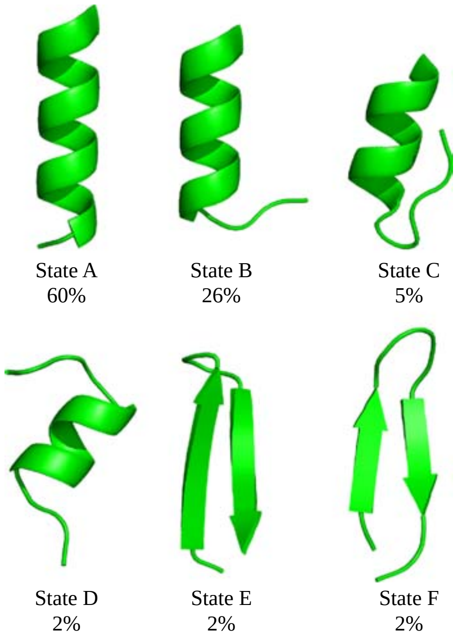

Figure 1. Conformational states found from REMD simulations. Percentages shown indicate the population of the configurational clusters extracted from 50 ns of simulation data accumulated from the last 10 ns each of the five independent 100 ns runs.

measure allows configurations to participate in more than one population, whereas the cluster version does not. Values for populations are shown in Table 1.

A question that immediately arises is whether or not potential energies predict the relative conformational populations. Table 1 and Figures 2 and 3 compare the minimised and averaged (at $T = 270$ K, through short MD simulations) potential energies of each cluster centroid with their cluster and trajectory populations. Importantly, these potential energies include the solvation free energy from the implicit model. As is clear in these results, the minimised potential energy completely fails to predict conformational population. The structures with the lowest energies are partially unfolded helical states, able to adopt more compact folds, but these structures represent small fractions of the overall ensemble. On the other hand, the average potential energy is able to discriminate the most populated cluster structure. However, for the non-dominant structures, this correlation fails, and allowing for a 2% error in the populations and several $k_{\mathrm{B}}T$ in the energies cannot account for this problem. Adding a harmonic free energy to the minimised potential energy, computed from the AMBER NMODE program, does not improve the correlation. This sum is also not able to capture the computed free energies described below, implying the existence of substantial non-harmonic contributions.

The failure of energetic measures to account for ensemble populations has several ramifications. First, it underscores the role of entropies in the determination of the stable structure. The conformational free energies, which relate directly to the equilibrium populations, include both a potential energy contribution and an entropic term. From these results, the failure of energies alone makes obvious that the folding entropies associated with each structure have substantial variations. These entropies likely stem from both backbone and side-chain fluctuations, although we do not attempt to make a distinction between the two here.

Second, these results strongly suggest that atomic physicochemical force fields such as the one used here cannot serve directly as structure prediction scoring functions, at least for peptides of this small size. The idea that the native structure lies at the global minimum of some interatomic scoring function has long been a perspective in

Table 1. Conformational states, populations and energies.

| State | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Cluster population (%) | 60 | 26 | 5 | 2 | 2 | 2 |
| Trajectory population (fraction within 2 Å basin) | 0.524 | 0.198 | 0.029 | 0.012 | 0.009 | 0.012 |
| Minimised potential energy | 0 | 6.08 | − 4.57 | − 10.84 | 7.20 | 17.56 |
| Average potential energy, $T = 270$ K | 0 | 8.94 | 3.17 | 6.53 | 18.60 | 26.77 |
| Harmonic free energy | 0 | − 2.26 | − 10.15 | − 5.47 | − 16.44 | − 25.16 |
| Minimised potential plus harmonic free energy | 0 | 3.82 | − 14.72 | − 16.31 | − 9.24 | − 7.60 |
| Free energy, $\delta\theta = 15°$ | 0 ± 0.36 | 3.27 ± 0.27 | 4.46 ± 0.14 | 7.04 ± 0.12 | 7.30 ± 0.29 | 6.06 ± 0.26 |
| Free energy, $\delta\theta = 45°$ | 0 ± 0.06 | 3.34 ± 0.10 | 5.43 ± 0.17 | 7.23 ± 0.38 | 6.47 ± 0.43 | 5.50 ± 0.59 |

Notes: All energies are expressed in units of $k_{\mathrm{B}}T$, with $T = 270$ K, and are normalised to zero for structure A. Standard errors in free energies are computed from time sequences of computed values for the final 30 ns of each simulation.

