PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

PCCP

RSCPublishing

ARTICLE

Toward Structure Prediction of Cyclic Peptides

Hongtao Yu[a] and Yu-Shan Lin[*a]

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Cyclic peptides are a promising class of molecules that can be used to target specific protein-protein interactions. A computational method to accurately predict their structures would substantially advance the development of cyclic peptides as modulators of protein-protein interactions. Here, we develop a computational method that integrates bias-exchange metadynamics simulations, a Boltzmann reweighting scheme, dihedral principal component analysis and a modified density peak-based cluster analysis, to provide converged structural description for cyclic peptides. Using this method, we evaluate the performance of a number of popular protein force fields on a model cyclic peptide. All the tested force fields seem to over-stabilize the α-helix and PPII/β regions in the Ramachandran plot, commonly populated by linear peptides and proteins. Our findings suggest that re-parameterization of a force field that well describes the full Ramachandran plot is necessary to accurately model cyclic peptides.

Introduction

Protein-protein interactions contribute to most aspects of biological processes, including signal transduction, membrane transport, and cell metabolism. Aberrant protein-protein interactions are involved in many human diseases.[1] Modulating protein-protein interactions thus offers a rich vein for therapeutic intervention. Unlike protein-ligand binding sites, which generally exhibit well-defined binding pockets that can be targeted with small molecules, protein-protein interfaces are relatively flat and large.[2-4] Cyclic peptides are a class of protein-protein interaction modulators with many promising pharmacokinetic characteristics. They are able to bind large protein surfaces with high affinity and specificity. In addition, they have enhanced metabolic stability and oral availability compared to their linear counterparts. Moreover, recent studies have demonstrated that the binding affinity and bioavailability of cyclic peptides can be further improved by, for example, N-methylation and solvent shielding by branched side chains.[5-8] Thus, cyclic peptides are a promising compound class for therapeutic modulation of challenging protein-protein interactions.

Several cyclic peptides are FDA-approved or in clinical trials as immunosuppressants,[9] artibiotics,[10, 11] antifungals[12] and potential antiviral[13] and anticancer therapeutics.[14] Despite these successful therapeutic applications of cyclic peptides, their potential remains underexplored – most potent cyclic peptides are simply natural products or their derivatives, rather than rationally designed.[15] A key impediment to fully exploring cyclic peptides as a promising class of therapeutics is that it is currently difficult to accurately predict the three-dimensional conformation that any given cyclic peptide will adopt. Furthermore, owing to their macrocyclic character, small modifications to cyclic peptides often cause massive conformational alterations. These challenges have rendered the optimization of cyclic peptides for biological targets a purely empirical pursuit, requiring brute force synthesis of many variants in hopes of finding one with appropriate conformational and target-binding properties.[7, 16]

In the past years, several methods have been proposed to predict the three-dimensional structure of cyclic peptides using computer simulations.[17-21] Most of these methods are based on the implicit solvent model. In aqueous solutions, hydrogen bonding with water plays a significant and often dominant role in determining peptide structures, which can be modelled in Molecular Dynamics (MD) simulations by using an explicit solvent model.[22, 23] However, the usefulness of MD simulations largely depends on the reliability of the force field implemented and the sufficiency of the sampling in relevant conformational space. In a MD simulation, Newton's equation of motion is numerically integrated to simulate the system's dynamics and theoretically, a Boltzmann-weighted ensemble can be obtained after a sufficiently long MD run. Unfortunately, due to the small size of MD time steps and the roughness of free energy landscapes in bimolecular systems, conventional MD simulations can be kinetically trapped in local free energy minima that are separated from the global minimum by large energy barriers. In the case of small cyclic peptides, their circular geometry may result in large free energy barriers between local minima and require coherent dihedral changes of multiple residues to sample a new conformation, thus making complete structure sampling even more challenging.

