Balancing Accuracy and Cost of Confinement Simulations by Interpolation and Extrapolation of Confinement Energies

François Villemot,[1] Riccardo Capelli,[2,3] Giorgio Colombo,[2] and Arjan van der Vaart[1,]*

[1] Department of Chemistry, University of South Florida, 4202 East Fowler Avenue, CHE 205, Tampa, Florida, 33620, U.S.A.

[2] Istituto di Chimica del Riconoscimento Molecolare, Consiglio Nazionale delle Ricerche, Via Mario Bianco 9, 20131 Milano, Italy

[3] Dipartimento di Fisica, Università degli Studi di Milano and INFN, via Celoria 16, 20133 Milano, Italy

Abstract

Improvements to the confinement method for the calculation of conformational free energy differences are presented. By taking advantage of phase space overlap between simulations at different frequencies, significant gains in accuracy and speed are reached. The optimal frequency spacing for the simulations is obtained from extrapolations of the confinement energy, and relaxation time analysis is used to determine time steps, simulation lengths, and friction coefficients. At post-processing, interpolation of confinement energies is used to significantly reduce discretization errors in the calculation of conformational free energies. The efficiency of this protocol is illustrated by applications to alanine *n*-peptides and lactoferricin. For the alanine-

n peptide errors were reduced between 2 and 10 fold and sampling times between 8 and 67 fold, while for lactoferricin the long sampling times at low frequencies were reduced 10-100 fold.

* Corresponding author. Email: avandervaart@usf.edu. Phone: +1-813-974-8762.

Introduction

Conformational free energy differences are of considerable importance to protein science, since they can be used to rationalize and predict the relative stabilities of different protein folds, to elucidate the mechanism of allostery and conformational transitions, and to explain the effect of mutations on these processes. Obtaining conformational free energies by computational methods is no easy task, however.[1-8] Most methods require a pathway in configuration space that connects the one state of interest to the other. Pathways with physiological relevance, that is, with high statistical weights and low free energy barriers, are hard to determine due to the enormous dimensionality of configurational space. Moreover, while many methods use a lower-dimensional order parameters to describe pathways, it is generally difficult to predict and verify if these order parameters are sufficient and correct.[9-10] Other pathways, like interpolated pathways, suffer from the occurrence of large free energy barriers, which impede sampling.[11-12]

The difficulties of calculating pathways and associated free energy profiles in configurational space can be circumvented by the use of nonphysical pathways to connect the states of interest. This principle is exemplified by the confinement method,[13-19] which connects the states of interest by artificial, uncoupled harmonic oscillator states. In the confinement method, the states of interest are slowly transformed into independent harmonic oscillators using harmonic restraints of increasing strength. Since the free energy cost of applying these restraints can be readily calculated, and the free energy of the harmonic oscillator is known, conformational free energy differences can be obtained in a relatively straight-forward manner, especially when using a modified harmonic oscillator state that is invariant to rigid body motions.[13-14] This calculation is normally carried out in an implicit solvent, but the method has recently been extended to include explicit solvation.[15]

By foregoing a path in configurational space, the confinement method provides an efficient and robust way to calculate conformational free energy differences, even for states that are highly dissimilar in structure. But the method has other computational benefits that can be exploited. The free energy of transforming the system to a set of independent harmonic oscillators is obtained through a series of restrained simulations, each with a different strength of the harmonic

restraint. While the strengths differ, the center of the restraint is the same in all of these simulations, and corresponds to the equilibrium structure of interest. This means that spatial overlap between the configurational space of the simulations is guaranteed, and that overlap can be particularly large for "neighboring" simulations for which the restraint force constants are closest in value. This overlap is illustrated in Fig. 1. The overlap in configurational space is closely associated with the overlap of energy distributions, which is crucial for the accurate estimate of thermodynamic properties.[20]

-- Figure 1 Here --

Here we exploit this overlap in order to maximize the efficiency and minimize the statistical error of confinement simulations. We will show that one can use the overlap in configurational space to accurately predict confinement energies at unsampled strengths, and that this interpolation significantly decreases the error in calculated free energies. We will also show that instead of sampling at given intervals, one can use the overlap to predict at which restraining strength to sample next for simultaneously optimal errors and costs. Finally, by coupling these interpolations and extrapolations to relaxation time analyses, we will introduce a robust protocol for optimal confinement simulations, which is illustrated by applications to the alanine *n*-peptide and lactoferricin.

Methods

Confinement method. The confinement method[13-19] aims to compute the free energy of macrostate Ω by transforming the system of interest with $3N$ degrees of freedom into a set of $3N$ non-interacting harmonic oscillators (HOs). This transformation is performed in a series of simulations by adding harmonic restraints of increasing strength to the physical potential. The restraints are centered at the positions of a reference structure X_0 belonging to Ω, and the simulations are performed until the system can be considered purely harmonic. A recent improvement,[13-14] also used here, removes the rigid-body motions by performing a mass-weighted best-fit alignment of the system onto X_0 at each simulation step. Removing the sampling of the translational and rotational degrees of freedom significantly speeds up

convergence, especially at low restraint strengths. The total number of degrees of freedom (N_{DOF}) then becomes $3N-5$ for systems with linear geometries and $3N-6$ otherwise. The vibrational free energy of the Ω macrostate can subsequently be computed by thermodynamic integration:[14]

$$G_\Omega = E(X_0) + \frac{N_{DOF}}{\beta}\log(\beta h\nu) - 2(\pi\nu)^2 M \int_0^1 \left\langle \rho_m^2(X,X_0)\right\rangle_{H_\lambda(\lambda)} d\lambda. \quad (1)$$

Here, β is the inverse temperature ($\beta = 1/k_bT$), h is Planck's constant, M is the total mass of the system, $\rho_m(X,X_0)$ is the mass-weighted root-mean-square distance (rmsd) of sampled configuration X with reference structure X_0, $E(X_0)$ is the potential energy of the reference structure, and $\langle . \rangle$ denotes an ensemble average. The Hamiltonian $H_\lambda(\lambda)$ depends on the parameter λ, which is used to switch from an unrestrained system ($\lambda = 0$) to a system with harmonic restraints of frequency ν ($\lambda = 1$). This frequency must be chosen high enough, so that at $\lambda = 1$ virtually all of the system's energy is due to the restraints. At that point, the system can be considered to be a set of non-interacting harmonic oscillators. For convenience, the integration in Eq. 1 can be transformed by the change of variable $\zeta = \lambda\nu^2$ and $H_\zeta(X;\zeta) = H_\lambda(X;\lambda\nu^2)$, so that the integration is done in frequency space. In addition, the mass-weighted rmsd can be expressed as a function of the confinement energy $U_{conf} = 2M\pi^2\nu^2\left\langle \rho_m^2(X,X_0)\right\rangle_{H_\zeta(\nu^2)}$:

$$G_\Omega = E(X_0) + \frac{N_{DOF}}{\beta}\log(\beta h\nu) - \int_0^{\nu^2} \frac{U_{conf}}{\nu^2} d\zeta. \quad (2)$$

For a harmonic oscillator, $\left\langle \rho^2(X,X_0)\right\rangle \propto \nu^{-2}$. A trapezoidal rule for numerical integration of Eq. 2, which interpolates linearly between successive points, would therefore be a bad choice. Since this relation becomes linear in logarithmic space, the integration is carried out using linear interpolation in logarithmic space instead;[21] in the following we will refer to the integrand as:

$$I_\nu \equiv \frac{U_{conf}}{\nu^2}. \quad (3)$$

In practice, simulations are carried out with increasing values of ν until the kinetic energy of the system ($N_{DOF}k_bT/2$) equals the confinement energy U_{conf} , as expected for a purely harmonic system, or equivalently:

$$\nu^2 \left\langle \rho_m^2(X,X_0) \right\rangle_{H_\zeta(\nu^2)} = \frac{N_{DOF}}{(2\pi)^2 \beta M}. \tag{4}$$

The averages $\left\langle \rho_m^2(X,X_0) \right\rangle_{H_\zeta(\nu^2)}$ are obtained from independent simulations that generate representative configurations at the given restraint frequencies.

Conformational free energies. By removing rigid body motions in the confinement procedure,[14] G_Ω of Eq. 1 and 2 represents a vibrational configurational free energy that lacks free energy contributions from the overall translational and rotational motions. The translational free energy can be obtained from the partition function of the ideal gas, while the rotational component can be obtained from the partition function of the rigid rotor.[22-23] In calculating conformational free energy differences (ΔG) between two states, the translational components will cancel, but rotational components generally won't, because of differences in the moments of inertia. Here, all reported ΔG values include the rotational component, while ΔG_Ω is used to denote a difference in vibrational free energy.

Reweighting and interpolation. Since the system is confined to the vicinity of the same reference structure X_0 in each simulation, there is large spatial overlap between these sets of configurations (Fig. 1). This means that the configurations obtained at a given frequency can be used to estimate ensemble averages at a different frequency. Consider N_i configurations obtained from a simulation with restraint frequency ν_i . The ensemble average of observable A at frequency ν_j is given by:

$$\langle A \rangle_{\nu_j} = \left\langle A e^{-\beta(U_j - U_i)} \right\rangle_{\nu_i} e^{-\beta \Delta F_{ij}}. \tag{5}$$

Here $\Delta F_{ij} = F_i - F_j$ is the free energy difference between the biased states at ν_i and ν_j; we use the symbol F to distinguish it from the configurational free energy of the unbiased state (G) of Eq. 1 and 2. U_j and U_i are the potential energy values corresponding to frequencies ν_j and ν_i, respectively. Since only the restraint differs between these potentials:

$$U_j(X_k) - U_i(X_k) = 2M\pi^2\rho_m^2(X_k, X_0)\nu_i^2\left[\left(\frac{\nu_j}{\nu_i}\right)^2 - 1\right]. \quad (6)$$

The accuracy of the ensemble average obtained by reweighting (Eq. 5 and 6) diminishes with increasing $|\nu_j - \nu_i|$. A more accurate estimation can be obtained by using additional statistics. This can be done by combining configurations obtained from all simulations at all frequencies. To combine samples from these multiple simulations, an estimation of the free energy difference between the states is needed, which can be obtained from the multistate Bennett acceptance ratio estimator (MBAR).[24]

With this reweighting we can interpolate averages between simulated frequencies, thereby increasing the accuracy of the thermodynamic integration of Eq. 2 in a cost-effective way. We also use it to extrapolate the value of the confinement energy for frequencies higher than the highest simulated thus far. As described below, this extrapolation is used to assess the frequency of the next simulation such that the overall cost and accuracy of the procedure is optimized. Finally, the reweighting also increases the accuracy at the simulated frequencies, by mixing in configurations from the other simulations in the computation of the confinement energy.