Table 2. Apparent unfolding temperature from heating simulations.

| State | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Temperature at which RMSD > 4.0 Å (K) | 373 | 345 | 322 | 287 | 347 | 347 |

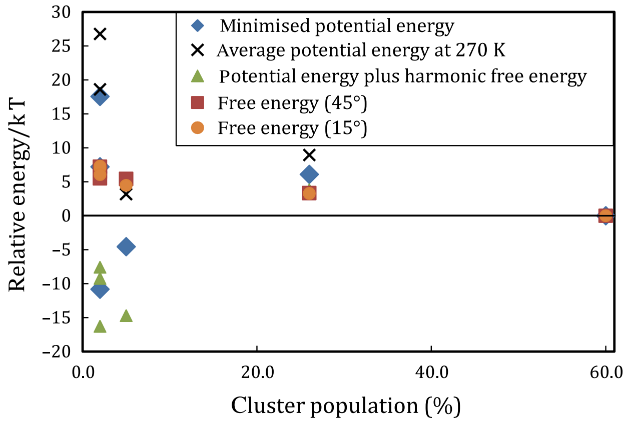

Figure 2. Energetic measures for six conformational states, as a function of cluster populations. Here, all energies are normalised to zero for the most populous cluster, state A. Energies are measured in units of $k_BT$.

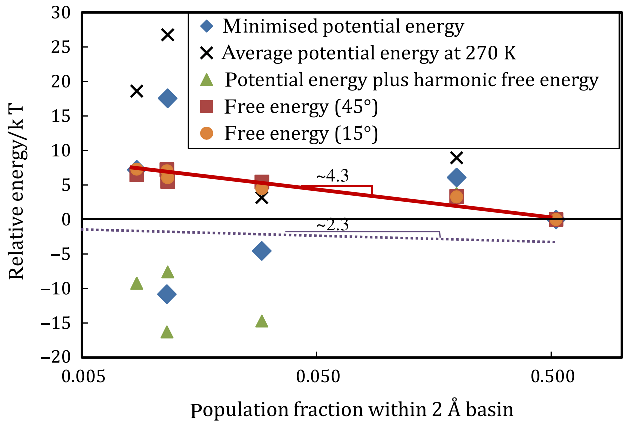

Figure 3. Energetic measures for six conformational states, as a function of cluster populations. Here, all energies are normalised to zero for the most populous cluster, state A. Energies are measured in units of $k_BT$. The comparison of the free energies to the theoretical expression $A^*/k_BT = -\ln \wp(\mathbf{r}^*) + \text{const}$ is shown by the dotted line.

the structure prediction community [38–41]. In particular, many bioinformatics-type efforts have parameterised such functions explicitly to attain this property. In contrast, these results emphasise that physicochemical force fields are *not* suitable for direct use as scoring functions. It is only by their coupling with canonical Boltzmann sampling and subsequent effective transformation to a free energy surface that the correct structure populations can be recovered. However, this does not rule out the possibility that appending *approximate* entropic terms to the interatomic force field might be effective in the development of a scoring function for minimisation.

### *4.2 Conformational free energies*

For each of the reference structures A–F, we used the restraint-based REMD approach described in Section 3 to compute their free energies, using 70 ns simulations with the final 30 ns for the WHAM analysis. Figures 2 and 3 show a comparison of the computed free energies with the cluster and trajectory populations. As expected, the free energies do an excellent job of rank-ordering the cluster and trajectory populations, particularly for the first four structures. The fact that the computed free energies capture qualitatively the trend in population between the structures is one important test that the method returns physically meaningful underlying free energies. Still, some differences in rank between the free energies and populations are found for the three lowest populated structures. For these structures, statistical errors may play a role. Both population measures are quite low and thus susceptible to greater error in their determination; moreover, the cluster and trajectory populations exhibit among themselves some slight discrepancies.