During the last decade, a number of advanced sampling techniques that aim to produce well-converged ensembles within a reasonable amount of simulation time have been developed.[24-35] Among these, replica-exchange molecular dynamics (REMD),[24, 36] also known as parallel tempering, is probably the most commonly used method. In standard REMD simulations, a series of non-interacting replicas of the system are simulated in parallel at different temperatures. Attempts are made at a regular time interval to swap the configurations of the neighbouring replicas based on the Metropolis acceptance criterion to enhance structure sampling. REMD thus exploits the more efficient conformational sampling in the higher-temperature replicas to enhance the conformational sampling of the lower-temperature replicas. In order to achieve the expected performance of REMD, there must be sufficient overlap in the energy space of adjacent replicas so that sufficient exchanges can be accepted. The number of replicas required depends on the number of degrees of freedom of the simulated system.[37]
Another popular enhanced sampling technique is bias-exchange metadynamics (BE-META). In a one-dimensional metadynamics (1D-META) simulation, the local free energy minima along a selected "reaction coordinate", often referred to as a "collective variable" (CV), are gradually filled with Gaussian hills to sample the full free-energy profile until converged.[38] In BE-META simulations,[35, 39, 40] *n* 1D-META simulations are performed simultaneously, each biased on one of the *n* CVs. During BE-META simulations, exchanges are attempted regularly between these *n* replicas to enhance sampling.

Protein force fields are generally parameterized by fitting to quantum chemistry calculations and experimental data for model systems. Over the past several decades, a number of force fields including AMBER,[41-46] OPLS-AA/L,[47] and GROMOS,[48-50] have been developed and widely used for simulations of bimolecular systems. The performance of these force fields has been extensively tested on *linear* peptide and protein systems, but their performance on highly constrained *cyclic* peptides remains to be determined. In this report, we first devise a method that integrates bias-exchange metadynamics simulations, a Boltzmann reweighting scheme, dihedral principal component analysis, and a modified density peak-based cluster analysis to provide converged structural description for cyclic peptides. Using this method, we then evaluate the performance of a number of popular protein force fields on a model cyclic peptide.

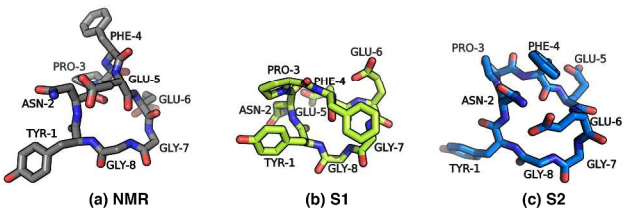

Fig. 1 The NMR structure (a) of cNPF1 and the initial structures (b, c) for MD simulations (hydrogen atoms not shown for clarity).

Methods

Model Peptide. A cyclic peptide cNPF1 (-YNPFEEGG-) was used as our benchmark peptide.[51] cNPF1 was designed to bind the EH domain of EHD1. The structure of cNPF1 in aqueous solution was determined using nuclear magnetic resonance (NMR) spectroscopy.[51] A tight ensemble of conformations were obtained from structure refinement using the NMR restraints,[51] and a representative structure is shown in **Fig. 1a**. The NMR spectroscopy revealed that the Asn2-Pro3-Phe4 motif of cNPF1 forms a type I β–turn configuration in aqueous solution (**Fig. 1a**). To perform MD simulations and monitor simulation convergence, two different initial structures were prepared from scratch using the Chimera molecular modelling package.[52] The first structure was built from an α-helix and the second from an extended conformation. The cyclic peptide was constructed by linking the N- and C-terminal residues of the linear peptide followed by an energy minimization. We will refer to these two initial structures as S1 and S2 (**Fig. 1b, c**). Convergence is considered achieved when simulations starting from the two different structures provide similar results.

Molecular Dynamics Simulation. Starting from the two input structures S1 and S2 (**Fig. 1b, c**), MD simulations were performed using the GROMACS 4.6.1 suite.[53] The initial structure (S1 or S2) was first immersed in a cubic box containing pre-equilibrated water molecules. The dimensions of the water box were chosen such that the minimum distance between any atoms of the peptide and the box walls is 1.0 nm. Two sodium ions were then added to neutralize the overall charge of the system. The solvated system was further energy optimized using the steepest descent algorithm to remove bad contacts. With the peptide heavy atoms restrained by a harmonic potential with a force constant of 1000 kJ/mol/nm, a 50 ps NVT (isochoric-isothermal) simulation and a subsequent 50 ps NPT (isobaric-isothermal) simulation were then implemented to equilibrate the solvent molecules and adjust the density. Before the production run, an additional 100 ps NVT simulation followed by a 100 ps NPT simulation without restraints was performed to equilibrate the whole system.