Extrapolation. In order to properly estimate the error on the extrapolated value, the following setup was used: we start from a set of simulations performed with different restraints, up to a frequency ν_{max}. We subsequently employ the extrapolation of Eq. 4 and 5 to estimate the confinement energy for a set of unsampled frequencies $\nu_i > \nu_{max}$. The computational cost of this extrapolation is low (much lower than the actual sampling), and nearly independent of the number of unsampled frequencies. We chose this direction, since in extrapolating towards higher frequencies, phase space is compressed. This means that all relevant areas of space for the

higher frequency restraint were sampled in the lower frequency simulation (but insufficiently). If we were to choose the other direction, that is, sampling at the higher frequency followed by extrapolation to the lower frequency, certain regions of space important for the low frequency restraint would be left unsampled. After the extrapolation, a confinement simulation is performed for each of the new frequencies in order to obtain the actual value of the confinement energy, and these calculated values are compared with those obtained from the extrapolation. The ratio between the extrapolated and actual confinement energy is a measure of the error of the extrapolation. We express this ratio as a function of the free energy difference ΔF between the simulations at ν_i and ν_{max}. ΔF can be obtained from Eq. 4 and 5, or, if simulations at multiple frequencies are used, from MBAR. The latter approach would yield somewhat more accurate extrapolations, since more data is used. However, here we used data from only one simulation and the former approach, in order to base all comparisons on the same amount of data. For a given ν_{max}, ΔF increases with $|\nu_{max} - \nu_i|$, and represents a meaningful quantity that can be compared across systems of different sizes.

Correlation times. The efficiency and accuracy can be improved further by considering the correlation time of the system. This correlation time is affected by the addition of harmonic restraints, especially at high frequencies, when the confinement energy accounts for a large portion of the total potential energy. In addition, these restraints limit the configurations accessible to the system to the ones close to the reference structure X_0, and the phase space to sample gets smaller as the frequency gets higher. In order to attain comparable sampling for each frequency, different sampling times are therefore needed, which can be estimated from the correlation time. These were estimated by block-averaging the confinement energies,[25] and also by calculating the auto-correlation function of the confinement energy.

Alanine *n*-peptide setup. We performed confinement simulations of capped alanine *n*-peptides (*n*=2, 4, 6, 8, and 10), with the general formula $CH_3CO\text{-}Ala_{n-1}\text{-}NHCH_3$. These simulations were carried out with the CHARMM program,[26] using the CHARMM polar hydrogen parameter set param19,[27] and the ACE implicit solvent model.[28] The two lowest-energy conformations of the alanine dipeptide are C_{7ax} and C_{7eq}, which for the force-field and implicit solvent method used, correspond to backbone dihedral angles of (ϕ, ψ) = (61.4, -71.4) and (-78.0, 138.7) degrees,

respectively. The C_{7ax} and C_{7eq} conformations were used as the reference structures for the alanine dipeptide. Larger alanine *n*-peptide systems behave nearly like independently linked alanine dipeptides when the peptides are in the C_{7ax} or C_{7eq} states.[12, 29] For instance, the energy minima (C_{7ax},C_{7ax}) and (C_{7eq},C_{7eq}) of the alanine tripeptide correspond to (ϕ_1, ψ_1, ϕ_2, ψ_2) = (61.1, -72.1, 59.6, -71.6) and (-77.4, 137.7, -76.7, 137.8) degrees, respectively. For all $n \geq 3$ alanine systems, the confinement reference structures were obtained by setting all the (ϕ,ψ) dihedral angles to the values of C_{7ax} and C_{7eq} of the alanine dipeptide, and performing an energy minimization. In the following, these configurations are simply named C_{7ax} and C_{7eq}, independently of the number of dihedral angles. The moments of inertia were obtained for the reference structures. The corresponding contribution to the free energy difference between C_{7ax} and C_{7eq} equals $\Delta G_{rotation} = \frac{k_b T}{2} \ln\left(\frac{I_{C_{7ax}}}{I_{C_{7eq}}}\right)$, where $I_{C_{7ax}}$ and $I_{C_{7eq}}$ represent the products of the three principal moments of inertia of C_{7ax} and C_{7eq}, respectively.[22-23]

All confinement simulations were performed using Langevin dynamics at 300 K, with friction coefficients of 1, 5, 10, or 20 ps^{-1} (see Results section). The time step had a maximum value of 1 fs, and was adjusted depending upon the restraint frequency. It was chosen so there are at least 80 time steps per harmonic oscillator period, resulting in smaller time steps for higher frequencies. Different time steps were tested, which showed that at least 40 steps per period are required to obtain accurate estimation of the confinement energy. A conservative value of 80 steps/period was then chosen. SHAKE[30] was not used in the simulations. To further restrict sampling to the state of interest, especially at the lowest frequencies, we also added flat-bottom dihedral restraints. These were centered on the energy-minimized values, with a force constant of 10 kcal/mol/rad^2, and a width of 2.5°. This value was chosen so that the states of interest are the same as the ones defined in the umbrella sampling. Interpolation of the confinement energy was done for 10 frequencies, equally spaced in log-space, between consecutive simulations. Adding more points didn't change the final free energy difference or the error bars.

For comparison, free energy differences between the C_{7ax} and C_{7eq} conformations were also obtained from one-dimensional umbrella sampling simulations.[31] For the alanine *n*-peptide, the transformation from C_{7eq} to C_{7ax} involves (*n*-1) ϕ and (*n*-1) ψ dihedral transformations. These angles were treated as reaction coordinates, and changed one at a time (in the order $\phi_1,\psi_1,\ldots,\phi_{n-1},\psi_{n-1}$), while keeping the others constant. For each dihedral angle, 50 equally spaced umbrella windows were used, with a force constant of 150 kcal/mol/rad^2, and a simulation time of 100 ns per window. To maintain the *trans* peptide configuration, flat-bottom dihedral restraining potentials were used for the ω backbone dihedral angles with a force constant of 10 kcal/mol/rad^2 and a width of 90°. Simulations were performed with Langevin dynamics at 300 K, using a 1 fs time step, no SHAKE,[30] param19[27] and ACE.[28] Potentials of mean force (PMF) were obtained from MBAR;[24] all free energies are reported in kcal/mol.

Lactoferricin setup. Bovine lactoferricin is 25-residue peptide cleaved from lactoferrin with anti-microbial properties.[32] In lactoferrin, the sequence is folded into an α-helix followed by a β-strand, while the cleaved peptide adopts a β-hairpin fold; the peptide contains one disulfide bond (Fig. 2).[33-34] No spontaneous conformational transitions were observed in long unbiased MD simulations.[35] Because of its size and the complexity of the transition, lactoferricin is a good test system for the confinement method, and representative of the more challenging biological systems that are the ultimate target for the method.

The α+β conformation was obtained from residues 17-41 of lactoferrin (PDB: 1BLF[34]), while the β-hairpin conformation was taken from lactoferricin in solution (PDB: 1LFC[33]). We used the CHARMM36 force field[36] with the GBMV implicit solvent model,[37] Langevin dynamics and no SHAKE.[30] A friction coefficient of 1 ps^{-1} was used for simulations with a frequency lower than 2 ps^{-1}, and a friction coefficient of 20 ps^{-1} for frequencies above. Interpolation of the confinement energy was done for 10 equally log-spaced frequencies between consecutive simulations. After an energy minimization, each conformation of the peptide was heated and equilibrated at 300 K. The reference structures used in the confinement simulations were obtained from rmsd-based clustering with a cut off of 3.5 Å of a 25 ns unrestrained trajectory. These trajectories were also used to obtain the principal moments of inertia. All simulations were conducted with the CHARMM program.[26]

-- Figure 2 Here --

Results

Alanine *n*-peptide. Fig. 3 shows the ratio between the extrapolated and actual confinement energies as a function of ΔF, for multiple values of ν_{max}. Curves for all alanine systems are provided; a value of one indicates that the extrapolation perfectly predicted the confinement energy. As expected, deviations from one strongly increased with the free energy difference, and the ratio was very close to one for small ΔF. A notable feature is that the extrapolation stayed accurate for larger free energy differences as ν_{max} increases. In other words, as the system becomes more harmonic, it becomes easier to predict the result of a new simulation. This is due to the fact that at larger frequencies, the harmonic restraints represent a larger portion of the total energy, and the configurational space is compactly distributed around X_0 in a predictable manner. In addition, we observed that the extrapolation is more accurate as the size of the system increases. Larger systems have narrower energy distributions, so that the weights in Eq. 5 are closer to one another. More configurations will therefore contribute significantly to the ensemble average at another frequency, thus lowering the error. This is a particularly encouraging feature, which will facilitate the application of the confinement method to larger and more complex systems.

-- Figure 3 Here --

The information of Fig. 3 can be used to extract the maximum value of ΔF for which the extrapolation error is below a desired threshold. We chose 5%, a fairly conservative value for this error, which corresponds to an extrapolated/actual confinement energy ratio of 0.95 or 1.05. In the following we will refer to this spacing as ΔF_{ext}. Fig. 4 shows ΔF_{ext} as a function of frequency for the alanine *n*-peptides. While the curves are bumpy (due to the fact that the ratios switched between 0.95 and 1.05, discretization of *v*, and finite sampling) the graph shows several

clear trends: consistent with the results of Fig. 3, ΔF_{ext} increased with both frequency and system size.

-- Figure 4 Here --

The physical relevance of ΔF_{ext} is the following. When the free energy difference between the sampled system at ν_{max} and the unsampled system at higher frequency is ΔF_{ext}, there is sufficient overlap in distribution functions to estimate, within some preselected error bound (here 5%), the confinement energy at the unsampled frequency from simulated data at ν_{max}. This means that after sampling at both frequencies, there will be sufficient overlap in distribution functions to accurately calculate confinement energies at frequencies inbetween. The accuracy of this interpolation will be higher than the accuracy of the extrapolation, since more data is available for the interpolation (one extra set of simulations). Furthermore, the error of interpolation can be reduced further by taking into account all simulated data, at all simulated frequencies. As shown below, this interpolation significantly reduced the overall error in calculating the configurational free energies. Thus, we propose to exploit ΔF_{ext} as guideline for selecting the frequency spacing of the simulations. The goal of this procedure is to pick the maximum spacing at which high quality interpolations remain feasible, thereby obtaining high accuracy at minimal computational costs. If we have a set of simulations up to frequency ν_{max}, the next frequency of simulation will be picked such that its free energy difference with the ν_{max} simulation is ΔF_{ext}.

Fig. 5 shows that interpolation can be performed to obtain confinement energies at non-simulated frequencies. In Fig. 5A, the value of I_v (Eq. 3) is shown in black for the alanine dipeptide at simulated frequencies of 0.021 and 6.6 ps^{-1}. The free energy difference between these simulations was 4.2 kcal/mol. The black line represents the value of I_v that would vary linearly with the logarithm of the frequency, which is the assumption made when performing the integration of Eq. 2 in log space,[18] and also the analytical solution for harmonic oscillators. The red curve corresponds to interpolated values using MBAR. Additional simulations at intermediate frequencies confirm the accuracy of the interpolation. The simulated values (blue dots) show that I_v does not follow a straight line in log space, but falls on the interpolated curve instead. The observed non-log-linear behavior is expected for this frequency range, since the

system is far from being purely harmonic. In fact, the harmonic terms contribute only 24% of the total energy at a frequency of 6.6 ps^{-1}. The interpolation accurately reproduced the observed behavior, which demonstrates that meaningful information about the system can be obtained through interpolation. It also shows how the interpolation can greatly increase the efficiency of the method: while all simulations (represented by black and blue symbols) would be needed to accurately compute the free energy difference over that frequency range, just two initial simulations (black point) are sufficient if the interpolation is used. Greater accuracy can also be achieved by reducing the discretization error arising from the frequency spacing, but this comes at additional computational costs. Fig. 5B and 5C illustrate how the interpolation for the alanine dipeptide performs at higher frequencies. The free energy differences between these two frequencies are comparable to the one corresponding to Fig. 5A (5.9 and 4.2 kcal/mol versus 4.2 kcal/mol). Again, the interpolation correctly estimated I_v for frequencies that were not simulated. At higher frequencies I_v varied more linearly with the log of the frequency, as expected for more harmonic system. The same behavior was observed for the other alanine systems, such as alanine 10-peptide (Fig. 5D).