For all but the dominant two structures, there are discrepancies up to ~ 1 kcal/mol in computed free energies between the two restraint cases $\delta\theta = 15°$ and 45°. Keeping in mind that the $\delta\theta$ parameter measures the extent of the conformational basin for which the free energy is calculated, these differences likely reflect the existence of significant structural fluctuations that extend to the borders of the metabasin in the 15° case and that make non-negligible contributions to the free energy. When the size of the basin expands to the 45° case, permitting the sampling of greater structural fluctuations, more such border region is included in the total free energy. These differences are consistent with the secondary structures of states C–F. States C and D are partially folded helices that have significant unstructured cap regions, which are likely to retain some conformational flexibility. States E and F are hairpin structures that have significantly less intra-peptide hydrogen bonding than the helix and thus are susceptible to both bending motions and transient ‘unzipping’ at the termini.

Figure 4 shows the time dependence of the computed free energies, by way of subdividing the long REMD runs into 10 ns windows and performing WHAM calculations on each independently. In the initial 10 ns, the structures with the lowest minimised potential energy (C and D) show the lowest computed free energy. However, these estimates rapidly change as the REMD simulations each reach equilibrium by about 25 ns. Errors in the computed free energies are extracted from variations among the values for the final 30 ns of these curves. As shown in Table 1, the errors are lower for the more populated structures and nearly always less than 0.5 $k_BT$. The relative error with respect to the range of free energies of the structures is about 5–10%. This error suggests quite

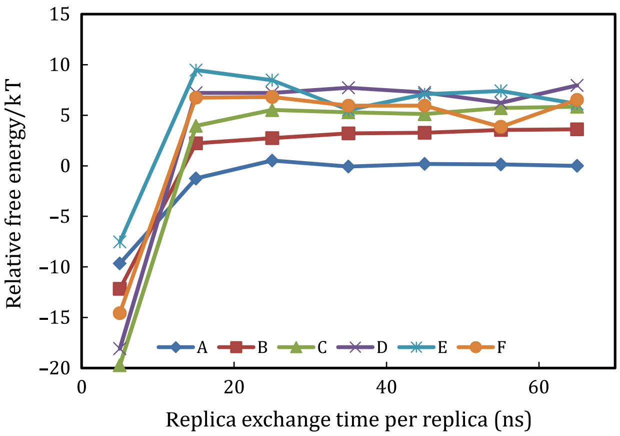

Figure 4. Time evolution of conformational free energies. Each point gives the computed free energies using the weighted histogram approach described in the text applied to a 10 ns window of REMD time around each time point. Results for the six conformations correspond to six separate 100 ns REMD runs using the restraint approach and the setting $\delta\theta = 45°$.

adequate performance of the method for determining relative free energies, and might be reduced by multiple or longer REMD runs.

To what extent do the computed free energies bear *quantitative* agreement with the populations? Equation (5) suggests that we should expect a linear relationship with the logarithm of trajectory population:

$$\frac{A^*}{k_{\mathrm{B}}T} = -\ln \wp(\mathbf{r}^*) + \text{const} \approx -2.30 \log_{10} \wp(\mathbf{r}^*) + \text{const}. \qquad (18)$$

In Figure 3, the computed slope of the dimensionless free energy vs. the base-10 logarithm of trajectory population for the first four structures is ∼4.3, a little less than twice what would be expected based on the theoretical expression. One possibility for this difference is that the use of an RMSD cut-off may not be an adequate measure of the cluster populations. On the one hand, the RMSD measure itself may be a poor way to partition configuration space around a basin, or should have structure-specific values. On the other hand, the cut-off value used may be too large and include configurations that should not be considered a part of each state's metabasin. To test this latter problem, we compared the computed free energies to trajectory populations determined from a series of increasingly stringent cut-offs, from 2 to 0.5 Å. Figure 5 shows the results of this test. With lower cut-offs, the trajectory populations for each state become increasingly smaller and thus susceptible to much increased error; consequently, in Figure 5, we do consider results for conformations with trajectory populations below 0.001. The data show that a linear fit of the free energies to the log of the populations becomes worse as the cut-off becomes more stringent, and is particularly difficult to assess at 0.5 Å, where only three structures admit populations above 0.001. However, a trend might still be inferred from these results in going from a cut-off of 2.0 to 1.0 Å, namely that as the cut-off becomes more stringent, the slope of the fit decreases and may approach that predicted by Equation (18) for more stringent cut-offs. A better approach here, which we leave to future work, might be to define the basin cut-off by the observed fluctuations in the restrained simulations, specific to each dihedral angle and each structure.