All production simulations were carried out in the NPT ensemble at a temperature of 300 K and a pressure of 1 bar. The temperature was maintained using the v-rescale thermostat[54] with a coupling time constant of 0.1 ps. To avoid the "hot solvent-cold solute" problem,[55-57] the peptide and solvent molecules were coupled to separate thermostats. The pressure was regulated using an isotropic Parrinello-Rahman barostat[58] with a coupling time of 2.0 ps and a compressibility of 4.5×10^{-5} bar^{-1}. The dynamics of the system were evolved using the leap-frog algorithm[59] with an integration time step of 2 fs. All bonds were constrained to their equilibrium values using the LINCS algorithm.[60] The non-bonded interactions (Lennard-Jones and electrostatic) were truncated at 1.0 nm. Long-range electrostatic interactions beyond the cut-off distance were calculated using the particle mesh Ewald (PME) method[61] with

a Fourier spacing of 0.12 nm and an interpolation order of 4. A long-range analytic dispersion correction was also applied to both the energy and pressure to account for the truncation of Lennard-Jones interaction.[62] The simulation trajectories were saved every 10 ps for subsequent analyses.

Replica-Exchange Molecular Dynamics Simulation. To improve the conformational sampling, we performed REMD simulations for cNPF1 using the equilibrated structure in the MD simulations. The numbers of replicas and temperatures implemented were determined by using the method proposed by Patriksson and van der Spoel.[37] The exchange attempts were made every 5 ps. All the REMD simulations were done in the NPT ensemble with separate thermostats for the peptide and solvent molecules. The simulation trajectories were saved every 1 ps for subsequent analyses.

Bias-Exchange Metadynamics Simulation. To enhance conformational sampling, we also performed BE-META simulations for cNPF1 using the PLUMED 2.0 plugin[63] for GROMACS. Eighteen CVs were biased in our simulations: TYR-1 $\varphi/\psi/\chi_1$, ASN-2 $\varphi/\psi/\chi_1$, PRO-3 ψ, PHE-4 $\varphi/\psi/\chi_1$, GLU-5 $\varphi/\psi/\chi_1/\chi_2$ and GLU-6 $\varphi/\psi/\chi_1/\chi_2$.The dihedral angles of GLY-7 and GLY-8 were not included due to the small size and large flexibility of glycine. The φ angle of PRO-3 was not used as a CV either, as biasing this dihedral can result in artificial *cis/trans* isomerization of the ASN-2/PRO-3 peptide bond. Exchanges between replicas were attempted every 5 ps. Gaussian hills of height 0.1 kJ/mol and width 0.314 rad were added every 4 ps. The simulation trajectories were saved every 1 ps for subsequent analyses. The free energy profile along each CV was recovered by summing the Gaussian hills added in the corresponding META simulation.

Structural Ensemble Analysis. Since biased potentials are added constantly during the course of BE-META simulations, the trajectories produced are in non-equilibrium and thus cannot be utilized directly for calculating the equilibrium properties of the simulated system. Recently, Laio and co-workers proposed a new method for recovering the equilibrium ensemble distribution from the biased MD simulation.[64] In their method, the simulation trajectories are first grouped into different clusters based on the pre-defined hypercubes; the population for each cluster is then determined through a complicated reweighting of the trajectory frames using the weighted histogram analysis method. In the present study, we considered an alternative approach of Boltzmann reweighting for acquiring an unbiased ensemble from a BE-META simulation. In our method, to obtain an equilibrium, unbiased ensemble from replica i (we used trajectory of 200−300 ns, after the free energy profile converged), frame k in replica i is either kept or discarded according to the Boltzmann probability:

$$p_k^i = e^{-\frac{\Delta G_i(s_k^i)}{k_B T}} \quad \begin{cases} \geq \rho & \text{kept} \\ < \rho & \text{discarded} \end{cases}$$

where s_k^i is the value of CV i in frame k, ΔG_i is the free energy profile along CV i (with the minimum free energy value shifted to 0), k_B is the Boltzmann constant and T is the temperature. A random number between 0 and 1, ρ, was generated and if $p_k^i \geq \rho$, frame k in replica i was kept; otherwise, this frame was discarded. The trajectories generated after Boltzmann reweighting thus obey the canonical distribution and can then be employed for subsequent analyses.