-- Figure 5 Here --

Fig. 6 shows the correlation time of the confinement energy for the alanine dipeptide and decapeptide, as a function of the restraint frequency. The correlation times were calculated by block-averaging[25] (indicated by circles) and from the autocorrelation function of the confinement energy (triangles). The two methods gave similar results, which indicates that the correlation time could be properly estimated. For the dipeptide, the correlation time was similar for all frequencies < 0.2 ps^{-1}, and irrespective of the friction coefficient, while for the decapaptide higher friction coefficients led to higher correlation times in this frequency range. This is likely due to the more complex landscape of the decapeptide, which has subbasins; visiting the various subbasins is hindered by large friction terms. Near a frequency of 0.2 ps^{-1} the correlation times dropped significantly for all systems. At this frequency, U_{conf} represents between 2 and 4% of the kinetic energy. Apparently, this energy is sufficient to limit the system to one subbasin, which explains the precipitous decline in correlation time. At high frequencies, low correlation times were observed, inversely proportional to the friction coefficient. We checked that the

average values of the confinement energies were not affected by the friction coefficient, so that lower friction coefficients indeed led to faster sampling. The simulation time needed to obtain a given number of independent measurements is proportional to the correlation time. Fig. 6 shows that in order to obtain uniform sampling across the frequencies, much smaller simulation times are needed at higher frequencies. In addition, by increasing the friction coefficient at high frequency, the simulation time can be reduced, thereby further increasing the efficiency of the calculation.

-- Figure 6 Here --

In order to demonstrate the increased performance of the confinement method through extrapolation, interpolation and assessment of correlation times, we computed the free energy difference between the C_{7ax} and C_{7eq} conformations of the alanine *n*-pepetides. For each frequency, the correlation time of the confinement energy was estimated at regular time intervals, and the simulation was stopped when the number of independent measurements (which is the simulation time divided by the correlation time) was at least 1000. The confinement simulations were run in an iterative manner. The first simulation was performed at a frequency $\nu_1 = 0.02$ ps^{-1}. The frequency of the next simulation was calculated by extrapolation. Several strategies were employed. The first used a constant free energy spacing of 5 kcal/mol. The other three strategies used the information of Fig. 4 to vary this spacing as a function of *v*. In the second strategy the spacing was system-dependent, and obtained from a log-linear bestfit of ΔF_{ext} to *v*. The third strategy was system-independent, and given by $\Delta F_{ext} = 3 + 0.388 \ln(\nu / \nu_1)$ (indicated by the lower dashed line in Fig. 4), a conservative estimate of $\Delta F_{ext}(\nu)$. In the last strategy, a more aggressive estimate was chosen (indicated by the upper dashed line in Fig. 4): $\Delta F_{ext} = 5 + 0.485 \ln(\nu / \nu_1)$. This iterative process of extrapolation and simulation was repeated until the convergence criterion of Eq. 4 was met.

Table 1 summarizes the free energy differences obtained with these strategies. For comparison, the table also shows the free energies obtained from 1-dimensional umbrella sampling (ΔG_{US}), and from confinement simulations according to the setup of Ovchinnikov *et al.* (ΔG_{Hom}), which

involved 17 simulations at frequencies equally spaced in log space, with a simulation time of 20 ns per simulation.[14] Finally, the total cost of the simulations are shown relative to the total cost of the simulations using homogeneous spacing in frequency space. Fig. 7 shows the frequency spacing and number of steps for the alanine 10-peptide for each of the simulation setups; the number of steps is indicated by the length of the bars (but the unit length represents 10^8 steps for the homogeneous and 10^6 steps for the other setups). Because of the small time step, the high frequency simulations are particularly costly in the homogeneous frequency setup. For this reason, Ovchinnikov *et al.* recommended simulating up to a frequency of 86 ps^{-1}, since the free energy difference for the alanine dipeptide is already converged at that frequency (even though the absolute free energies of the C_{7eq} and C_{7ax} configurations are not). A converged free energy difference at a lower frequency implies that the anharmonicity of the system at higher frequencies is the same for both configurations. However, this is not necessarily the case for large conformational changes, especially if new interaction were formed. While we also observed a convergence of the free energy difference for the alanine dipeptide at 86 ps^{-1}, omitting the high frequency portion led to an error in ΔG of between 0.10 and 0.31 kcal/mol for the alanine decapeptide, and 0.98 kcal/mol for lactoferricin. Due to these errors, we included all frequencies until the absolute free energies of the C_{7eq} and C_{7ax} states were converged.

-- Table 1 Here --

-- Figure 7 Here --

The free energies obtained by umbrella sampling showed good agreement with the confinement free energies for all the alanine systems. The confinement simulations gave free energies of ~3.3-3.5 kcal/mol per (ϕ, ψ) dihedral angles, which shows the lack of correlations between (ϕ, ψ) backbone angles. Backbone rotation in the larger systems act as backbone rotation in independent alanine dipeptide systems, as observed before for the alanine tripeptide.[12, 29] The four extrapolation strategies gave similar ΔG values, with error barrs 2 to 10 times smaller than with the homogeneous setup, which shows that all strategies could be used with good accuracy. The low error bars came from interpolation, which reduced the discretization error, the use of correlation times, which ensured sufficient sampling, and the use of MBAR. While the use of interpolation does not significantly affect the free energy differences, it significantly contributes

to the low error bars. With these strategies, the free energy spacing between consecutive simulations was small enough, so that configurations from multiple simulations could be used to increase the statistics at a given frequency. If the free energy spacing would be too high, the weights in Eq. 5 would be very small, which would then effectively prevent the mixing of configurations, and increase the error on the interpolated values. Use of MBAR and interpolation is therefore only useful when the frequencies are chosen judiciously. While all strategies gave values that were relatively close with low error bars, not all free energies of the various strategies overlap within their error bars, which indicates that the error bars are underestimated. This is likely due to insufficient sampling, which is not taken into account by the error bars. The problem of insufficient sampling cannot be easily solved, as one cannot quantify missing information.

Because each simulation was run until a fixed number of independent frames was obtained, the simulation time was different for each frequency. The low frequency simulations required the highest number of simulation steps, because of large correlation times (Fig. 6, 7). This correlation time was system-dependent, since at low frequencies the harmonic restraints were fairly weak and the system dynamics were only slightly affected by the restraints. Upon increasing the frequency, the simulation time dropped significantly, because of a drop in correlation times. At frequencies above ~12.5 ps^{-1} smaller time steps were required, so that even though correlation times were roughly constant at high frequencies, the required number of steps increased (Fig. 7). The various extrapolation strategies had non-constant frequency spacings that were larger than the homogeneous setup at low frequencies, but smaller at high frequencies. The difference in spacing is due to the free energy difference between neighboring simulations (ΔF of Eq. 5) which increases with frequency for a given frequency spacing. In addition, the total number of simulations increased with the size of the system. This makes sense, since for a purely harmonic system, the free energy difference between two frequencies is proportional to the number of degrees of freedom. The number of simulations at high frequencies, where the system is largely harmonic, will therefore scale ~linearly with the number of atoms. The cost of the first strategy, which is based on a constant free energy spacing of 5 kcal/mol between consecutive simulations, indeed increased with system size (Table 1).

The four extrapolation strategies led to much lower computational costs than the setup with homogeneous sampling in log-frequency space (between 1.5 and 12.3% of the cost). This was mostly due to much shorter simulation lengths at high frequencies. The computational cost of the three system-independent setups (strategy 1, 3 and 4 in Table 1) increased with the size of the system (as discussed above). The setup with a constant free energy spacing of 5 kcal/mol was the most expensive of the four strategies, as it required the most simulations at high frequency. The setup based on best-fits of ΔF_{ext} was the cheapest overall as it used the largest free energy spacing. This advantage was especially pronounced for the alanine decapeptide, for which extrapolations to high free energy differences were possible (Fig. 4). For future applications, obtaining system-dependent expressions for ΔF is not practical due to the simulation costs associated with estimating this expression. System-independent strategies are much more practical, and even the most aggressive strategy (strategy 4) presented here was accurate, as well as cost efficient.

While the optimized protocol consists of a combination of interpolation, extrapolation, optimized friction coefficients, and correlation analysis to determine simulation lengths, the contribution of the interpolation and extrapolation to the decrease in error was estimated for the alanine decapeptide by calculating ΔG using the 17 windows of the homogeneous setup and optimizing the friction coefficients, simulation length, and time steps only. This resulted in a free energy difference of -31.55 ± 0.44 kcal/mol, at 1.0% of the cost of the homogeneous setup. Relative to the umbrella sampling results, the partially optimized 17-window strategy led to a larger shift in the free energy than the fully optimized strategies, while the statistical error was also significantly larger (8.8, 4.9, 2.8, and 4.9 times larger than the fully optimized schemes, respectively). When taking the difference in simulation lengths into account, these statistics suggest that the interpolation/extrapolation strategy reduces the error two-fold for the alanine decapeptide. In fact, when calculating ΔG using the fully optimized schemes but at exactly the same cost of the partially optimized 17-window strategy (by using less frames), free energies of -31.18 ± 0.17 kcal/mol, -31.61 ± 0.18 kcal/mol, -29.58 ± 0.28 kcal/mol, and -30.24 ± 0.22 kcal/mol were obtained for the four schemes, respectively. Thus, for the alanine decapeptide, errors were about factor of two (2.6, 2.4, 1.6, and 2.0, respectively) lower when

interpolation/extrapolation was used. Lactoferricin results suggest that interpolation/extrapolation might become more important for larger systems though.

Lactoferricin. As shown by the alanine systems, larger molecules require simulations at more frequencies to maintain high accuracy. Performing these simulations sequentially, as was done for the alanine *n*-peptides, can be impractical if a high number of CPUs is available: running multiple simulations at the same time is then usually more wall-clock time efficient. In this case, two strategies can be employed. One can either use the extrapolation protocol to estimate the optimal frequencies from short test simulations, and then extend these simulations in a parallel fashion. Or one can start multiple simulations at pre-determined frequencies (estimated from experience), calculate the free energy differences between these simulations using MBAR, and insert extra simulations to obtain the desired free energy spacing between consecutive simulations. This second approach was used here in order to compute the free energy difference between the two lactoferricin conformations. The free energy spacing used corresponds to strategy 4 of Table 1, which is the most aggressive. Since we observed that extrapolation and interpolation are more accurate for larger system, we expect this strategy to be very accurate for lactoferricin. Simulations were run until 1000 independent frames were obtained, except when using replica exchange, for which we used 100 frames per replica (500 frames total).

Similar to the alanine systems, the correlation time of the confinement energy for lactoferricin was much higher at low frequencies than at high frequencies (Fig. 8A), because at low frequencies the harmonic restraints had a small impact on the overall dynamics of the molecule. For lactoferricin, this correlation time was as high as 10 ns for some frequencies, which would require extremely long simulations to obtain sufficient uncorrelated data. The long correlation times are likely a general feature for more complex biomolecules. To gain efficiency, it is therefore highly desirable to lower the correlation times at low frequencies. Multiple strategies can be employed: a careful choice of the reference structure, the use of additional restraints to eliminate subbasin hopping, or the use of replica exchange.