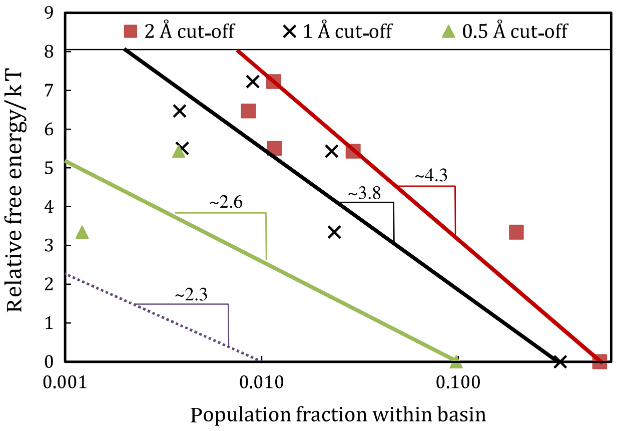

Figure 5. Relationship of free energies to populations in basins varying radii. The $x$-axis shows the fraction of the equilibrium trajectory configurations corresponding to conformations within a prescribed RMSD cut-off radius of the states shown in Figure 1. The three curves give the results for different cut-off radii, for $\delta\theta = 45°$; results for $\delta\theta = 15°$ are similar (not shown). The comparison of the free energies to the theoretical expression $A^*/k_{\mathrm{B}}T = -\ln \wp(\mathbf{r}^*) + \text{const}$ is shown by the dotted line.

The replica-exchange temperature schedule optimisation also reveals important characteristics of the conformational states. Figure 6 shows the final temperature

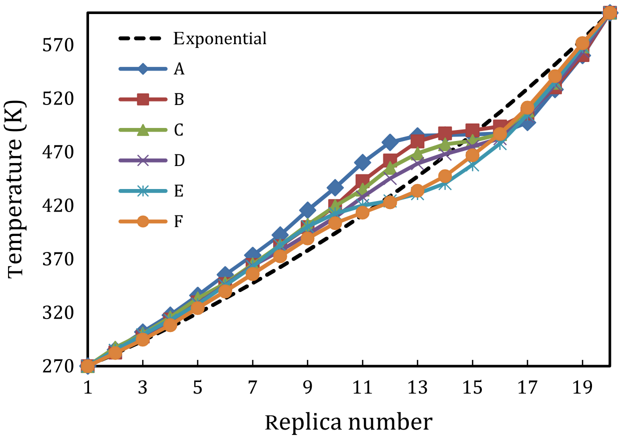

Figure 6. REMD temperature distributions in free energy calculations. Temperatures are updated based on acceptance probability statistics every 1 ns for the first 10 ns of each replica-exchange run. Final distributions are shown for the case $\delta\theta = 45°$; results for $\delta\theta = 15°$ are similar (not shown).

schedule for each of the structures, after the ten 1 ns updates based on the initial stages of the simulation. What is interesting about these results is that each state has a similar form that shows a ‘flattening’ of the schedule at intermediate replica numbers; above and below this flat region, the temperature schedule appears to follow the usual exponential distribution. Such behaviour indicates that the optimisation procedure places many replicas of different restraint strengths near the same temperature. This temperature corresponds to the point at which each structure unfolds (in the presence of restraints), and is the result of a balance between the increasing temperature and decreasing restraint strengths as one moves up the replica cascade. Remarkably, the degree of flattening almost exactly parallels the populations: the highest populated states have many replicas near the unfolding temperature while the least populous have a distribution that more closely resembles the purely exponential one. This is consistent with the fact that the free energy difference between the high-temperature, unrestrained reference state and the low-temperature, structure-specific one is greatest for the lowest free energy states and thus, to some extent, the most cooperative.