To provide structural insights into the cyclic peptide, principal component analysis (PCA) in combination with cluster analysis was carried out using the REMD trajectories and/or the Boltzmann reweighted BE-META trajectories. PCA is a well-established approach for dimensionality reduction without significant information loss. When performing using Cartesian coordinates, PCA reduces the highly correlated 3*N* atomic coordinates to a few uncorrelated collective degrees of freedom that contribute most to the essential dynamics of the system. The Cartesian PCA (cPCA) method is based on the *3N* dimensional covariance matrix:

$$c_{ij} = \langle (x_i - \langle x_i \rangle) \times (x_j - \langle x_j \rangle) \rangle$$

where $x_i, \cdots x_{3N}$ are the atomic coordinates and $\langle \cdots \rangle$ denotes the average over all sampled conformations. The modes of collective motion and their amplitudes are described by the eigenvectors (principal components, PCs) and eigenvalues of this covariance matrix. The PC associated with the largest eigenvalue accounts for the direction along which the system shows greatest variation. The PC with the second largest eigenvalue is orthogonal to the first one and describes the direction of second greatest variation. If the signal-to-noise is high in the data set, a large part of the system's variation can be represented by only the first few PCs. In recent years, PCA have been applied successfully for analyzing the MD trajectories of biomolecular simulations.[21, 65-73] A recent study by Stock et al. showed that PCA using internal coordinates such as dihedrals is free of the rotational fitting problem encountered in cPCA, and can provide well-resolved energy landscapes for, for example, villin headpiece HP35 and bovine pancreatic trypsin inhibitor.[74] Considering that RMSD in Cartesian coordinates may not be a sensitive metric to separate conformations for a cyclic peptide owing to its circular nature, the dihedral angle PCA (dPCA) using the φ/ψ angles of all the eight residues in cNPF1 was employed in this study.

Cluster analysis is another unsupervised dimensionality reduction technique for finding patterns within data. It organizes objects in such a way that objects within the same cluster are more similar to each other than to those in other clusters. To further analyze the structural ensemble of cNPF1, we conducted a cluster analysis in the two-dimensional (2D) principal subspace using a density peak-based clustering algorithm recently developed by Rodriguez and Laio (hereafter referred to as the RL algorithm).[75] It should be pointed out that the original RL algorithm requires constructing a triangular matrix recording the distances between all data points in order to compute local density around each data point. Constructing such a large distance matrix may be computationally intensive when dealing with MD simulation trajectories, which usually contain a large number of data points. Here we adopted a