While any configuration that belongs to the basin of interest can be used as a reference structure, depending on the free energy landscape, different configurations may result in different

correlation times. An energy-minimized configuration is a straightforward choice, but there is no guarantee that this structure is most representative of the basin. A more representative configuration can easily be obtained by performing rmsd-based clustering of an unrestrained MD trajectory, which is then expected to yield lower correlation times. This procedure was used here to obtain the reference structures for both lactoferricin conformations.

At low frequencies, the dynamics of the system are only slightly affected by the harmonic restraints. The correlation time of the confinement energy is therefore very close to the correlation time of the rmsd for an unrestrained simulation of the same molecule. Restricting the motion of the molecule, by using additional restraints, will therefore lower the correlation time. However, the free energy might also be affected, depending on whether the definition of each basin is modified. Analysis of the α+β conformation trajectories showed that some backbone dihedral angles switched between two different values, with a long correlation time. We therefore added flat-bottom dihedral angle restraints to confine these angles near the values of the reference state. Fraying was observed in simulations of the β-hairpin conformation, which was prevented by the addition of NOE restraints to maintain hydrogen bonding between residues 2 and 24. While these restraints indeed restricted the peptide to a single basin, the effect on the correlation times was marginal (Fig. 8A).

Finally, correlation times can be broken and sampling can be enhanced by using temperature replica exchange. In replica exchange, multiple independent simulations are run at different temperatures, and at given time intervals coordinates between the different simulations are swapped based on a criterion that preserves detailed balance.[38] Sampling is enhanced by the use of elevated temperatures, while correlations times are broken after swapping. In order not to unfold the peptide, we only used a small temperature range. For each simulation with a harmonic oscillator frequency lower than 0.72 ps^{-1}, replica exchange with 5 replicas at temperatures of 300, 312, 324, 337, and 350 K was used. This setup reduced the correlation times by a factor ~10-100 (Fig. 8A). In addition, extra efficiency was gained since data from all the temperatures could be combined using MBAR.

Using these approaches, we obtained a free energy difference of 2.13 ± 0.05 kcal/mol in favor of the β-hairpin, which is indeed the stable form in solution. While many simulations were needed for each conformation (Fig. 8B), most of these were at high frequencies, and only required short simulation times (typically 10^5-10^7 steps per simulation). Most simulation steps were needed at frequencies between 0.7 and 2 ps^{-1}, but simulation times could likely have been reduced by also using replica exchange for these frequencies. When excluding interpolation/extrapolation from the optimization protocol (i.e. using the 17 frequencies of the homogeneous setup, but optimizing friction coefficients, lengths and time steps), a free energy difference of of 4.67 ± 4.19 kcal/mol was obtained, indicating the importance of the interpolation and extrapolation strategy for larger systems.

Conclusion

We showed that the accuracy and efficiency of the confinement method can be greatly increased by the use of interpolation and extrapolation of the confinement energies, and the careful consideration of correlation times. Interpolation can be used to obtain confinement energies at unsampled frequencies, which significantly reduces the discretization error. The free energy difference between two consecutive simulations must stay below a certain value for accurate interpolations; however, this difference can be increased for larger systems, and at higher frequencies. Extrapolated free energy differences between simulated and unsimulated frequencies can also be used as a guide to select the optimal frequencies of the simulations. Cost and accuracy can be further optimized by basing the duration of each simulation on correlation times, costs can be decreased by increasing the friction coefficient at high frequencies, and accuracy can be increased by combining all data from multiple simulations. This setup proved to be efficient, as it led to proper estimations of conformation free energy differences for alanine *n*-peptides, with significantly increased accuracy (factor of 2-10) and greatly decreased computational costs (factor of 8-67) compared to homogeneous sampling. Additional techniques were used to speed up sampling for lactoferricin, a much more complex system with very long correlation times at low frequencies. Correlation times were significantly reduced by the use of temperature replica exchange (factor ~10-100). They were also reduced by using a reference structure obtained from rmsd-based clustering of unrestrained simulations, which prevented

excessive subbasin hopping, and by the application of additional restraints to restrict the configurational space.

Our analysis revealed promising features for application of the confinement method to large systems. As illustrated by our alanine *n*-peptide and lactoferricin simulations, large systems will clearly take longer sampling times, since their configurational space is larger. To maintain accuracy the total number of simulations grows with system size, but most of these simulations are at high frequencies where sampling is relatively short (Fig. 8B). Moreover, the growth in the number of simulations is partly counteracted by the fact that at a given accuracy, the spacing in free energy can be larger for larger systems. It is likely that large and complex systems will suffer from long correlation times at low frequencies, as also observed for lactoferricin. We showed however, that sampling at low frequencies can be significantly reduced by temperature replica exchange and other strategies to reduce the correlation times. While treatment of large systems will be computationally expensive, our study provides effective ways by which costs and accuracy can be managed and controlled.

Acknowledgments

We thank Dr. Victor Ovchinnikov for stimulating discussions and assistance. Computer time at USF was provided by USF Research Computing, supported in part by NSF MRI CHE-1531590. This project was funded by Regione Lombardia and the Cariplo foundation, project 2013-1702.

References

(1) Mitsutake, A.; Mori, Y.; Okamoto, Y., Enhanced sampling algorithms. In *Methods in Molecular Biology*, Springer: New York, 2013; Vol. 924, pp 153-195.
(2) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta* **2015,** *1850*, 872-877.

(3) Doshi, U.; Hamelberg, D. Towards fast, rigorous and efficient conformational sampling of biomolecules: Advances in accelerated molecular dynamics. *Biochim. Biophys. Acta* **2015,** *1850*, 878-888.

(4) Spiriti, J.; Kamberaj, H.; Van Der Vaart, A. Development and Application of Enhanced Sampling Techniques to Simulate the Long-Time Scale Dynamics of Biomolecular Systems. *Int. J. Quantum Chem.* **2012,** *112*, 33-43.

(5) Christen, M.; Van Gunsteren, W. F. On searching in, sampling of, and dynamically moving through conformational space of biomolecular systems: A review. *J. Comput. Chem.* **2008,** *29*, 157-166.

(6) Zuckerman, D. M., Equilibrium Sampling in Biomolecular Simulations. In *Annual Review of Biophysics*, Rees, D. C.; Dill, K. A.; Williamson, J. R., Eds. 2011; Vol. 40, pp 41-62.

(7) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005,** *7*, 3910-3916.

(8) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *WIREs Comput. Mol. Sci.* **2011,** *1*, 826-843.

(9) Weinan, E.; Ren, W. Q.; Vanden-Eijnden, E. Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chem. Phys. Lett.* **2005,** *413*, 242-247.

(10) Ma, A.; Dinner, A. R. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B* **2005,** *109*, 6769-6779.

(11) van der Vaart, A. Simulation of conformational transitions. *Theor. Chem. Acc.* **2006,** *116*, 183-193.

(12) van der Vaart, A.; Karplus, M. Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations. *J. Chem. Phys.* **2007,** *126*, 164106.

(13) Cecchini, M.; Krivov, S. V.; Spichty, M.; Karplus, M. Calculation of Free-Energy Differences by Confinement Simulations. Application to Peptide Conformers. *J. Phys. Chem. B* **2009,** *113*, 9728-9740.

(14) Ovchinnikov, V.; Cecchini, M.; Karplus, M. A Simplified Confinement Method for Calculating Absolute Free Energies and Free Energy and Entropy Differences. *J. Phys. Chem. B* **2013,** *117*, 750-762.

(15) Esque, J.; Cecchini, M. Accurate Calculation of Conformational Free Energy Differences in Explicit Water: The Confinement-Solvation Free Energy Approach. *J. Phys. Chem. B* **2015,** *119*, 5194-5207.

(16) Roy, A.; Perez, A.; Dill, K. A.; MacCallum, J. L. Computing the Relative Stabilities and the Per-Residue Components in Protein Conformational Changes. *Structure* **2014,** *22*, 168-175.

(17) Stoessel, J. P.; Nowak, P. Absolute free-energies in biomolecular systems. *Macromolecules* **1990,** *23*, 1961-1965.

(18) Tyka, M. D.; Clarke, A. R.; Sessions, R. B. An efficient, path-independent method for free-energy calculations. *J. Phys. Chem. B* **2006,** *110*, 17212-17220.

(19) Capelli, R.; Villemot, F.; Moroni, E.; Tiana, G.; van der Vaart, A.; Colombo, G. Assessment of mutational effects on peptide stability through confinement simulations. *J. Phys. Chem. Lett.* **2016,** *7*, 126-130.

(20) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*. Academic Press (Elsevier USA): San Diego, 2002.

(21) Tyka, M. D.; Sessions, R. B.; Clarke, A. R. Absolute free-energy calculations of liquids using a harmonic reference state. *J. Phys. Chem. B* **2007,** *111*, 9571-9580.

(22) Hill, T. L. *Statistical Mechanics*. McGraw-Hill: New York, 1956.

(23) Tidor, B.; Karplus, M. The contribution of vibrational entropy to molecular association - the dimerization of insulin. *J. Mol. Biol.* **1994,** *238*, 405-414.

(24) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008,** *129*, 124105.

(25) Grossfield, A.; Zuckerman, D. M. Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. *Annu. Rep. Comp. Chem.* **2009,** *5*, 23-48.

(26) Brooks, B. R.; Brooks, C. L., III; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S., et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009,** *30*, 1545-1614.

(27) Neria, E.; Fischer, S.; Karplus, M. Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **1996,** *105*, 1902-1921.

(28) Schaefer, M.; Karplus, M. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.* **1996,** *100*, 1578-1599.

(29) Maragakis, P.; van der Vaart, A.; Karplus, M. Gaussian-mixture umbrella sampling. *J. Phys. Chem. B.* **2009,** *113*, 4664-4673.

(30) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977,** *23*, 327-341.

(31) Torrie, G. M.; Valleau, J. P. Non-physical sampling distributions in Monte-Carlo free-energy estimation - Umbrella sampling. *J. Comput. Phys.* **1977,** *23*, 187-199.

(32) Bellamy, W.; Takase, M.; Yamauchi, K.; Wakabayashi, H.; Kawase, K.; Tomita, M. Identification of the bactericidal domain of lactoferrin. *Biochim. Biophys. Acta* **1992,** *1121*, 130-136.

(33) Hwang, P. M.; Zhou, N.; Shan, X.; Arrowsmith, C. H.; Vogel, H. J. Three-dimensional solution structure of lactoferricin B, an antimicrobial peptide derived from bovine lactoferrin. *Biochemistry* **1998,** *37*, 4288-4298.

(34) Moore, S. A.; Anderson, B. F.; Groom, C. R.; Haridas, M.; Baker, E. N. Three-dimensional structure of diferric bovine lactoferrin at 2.8 angstrom resolution. *J. Mol. Biol.* **1997,** *274*, 222-236.

(35) Zhou, N.; Tieleman, D. P.; Vogel, H. J. Molecular dynamics simulations of bovine lactoferricin: turning a helix into a sheet. *Biometals* **2004,** *17*, 217-223.

(36) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi(1) and chi(2) Dihedral Angles. *J. Chem. Theory Comput.* **2012,** *8*, 3257-3273.

(37) Lee, M. S.; Feig, M.; Salsbury, F. R.; C.L. Brooks, I. New analytical approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **2003,** *25*, 265-284.