As an independent test of the computed free energies, we evaluated unfolding temperatures of each state A–F by non-equilibrium Langevin dynamics MD simulations (Figure 7, Table 2). Each run was seeded with the initial energy-minimised state at $T = 250$ K and used a heating rate of 3.5 K/ns until a temperature of 600 K was reached. The unfolding temperature was determined as the point when the RMSD of a system from its initial structure first exceeded 4 Å during this process. The various energetic and free-energetic measures in Figures 2 and 3 are plotted as a function of these unfolding temperatures. As before, none of the potential energy metrics exhibits anticorrelation with the folding temperature. In fact, the potential energy plus the harmonic free energy approximation erroneously shows the *opposite* trend, suggesting that lower free energy structures have lower unfolding temperatures and are thus *less* stable. On the other hand, the computed free energies from this approach show a reasonably good anticorrelated dependence with unfolding temperatures. For the two hairpin structures E and F, the unfolding temperature is somewhat higher than would be anticipated based on the overall relationship. This is likely either an artefact of the non-equilibrium nature of the unfolding simulations or the use of a common 4.0 Å RMSD cut-off for both alpha and beta structures in determining the unfolding temperature. In any case, the qualitative consistency of the unfolding temperatures and computed free energies provides further validation of the method.

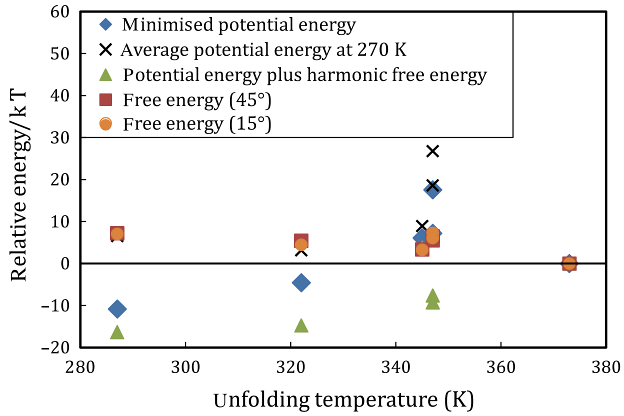

Figure 7. Apparent unfolding temperature vs. energetic measures. Unfolding temperatures for each conformation are determined from 100 ns Langevin dynamics simulations starting at $T = 250$ K using a heating rate of 3.5 K/ns until a temperature of 600 K is reached.

## 5. Conclusions

In this paper, we developed an enhanced sampling simulation method to compute the free energies associated with different peptide conformations. The method uses REMD, histogram reweighting and temperature schedule optimisation techniques to produce free energy estimates with statistical fluctuations of around 0.5 $k_B T$ or less. We applied this approach to a chameleon sequence displaying a variety of alpha and beta structures at equilibrium and found that the computed free energies were able to correctly rank the structure populations, while combinations of minimised and averaged potential energies were not, even when corrected with a harmonic free energy term. The computed free energies also showed a consistent relationship with unfolding temperatures for each structure, measured from non-equilibrium heating simulations. In total, these results suggest that the method is able to compute conformation free energies that underlie structural stability and populations.

The inability of potential energy metrics (even including the solvation free energy) to describe conformational populations is perhaps one of the most important ramifications of these results. Differences between the potential energy and the computed free energies show that entropic contributions to structural population and stability are significant here, likely stemming from a combination of backbone and side-chain structural fluctuations. Thus, physicochemical force fields such as the one used here are not directly useful as structure prediction scoring functions in the same way that database-parameterised scoring functions are, at the very least for peptides of this size. Instead, canonical sampling and inclusion of structural fluctuations due to conformational space regions surrounding potential energy minima is the key to the connection between force fields and structure stability.