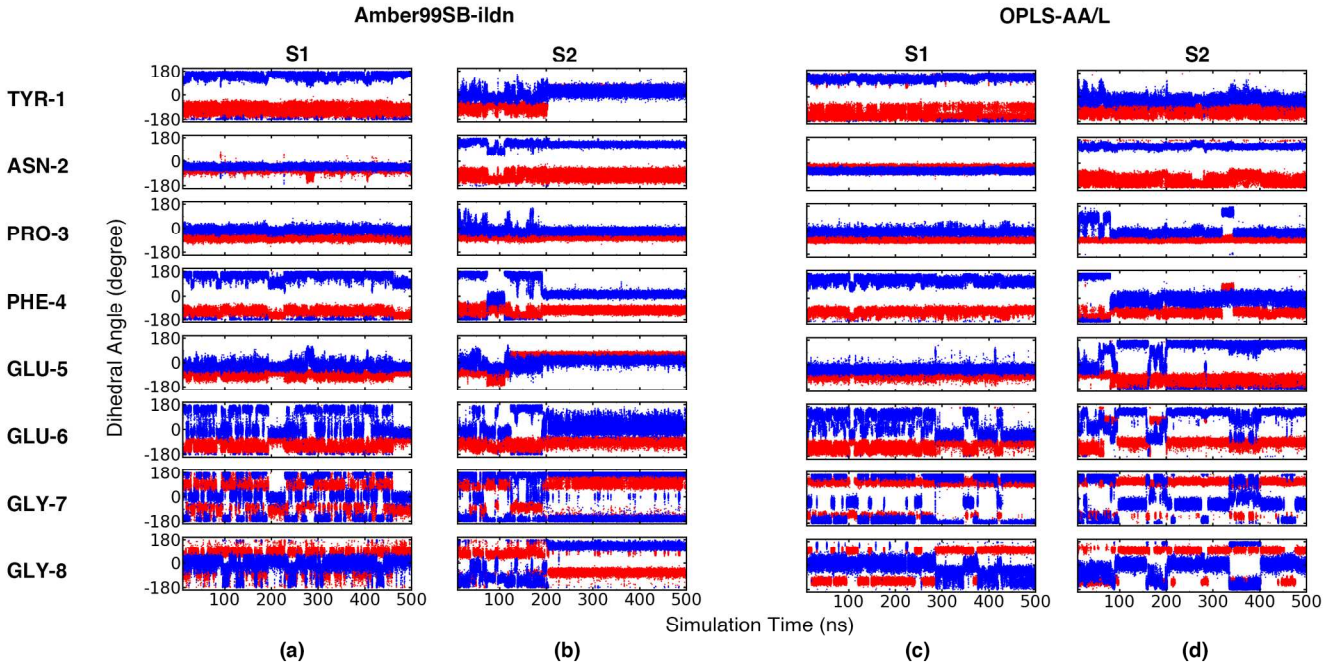

Fig. 2 Trajectories of φ (in red) and ψ (in blue) from MD simulations of cNPF1 at 300 K/1 atm. (a) Amber99SB-ildn simulation using input structure S1; (b) Amber99SB-ildn simulation using input structure S2; (c) OPLS-AA/L simulation using input structure S1; (d) OPLS-AA/L simulation using input structure S2.

variant of the RL algorithm. Instead of clustering the raw PCA data points, we first divided the 2D principal subspace into small 2D grids, and computed the data point population within each grid. The local density of grid i is then estimated by

$$\rho_i = \sum_{j\in\Omega} w_j e^{-\frac{d_{ij}}{d_c}}$$

where w_j is the data point population with grid j, d_{ij} is the distance between grids i and j, and d_c is a cutoff distance. The sum over j is over all the grids (including grid i). Using the grids and their local densities, centre grids were identified if their densities are higher than their neighbours and their distances to grids with higher densities are relatively large. Once the grids are properly clustered, the original PCA data points can then be assigned based on the cluster IDs of the grids to which they belong.

Results and Discussion

Conventional MD Simulations of cNPF1. Starting from the two input structures S1 and S2 (**Fig. 1b, c**), 500 ns NPT MD simulations at 300 K/1 bar were performed for cNPF1 using the Amber99SB-ildn force field[46] coupled with the TIP3P water model[76] and the OPLS-AA/L force field[47] coupled with the TIP4P water model[76, 77]. Owing to the intrinsic backbone rigidity of proline and flexibility of glycine, the region around PRO-3 is expected to be the most stable and the region around GLY-7 and GLY-8 the most flexible. In **Fig. 2**, we plot how the eight sets of φ/ψ angles of cNPF1 changed during the 500 ns simulations. During the simulations starting from input structure S1, the region of residues 6−8 (Glu6-Gly7-Gly8) was indeed the most flexible in both the Amber99SB-ildn and OPLS-AA/L simulation, while all the other dihedrals (residues 1−5) were stuck in the initial conformation and their values rarely changed (**Fig. 2a, c**). During the simulations starting from input structure S2, the peptide conformation in the Amber99SB-ildn simulation relaxed in the first 200 ns and then locked into a rather stable conformation for the remaining 300 ns (**Fig. 2b**). The simulation for S2 with the OPLS-AA/L force field showed a similar behaviour as the Amber99SB-ildn S2 simulation: The peptide relaxed in the first 100 ns and then locked into a relatively stable conformation (**Fig. 2d**). By comparing the φ/ψ trajectories of S1 and S2 simulations in **Fig. 2**, it is clear that constant-temperature MD simulations do not provide a converged description for cNPF1 within 500 ns simulation time, either with the Amber99SB-ildn or the OPLS-AA/L force field.

Replica-Exchange MD Simulations of cNPF1. REMD is a widely used method for enhancing sampling in MD simulations. To improve conformational sampling, we performed REMD simulations for cNPF1 starting from structures S1 and S2 using both the Amber99SB-ildn (with TIP3P water) and OPLS-AA/L (with TIP4P water) force fields. In order to acquire sufficient overlaps in potential energy space between neighbouring replicas and consequently reasonable exchange acceptance ratios, 59 and 51 replicas were used for the Amber99SB-ildn and OPLS-AA/L simulations respectively. The potential energy distributions obtained from these two sets of simulations are presented in **Fig. S1**. Throughout the entire