(38) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999,** *314*, 141-151.

Tables

Table 1. Free energy differences between the C_{7eq} and C_{7ax} conformations of the alanine *n*-peptide in kcal/mol. ΔG_{US} corresponds to the free energy difference calculated from 1-dimensional umbrella sampling. ΔG_{Hom} corresponds the free energy obtained from confinement at frequencies that are equally spaced in log space. A total of 17 of these were performed per conformation, with a constant simulation time of 20 ns per simulation, according to the setup of Ref. [14]. In the strategies 1-4, the simulations are spaced in free energy space according to a given relation (see text). All free energies are in kcal/mol.

			Strategy							
			1[a]		2[b]		3[c]		4[d]	
System	ΔG_{US}	ΔG_{Hom}	ΔG	Cost[e]	ΔG	Cost[e]	ΔG	Cost[e]	ΔG	Cost[e]
Ala_2	-3.42 ± 0.06	-3.57 ± 0.27	-3.25 ± 0.06	2.57%	-3.15 ± 0.06	2.41%	-3.77 ± 0.07	2.19%	-3.60 ± 0.11	1.50%
Ala_3	-7.01 ± 0.09	-6.84 ± 0.35	-6.91 ± 0.05	3.54%	-7.03 ± 0.06	2.20%	-6.99 ± 0.07	2.29%	-6.67 ± 0.07	1.98%
Ala_4	-9.38 ± 0.21	-10.03 ± 0.39	-9.72 ± 0.05	4.94%	-10.15 ± 0.08	2.33%	-10.38 ± 0.07	3.12%	-10.10 ± 0.08	2.86%
Ala_6	-16.25 ± 0.25	-17.21 ± 0.48	-16.96 ± 0.05	6.72%	-16.75 ± 0.09	2.88%	-17.20 ± 0.06	4.46%	-17.47 ± 0.10	3.67%
Ala_8	-22.6 ± 0.31	-23.41 ± 0.46	-23.47 ± 0.05	9.24%	-23.68 ± 0.09	3.27%	-23.85 ± 0.06	5.50%	-23.74 ± 0.09	4.47%
Ala_{10}	-28.98 ± 0.35	-29.87 ± 0.33	-30.94 ± 0.05	12.28%	-30.69 ± 0.09	3.67%	-29.60 ± 0.16	6.87%	-30.30 ± 0.09	5.39%

a. $\Delta F = 5$ kcal/mol.

b. ΔF from system-dependent best-fits of to ΔF_{ext} (Fig. 4). These fits are:

$\Delta F^{Ala_2} = 3.19 + 0.54\log(\nu/\nu_1)$, $\Delta F^{Ala_3} = 1.44 + 1.02\log(\nu/\nu_1)$, $\Delta F^{Ala4} = 0.95 + 1.27\log(\nu/\nu_1)$,

$\Delta F^{Ala_6} = 3.36 + 1.12\log(\nu/\nu_1)$, $\Delta F^{Ala_8} = 3.27 + 1.38\log(\nu/\nu_1)$, $\Delta F^{Ala_{10}} = 4.87 + 1.55\log(\nu/\nu_1)$.

c. $\Delta F_{ext} = 3 + 0.388\ln(\nu/\nu_1)$, shown by the lower dotted line in Fig. 4.

d. $\Delta F_{ext} = 5 + 0.485\ln(\nu/\nu_1)$, shown by the upper dotted line in Fig. 4.

e. Total computational cost as a percentage of the total computational cost when using homogeneous spacing in log frequency.

Captions for the figures

Figure 1. Illustration of confinement simulations and their spatial overlap. A) Sampled configurations using a very small force constant for the harmonic restraint. Sampling is mostly governed by the unbiased, original potential. B) In red: sampled configurations using a medium force constant for the harmonic restraint. Sampling is largely due to the confining harmonic potentials. C) In yellow: sampling using a high force constant. Sampling is (virtually) purely due to the confining harmonic potentials.

Figure 2. Structure of lactoferricin. Solution structure on right, structure of the lactoferricin sequence within the lactoferrin protein on left, and disulfide bonds shown as ball and stick.

Figure 3. Ratio between the confinement energy computed from extrapolation and from an actual simulation, for alanine 2-peptide (A), 3-peptide (B), 4-peptide (C), 6-peptide (D), 8-peptide (E), and 10-peptide (F) . Simulations with different restraints up to a frequency ν_{max} were used to extrapolate the confinement energy at higher frequencies. These extrapolated states have a higher free energy than the state corresponding to ν_{max}, with a difference of ΔG.

Figure 4. Free energy difference corresponding to an error of 5% on the confinement energy obtained by extrapolation. This free energy difference is between the simulated system having the highest harmonic restraints, and the one corresponding to the extrapolated frequency.

Figure 5. Interpolation of the confinement energy for alanine 2-peptide for different restraint frequency ranges: A) low frequencies (0.02-6.6 ps^{-1}), B) intermediate frequencies (6.6-18.4 ps^{-1}), and C) high frequencies (26.6-38.8 ps^{-1}), as well as D) for alanine 10-peptide at low frequencies (0.02-0.37 ps^{-1}). The value of the confinement energy is interpolated (red) for frequencies between two initial simulations (black). The thickness of the red line represents the error bar. Additional simulations (blue) are added thereafter to estimate the accuracy of the interpolation.

Figure 6. Correlation times for A) alanine 2-peptide and B) alanine 10-peptide for different restraint frequencies, calculated from the autocorrelation function of the confinement energy

(triangle symbols) and from block analysis (circle symbols). Multiple friction coefficients (from 1 to 20 ps^{-1}) were used as parameters for the Langevin thermostat.

Figure 7. Number of MD steps used in the confinement simulations of the alanine 10-peptide in the C_{7eq} conformation. Each vertical bar represents a simulation, and its length is linearly proportional to the number of steps. Different strategies were employed (see text), and led to different number of simulations.

Figure 8. A) Correlation times for lactoferricin in the β conformation, using a reference structure obtained from clustering (blue), the same reference with the addition of restraints (black), and temperature replica exchange in addition to the restraints (red). B) Number of MD steps used in the confinement simulations. Each vertical bar represents a simulation, and its length is linearly proportional to the number of steps.

Figures

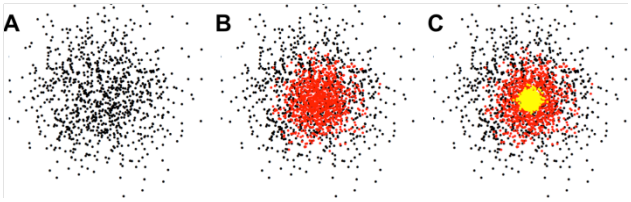


Figure 1.

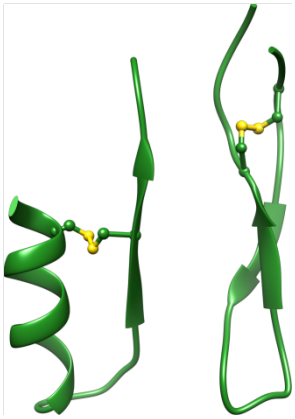


Figure 2

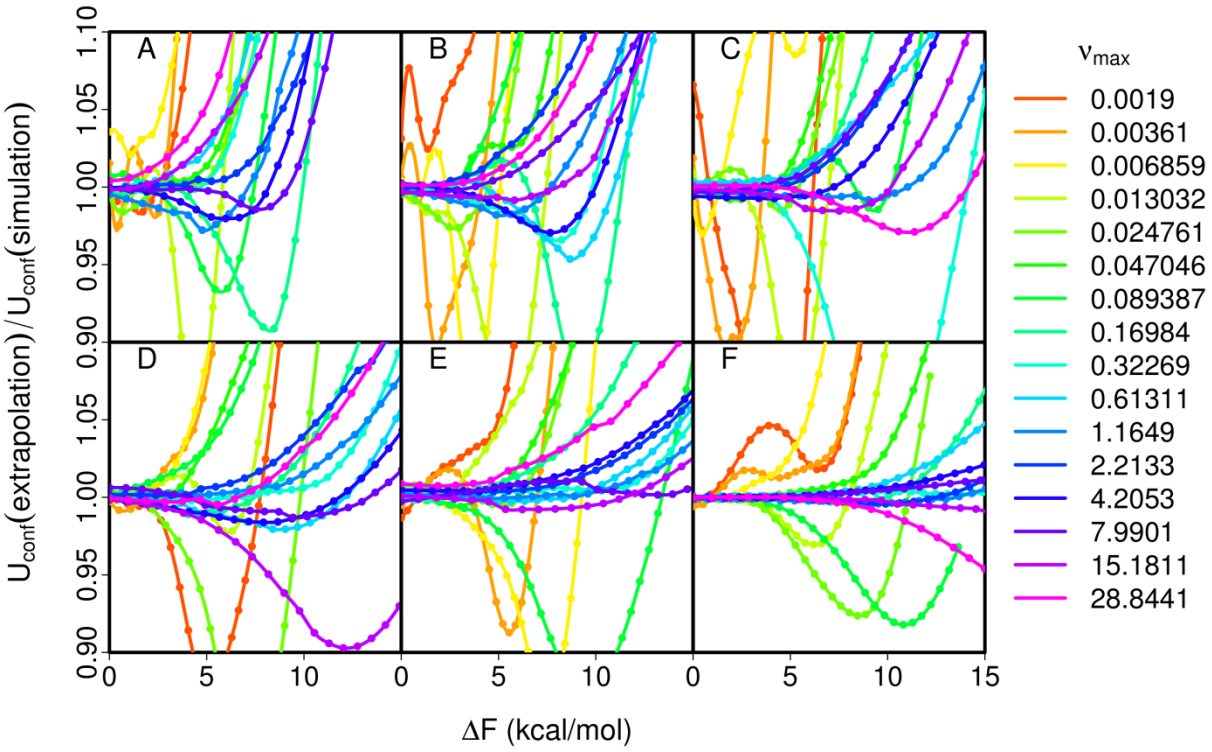

A
B
C
D
E
F
U_{conf}(extrapolation)/U_{conf}(simulation)
0.90
0.95
1.00
1.05
1.10
0
5
10
15
ΔF (kcal/mol)
ν_{max}
0.0019
0.00361
0.006859
0.013032
0.024761
0.047046
0.089387
0.16984
0.32269
0.61311
1.1649
2.2133
4.2053
7.9901
15.1811
28.8441