The methodology developed here should be useful in a range of applications. The development and testing of molecular force fields may be aided by computing stabilities of various structures of test peptides that are well characterised experimentally, rather than by computing dominant structures alone. In particular, effects on the stability of a structure by changes to force-field parameters could be readily evaluated using this approach. In other areas, conformational free energies may shed light on driving forces and dissociation constants of natural and designed peptide ligands. They might also be able to elucidate mutational changes to local secondary structure stability in larger proteins, an issue that has been identified as relevant to aggregation propensity [5]. In principle, this approach might also be applied to larger protein systems, which would provide a method for mapping out folding free energy landscapes, although it remains for future work to characterise the computational requirements as the system size increases.

## Acknowledgements

The author gratefully acknowledges the support of the Camille and Henry Dreyfus Foundation and the National Science Foundation (Award No. CBET-0845074).

## References

[1] J.F. Swain and L.M. Gierasch, *The changing landscape of protein allostery*, Curr. Opin. Struct. Biol. 16 (2006), pp. 102–108.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, Garland Science, New York, 2007.

[3] R.W. Carrell and B. Gooptu, *Conformational changes and disease – Serpins, prions and Alzheimer's*, Curr. Opin. Struct. Biol. 8 (1998), pp. 799–809.

[4] D. Thirumalai, D.K. Klimov, and R.I. Dima, *Emerging ideas on the molecular basis of protein and peptide aggregation*, Curr. Opin. Struct. Biol. 13 (2003), pp. 1–14.

[5] K. Pagel, T. Vagt, and B. Koksch, *Directing the secondary structure of polypeptides at will: From helices to amyloids and back again?* Org. Biomol. Chem. 3 (2005), pp. 3843–3850.

[6] Y. Sugita and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding*, Chem. Phys. Lett. 314 (1999), pp. 141–151.

[7] S. Kumar, D. Bouzida, R.H. Swendsen, P.A. Kollman, and J.M. Rosenberg, *The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method*, J. Comput. Chem. 13 (1992), pp. 1011–1021.

[8] K. Ikeda and J. Higo, *Free-energy landscape of a chameleon sequence in explicit water and its inherent alpha/beta bifacial property*, Protein Sci. 12 (2003), pp. 2542–2548.

[9] D. Eisenberg and A.D. McLachlan, *Solvation energy in protein folding and binding*, Nature 319 (1986), pp. 199–203.

[10] W.C. Still, A. Tempczyk, R.C. Hawley, and T. Hendrickson, *Semianalytical treatment of solvation for molecular mechanics and dynamics*, J. Am. Chem. Soc. 112 (1990), pp. 6127–6129.

[11] P.G. Debenedetti, *Metastable Liquids: Concepts and Principles*, Princeton University Press, Princeton, NJ, 1996.

[12] F.H. Stillinger, *A topographic view of supercooled liquids and glass-formation*, Science 267 (1995), pp. 1935–1939.

[13] T.L. Hill, *An Introduction to Statistical Thermodynamics*, Addison-Wesley, Reading, MA, 1960.

[14] D.A. Case, *Normal mode analysis of protein dynamics*, Curr. Opin. Struct. Biol. 4 (1994), pp. 285–290.

[15] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman, *AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules*, Comput. Phys. Commun. 91 (1995), pp. 1–41.

[16] D.A. Case, T.E. Cheatham, III, T. Darden, H. Gohlke, R. Luo, K.M. Merz, Jr, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods, *The Amber biomolecular simulation programs*, J. Comput. Chem. 26 (2005), pp. 1668–1688.

[17] P.A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, and A. Pohorille, *The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of* ab initio *calculations and experimental data*, Comput. Sim. Biomol. Sys. 3 (1997), pp. 83–96.

[18] A. Onufriev, D. Bashford, and D.A. Case, *Modification of the generalized Born model suitable for macromolecules*, J. Phys. Chem. B 104 (2000), pp. 3261–3429.

[19] M.S. Shell, R. Ritterson, and K.A. Dill, *A test on peptide stability of AMBER force fields with implicit solvation*, J. Phys. Chem. B 112 (2008), pp. 6878–6886.