Figure 3

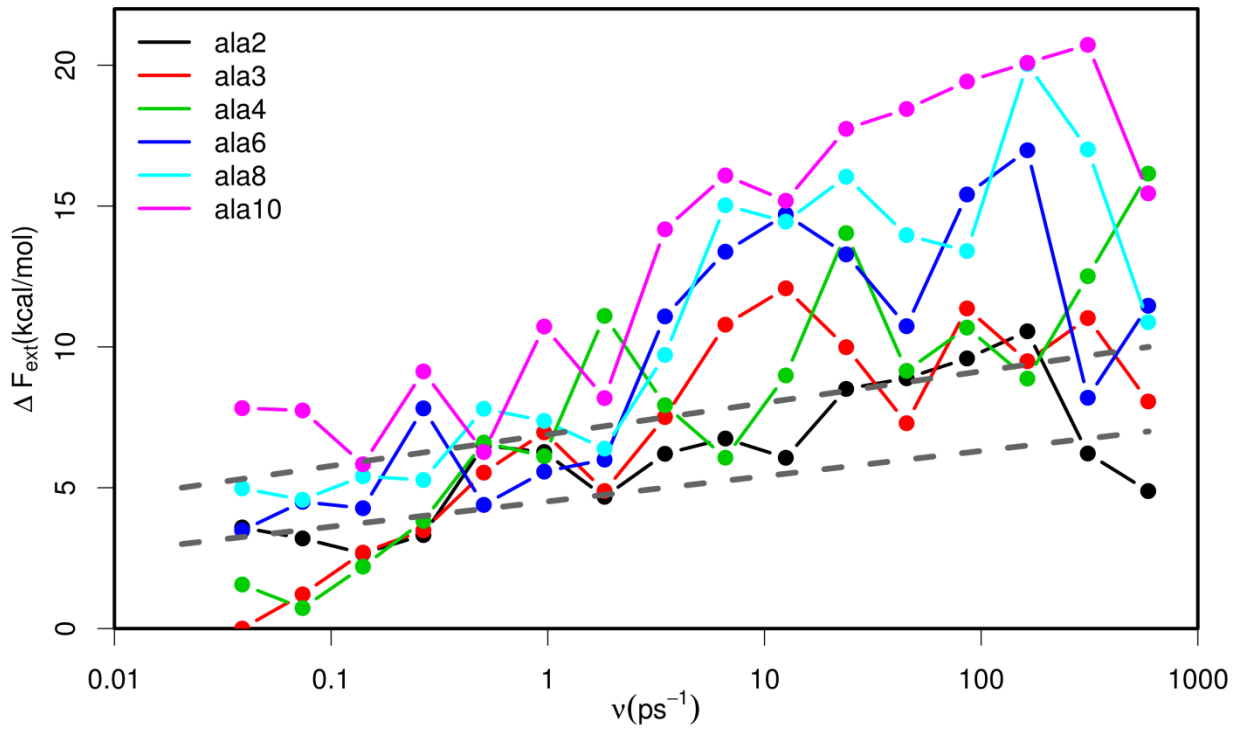

ala2
ala3
ala4
ala6
ala8
ala10
Δ F_ext(kcal/mol)
0
5
10
15
20
0.01
0.1
1
10
100
1000
ν(ps^-1)

Figure 4

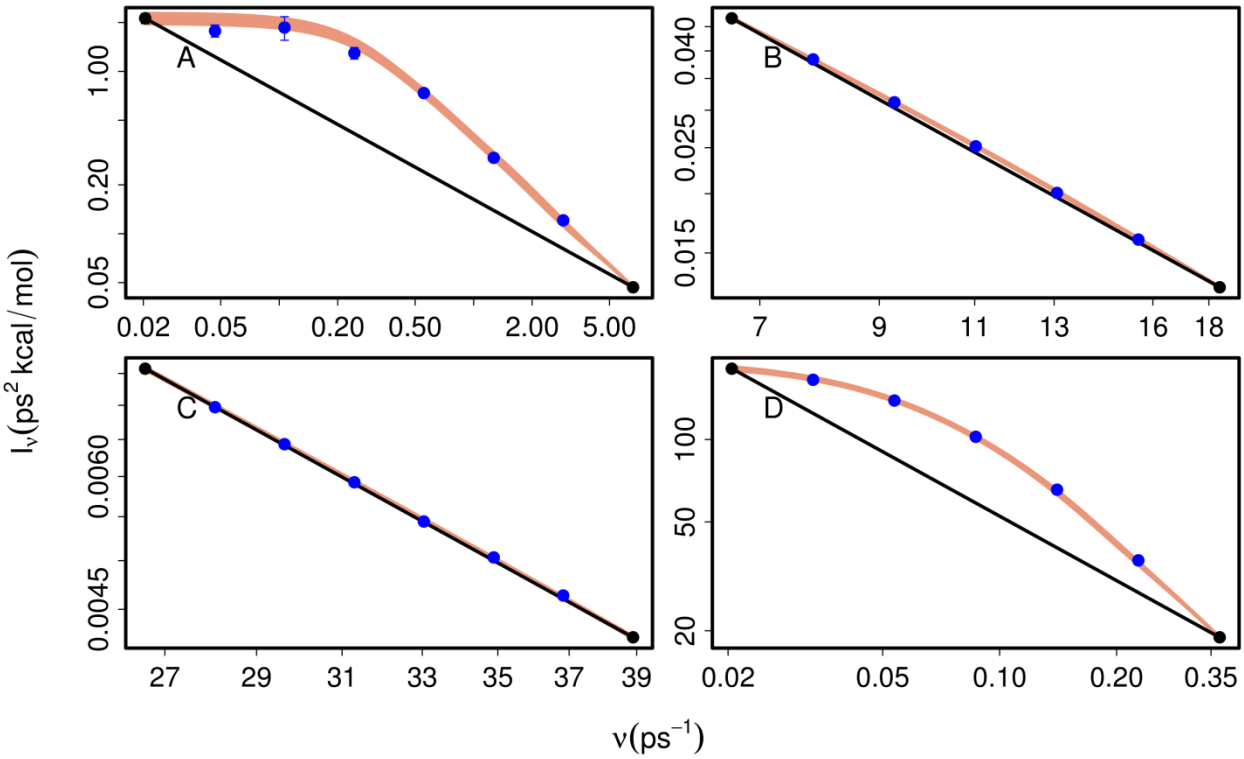

A
B
C
D
0.05
0.20
1.00
0.02
0.05
0.20
0.50
2.00
5.00
0.015
0.025
0.040
7
9
11
13
16
18
0.0045
0.0060
27
29
31
33
35
37
39
20
50
100
0.02
0.05
0.10
0.20
0.35
$I_\nu(ps^2\ kcal/mol)$
$\nu(ps^{-1})$

Figure 5

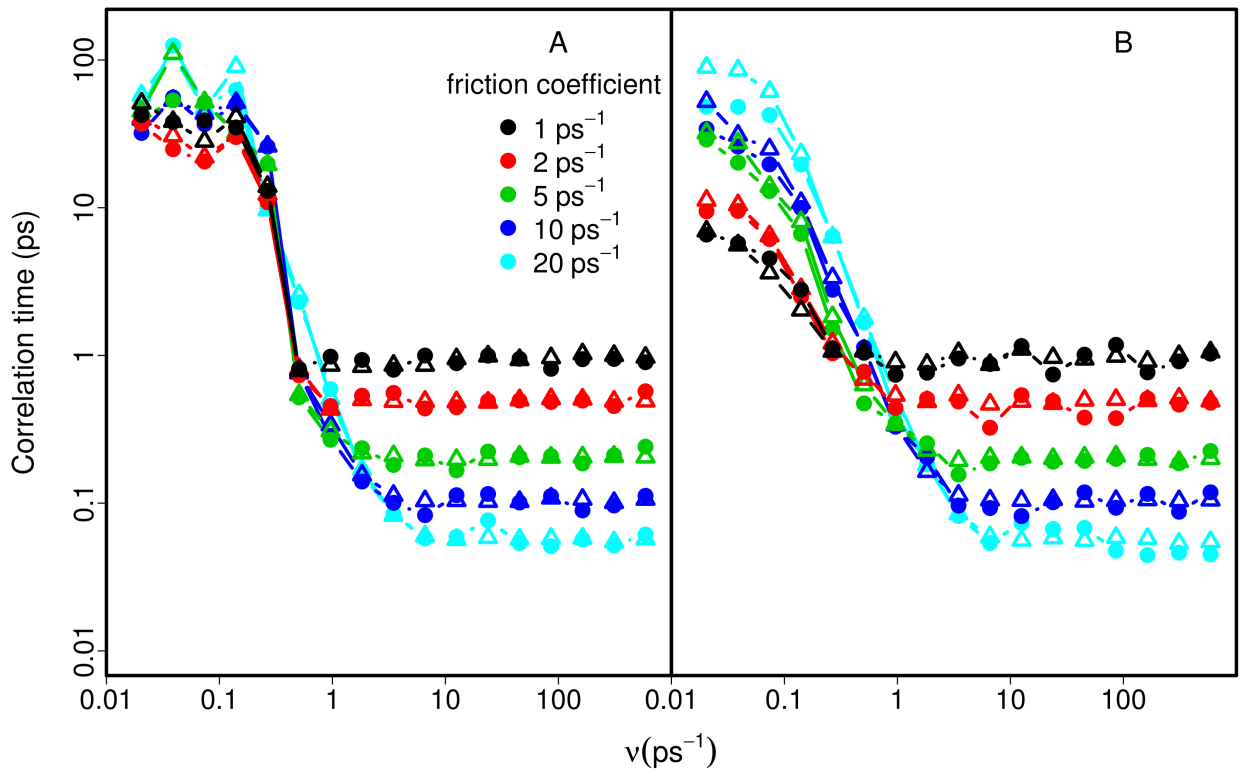

A
B
friction coefficient
1 ps^{-1}
2 ps^{-1}
5 ps^{-1}
10 ps^{-1}
20 ps^{-1}
Correlation time (ps)
0.01
0.1
1
10
100
ν(ps^{-1})

Figure 6

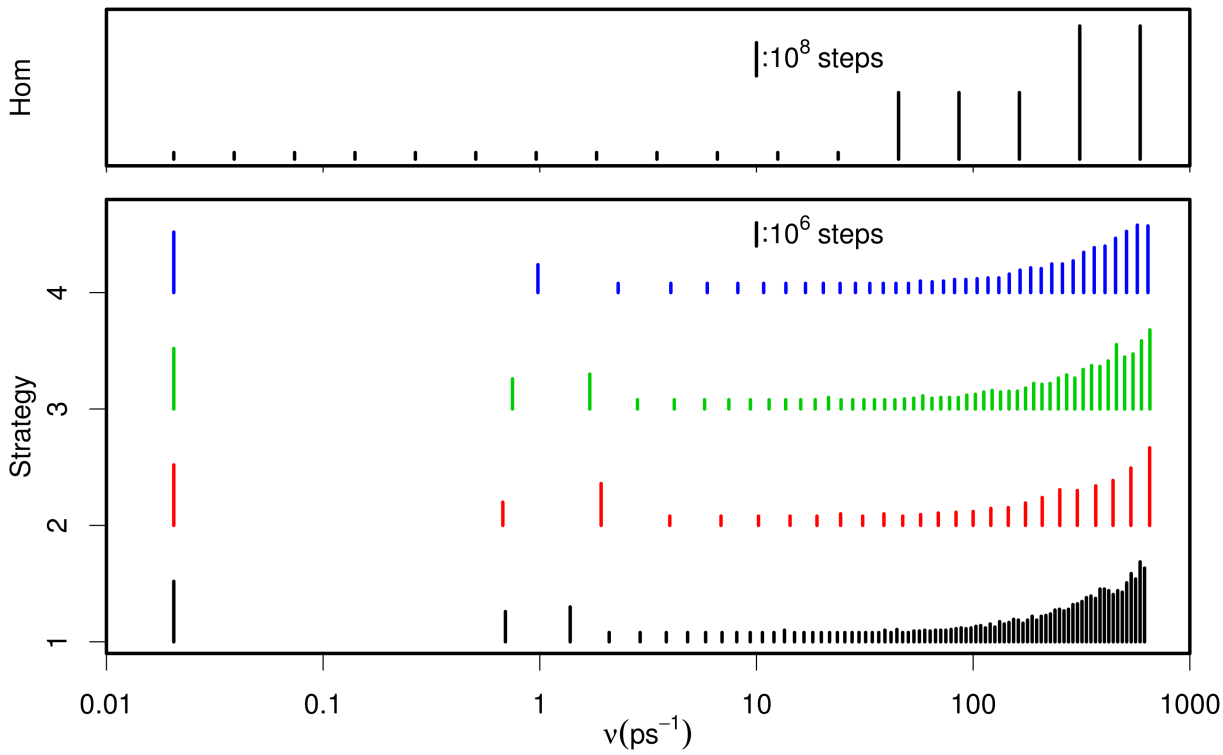

Hom
:10^8 steps
:10^6 steps
Strategy
4
3
2
1
0.01
0.1
1
10
100
1000
ν(ps^-1)

Figure 7

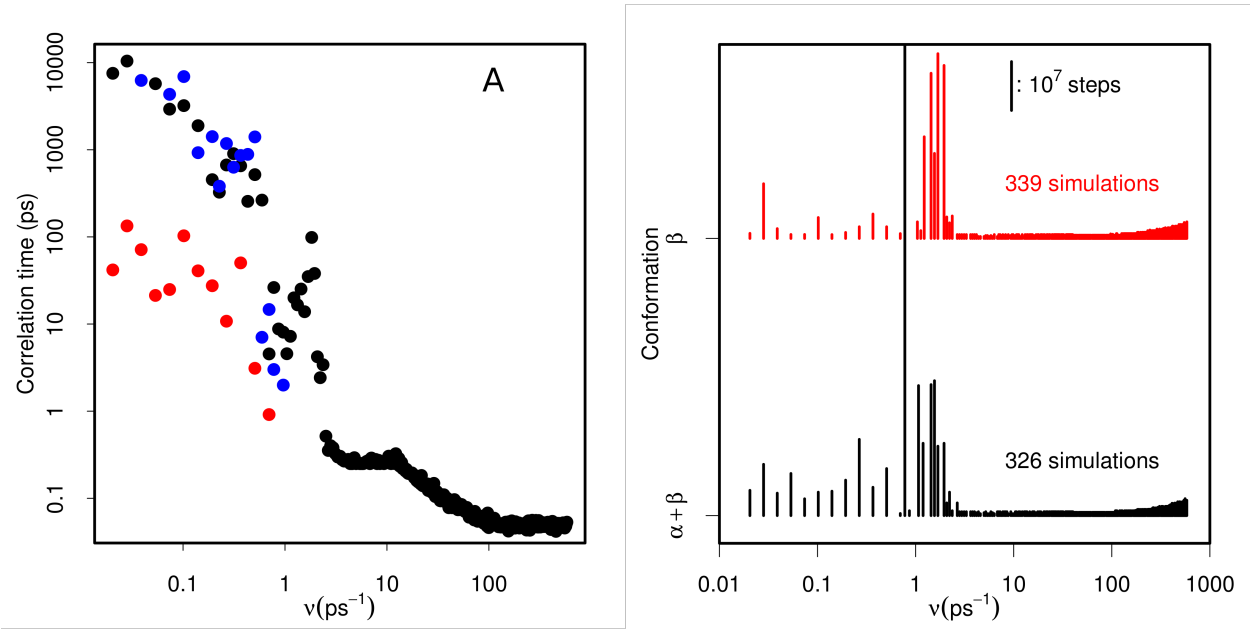
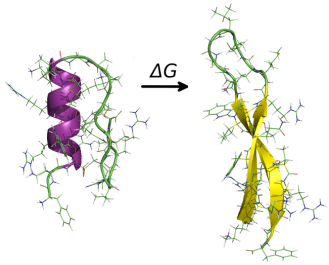

A
Correlation time (ps)
0.1
1
10
100
1000
10000
0.1
1
10
100
ν(ps^{-1})
: 10^7 steps
339 simulations
326 simulations
Conformation
β
α+β
0.01
0.1
1
10
100
1000
ν(ps^{-1})

Figure 8

TOC GRAPHIC



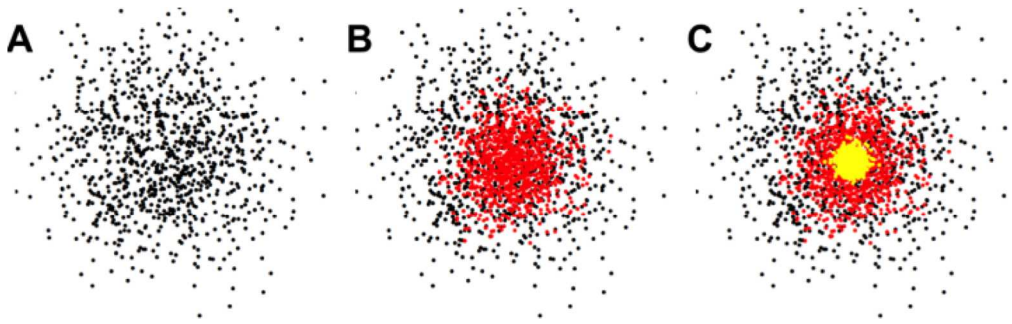


Figure 1. Illustration of confinement simulations and their spatial overlap. A) Sampled configurations using a very small force constant for the harmonic restraint. Sampling is mostly governed by the unbiased, unharmonic potential. B) In red: sampled configurations using a medium force constant for the harmonic restraint. Sampling is largely harmonic. C) In yellow: sampling using a high force constant. Sampling is (virtually) purely harmonic.
1030x319mm (72 x 72 DPI)

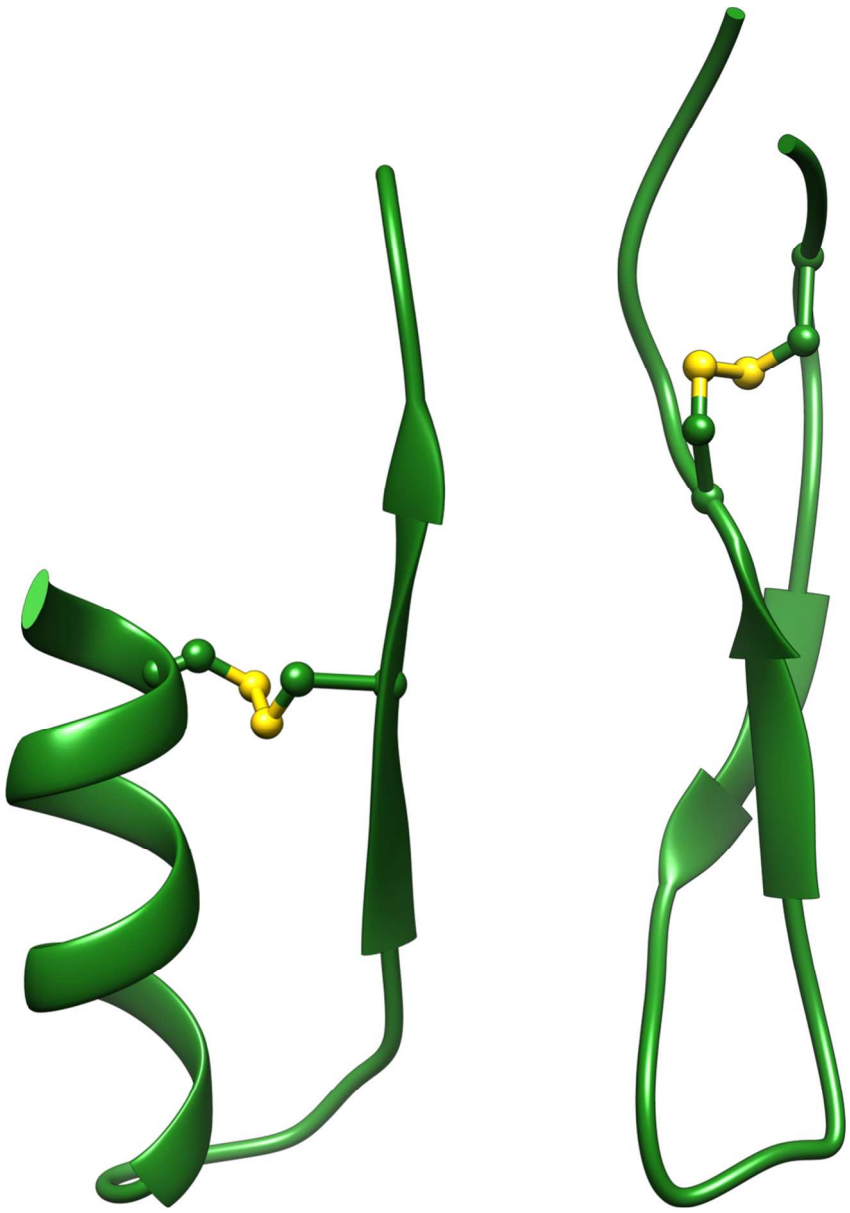


Figure 2. Structure of lactoferricin. Solution structure on right, structure of the lactoferricin sequence within the lactoferrin protein on left, and disulfide bonds shown as ball and stick.
377x539mm (72 x 72 DPI)

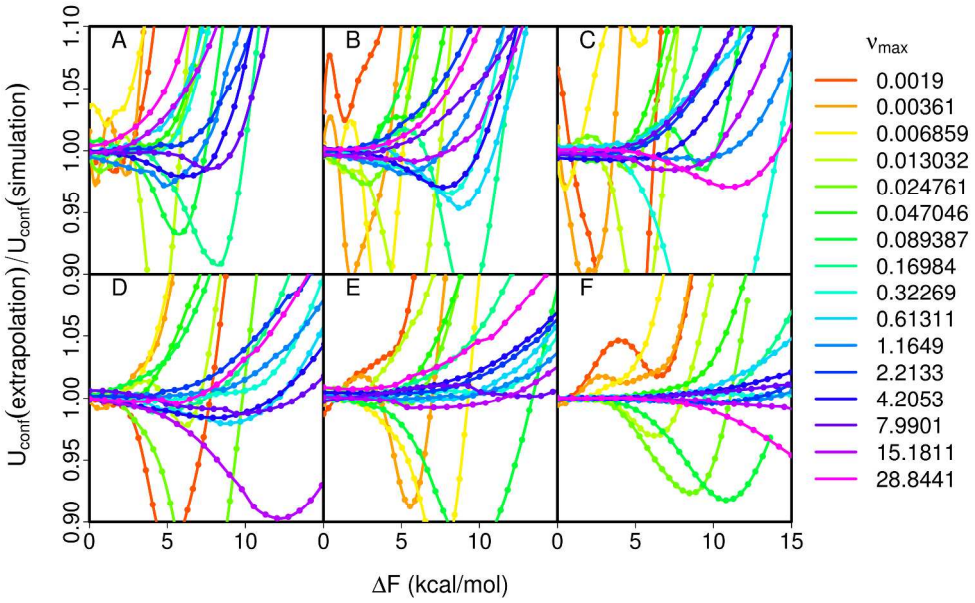


Figure 3. Ratio between the confinement energy computed from extrapolation and from an actual simulation, for alanine 2-peptide (A), 3-peptide (B), 4-peptide (C), 6-peptide (D), 8-peptide (E), and 10-peptide (F) . Simulations with different restraints up to a frequency were used to extrapolate the confinement energy at higher frequencies. These extrapolated states have a higher free energy that the state corresponding to nu_max , with a difference of delta G.
286x177mm (300 x 300 DPI)