[20] E. Lin and M.S. Shell, *Convergence and heterogeneity in peptide folding with replica exchange molecular dynamics*, J. Chem. Theory Comput. 5 (2009), pp. 2062–2073.

[21] S.B. Ozkan, G.H.A. Wu, J.D. Chodera, and K.A. Dill, *Protein folding by zipping and assembly*, Proc. Natl Acad. Sci. USA 104 (2007), pp. 11987–11992.

[22] M.S. Shell, S.B. Ozkan, V.A. Voelz, G.H.A. Wu, and K. Dill, *Blind test of physics-based prediction of protein structures*, Biophys. J. 96 (2009), pp. 917–924.

[23] S. Ono, N. Nakajima, J. Higo, and H. Nakamura, *Peptide free-energy profile is strongly dependent on the force field: Comparison of C96 and AMBER95*, J. Comput. Chem. 21 (2000), pp. 748–762.

[24] A.E. García and K.Y. Sanbonmatsu, *Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds*, Proc. Natl Acad. Sci. USA 99 (2002), pp. 2782–2787.

[25] S. Gnanakaran and A.E. Garcia, *Folding of a highly conserved diverging turn motif from the SH3 domain*, Biophys. J. 84 (2003), pp. 1548–1562.

[26] Y. Mu, D.S. Kosov, and G. Stock, *Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments*, J. Phys. Chem. B 107 (2003), pp. 5064–5073.

[27] R. Zhou, *Free energy landscape of protein folding in water: Explicit vs. implicit solvent*, Proteins 53 (2003), pp. 148–161.

[28] T. Yoda, Y. Sugita, and Y. Okamoto, *Comparisons of force fields for proteins by generalized-ensemble simulations*, Chem. Phys. Lett. 386 (2004), pp. 460–467.

[29] T. Yoda, Y. Sugita, and Y. Okamoto, *Secondary structure preferences of force fields for proteins evaluated by generalized-ensemble simulations*, Chem. Phys. 307 (2004), pp. 269–283.

[30] R. Geney, M. Layten, R. Gomperts, V. Hornak, and C. Simmerling, *Investigation of salt bridge stability in a generalized Born solvent model*, J. Chem. Theory Comput. 2 (2006), pp. 115–127.

[31] V.A. Voelz, M.S. Shell, and K.A. Dill, *Predicting peptide structures in native proteins from physical simulations of fragments*, PLoS Comput. Biol. 5 (2009), e1000281.

[32] M.R. Shirts and J.D. Chodera, *Statistically optimal analysis of samples from multiple equilibrium states*, J. Chem. Phys. 129 (2008), pp. 124105–124110.

[33] H.G. Katzgraber, S. Trebst, D.A. Huse, and M. Troyer, *Feedback-optimized parallel tempering Monte Carlo*, J. Stat. Mech.: Theory Exp. (2006), P03018.

[34] S. Trebst, M. Troyer, and U.H.E. Hansmann, *Optimized parallel tempering simulations of proteins*, J. Chem. Phys. 124 (2006), pp. 174903–174906.

[35] D. Gront and A. Kolinski, *Efficient scheme for optimization of parallel tempering Monte Carlo method*, J. Phys.: Condens. Matter. (2007), 036225.

[36] H. Kamberaj and A.v.d. Vaart, *An optimized replica exchange molecular dynamics method*, J. Chem. Phys. 130 (2009), 074906.

[37] R. Denschlag, M. Lingenheil, and P. Tavan, *Optimal temperature ladders in replica exchange simulations*, Chem. Phys. Lett. 473 (2009), pp. 193–195.

[38] P. Koehl and M. Levitt, *A brighter future for protein structure prediction*, Nat. Struct. Biol. 6 (1999), pp. 108–111.

[39] D. Baker and A. Sali, *Protein structure prediction and structural genomics*, Science 294 (2001), pp. 93–96.

[40] C. Venclovas, A. Zemla, K. Fidelis, and J. Moult, *Assessment of progress over the CASP experiments*, Proteins Struct. Funct. Genet. 53 (2003), pp. 585–595.

[41] J. Moult, *A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction*, Curr. Opin. Struct. Biol. 15 (2005), pp. 285–289.