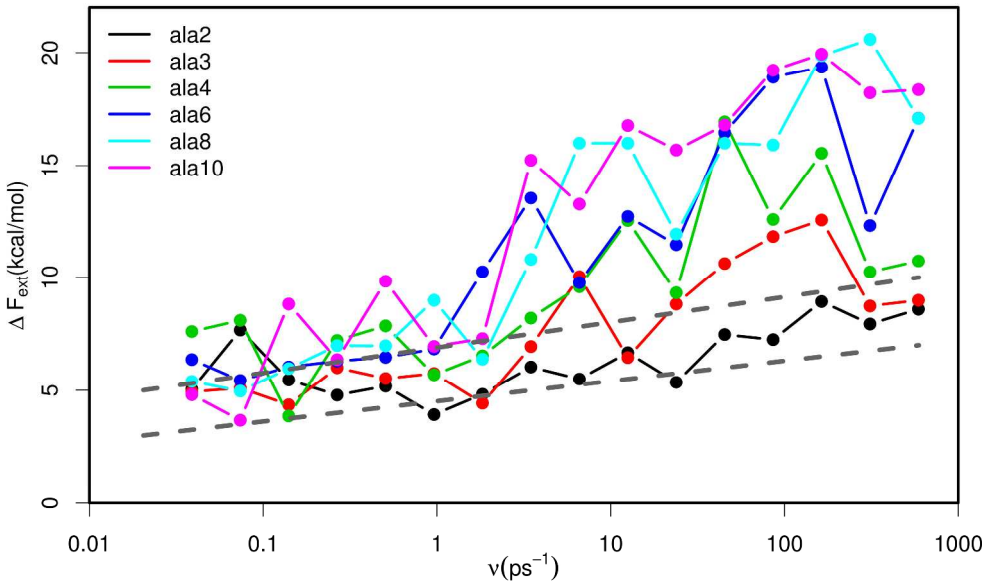


Figure 4. Free energy difference corresponding to an error of 5% on the confinement energy obtained by extrapolation. This free energy difference is between the simulated system having the highest harmonic restraints, and the one corresponding to the extrapolated frequency.
286x177mm (300 x 300 DPI)

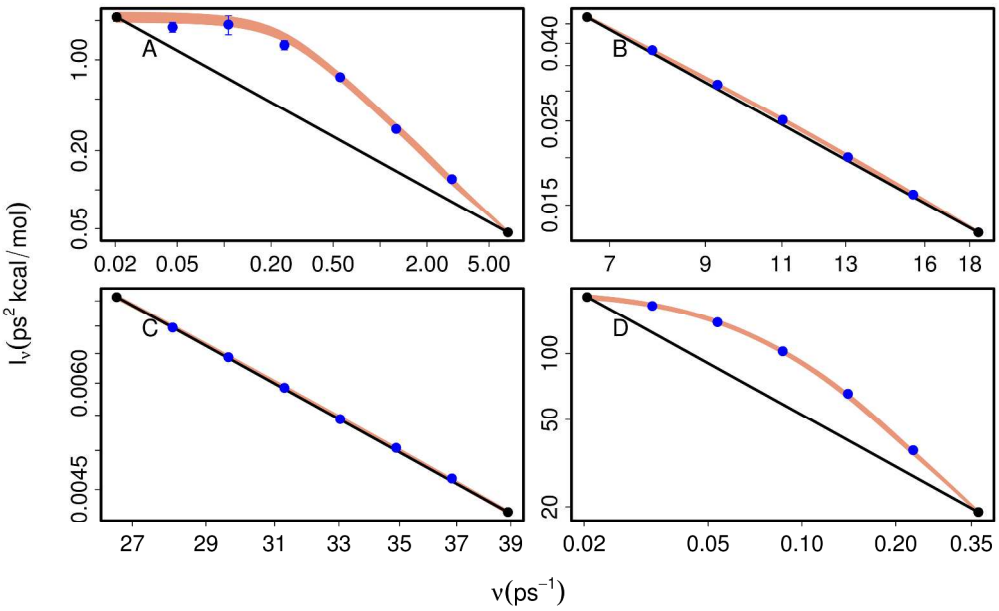


Figure 5. Interpolation of the confinement energy for alanine 2-peptide for different restraint frequency ranges: A) low frequencies (0.02-6.6 ps-1), B) intermediate frequencies (6.6-18.4 ps-1), and C) high frequencies (26.6-38.8 ps-1), as well as D) for alanine 10-peptide at low frequencies (0.02-0.37 ps-1). The value of the confinement energy is interpolated (red) for frequencies between two initial simulations (black). The thickness of the red line represents the error bar. Additional simulations (blue) are added thereafter to estimate the accuracy of the interpolation.
286x177mm (300 x 300 DPI)

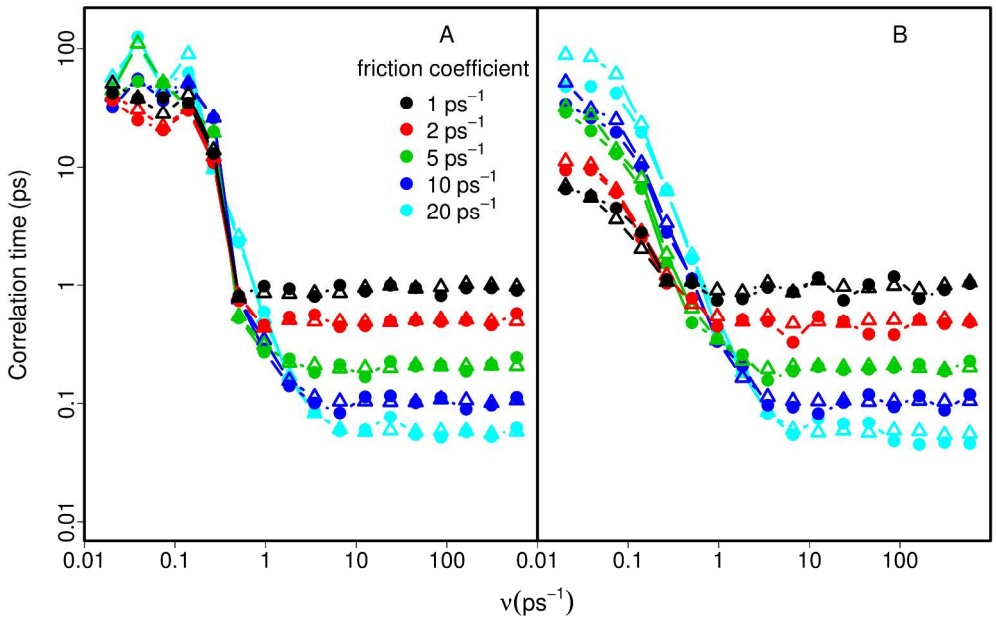


Figure 6. Correlation times for A) alanine 2-peptide and B) alanine 10-peptide for different restraint frequencies, calculated from the autocorrelation function of the confinement energy (triangle symbols) and from block analysis (circle symbols). Multiple friction coefficients (from 1 to 20 ps-1) were used as parameters for the Langevin thermostat.
286x177mm (300 x 300 DPI)

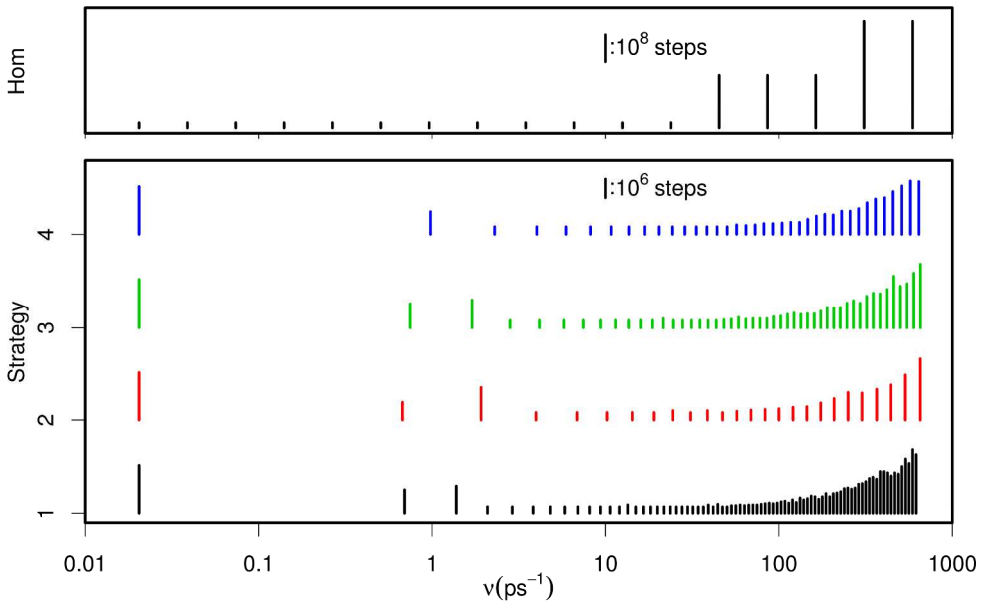


Figure 7. Number of MD steps used in the confinement simulations of the alanine 10-peptide in the C7eq conformation. Each vertical bar represents a simulation, and its length is linearly proportional to the number of steps. Different strategies were employed (see text), and led to different number of simulations.
286x172mm (300 x 300 DPI)

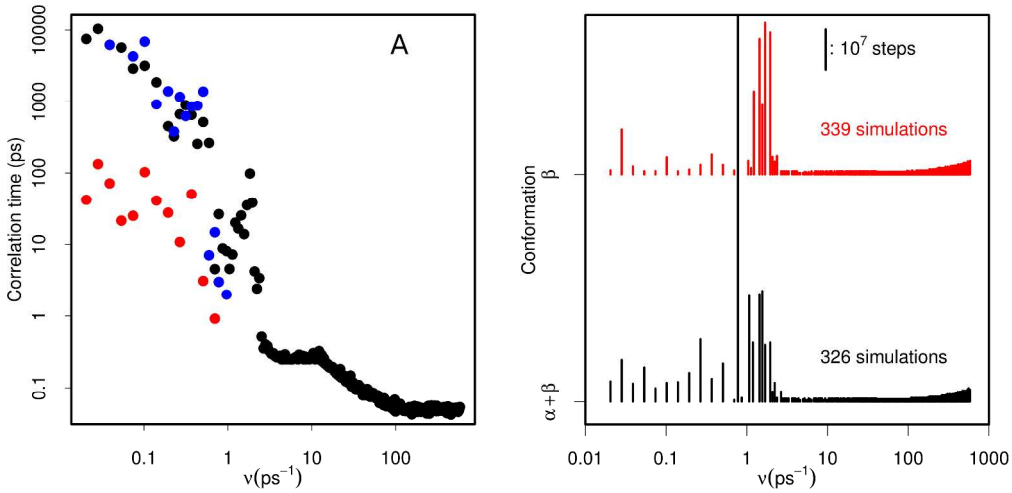


Figure 8. A) Correlation times for lactoferricin in the β conformation, using a reference structure obtained from clustering (blue), the same reference with the addition of restraints (black), and temperature replica exchange in addition to the restraints (red). B) Number of MD steps used in the confinement simulations. Each vertical bar represents a simulation, and its length is linearly proportional to the number of steps.
346x167mm (300 x 300 DPI)