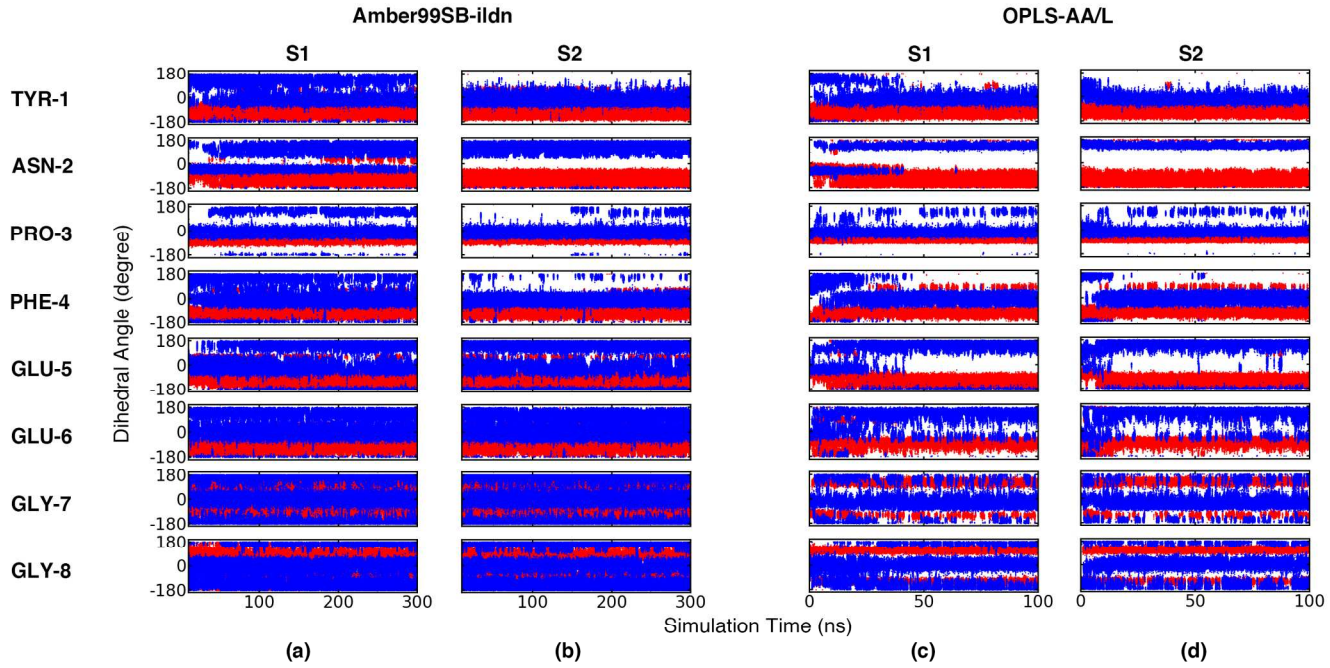

Fig. 3 Trajectories of φ (in red) and ψ (in blue) from the 300 K replicas of REMD simulations. (a) Amber99SB-ildn simulation using input structure S1; (b) Amber99SB-ildn simulation using input structure S2; (c) OPLS-AA/L simulation using input structure S1; (d) OPLS-AA/L simulation using input structure S2.

temperature range, the exchange acceptance probabilities were 45-55% and 32-42% for the Amber99SB-ildn and OPLS-AA/L simulations, respectively. **Fig. 3** shows the trajectories of the eight sets of φ/ψ dihedrals of cNPF1 in the 300 K replicas of the REMD simulations. We observe more changes in the dihedrals when compared to the conventional MD simulations (**Fig. 2**), suggesting that structures sampling at 300 K were indeed enhanced by the REMD simulations. In the simulations using the OPLS-AA/L force field, the φ/ψ angles of the two sets of simulations starting from input structures S1 (**Fig. 3c**) and S2 (**Fig. 3d**) reached convergence after approximately 50 ns. However, the simulations using the Amber99SB-ildn force field remained unconverged after 300 ns run (e.g., note the TYR-1 ψ angles in **Fig. 3a, b, top row, blue**).

To further verify the simulation convergence and provide structural insights for cNPF1 in the OPLS-AA/L REMD simulations, (**Fig. 3c, d**) we carried out dPCA and cluster analysis using the last 50 ns of the REMD simulation trajectories. The eigenvectors (PCs) of the covariance matrix were calculated using the combined trajectories of the S1 and S2 simulations, and these two simulations were then projected individually onto the first two largest PCs (PC1 and PC2). **Fig. 4a, b** show the conformational density profiles projected onto PC1 and PC2. Both the S1 and S2 simulations reveal multiple low-population conformational states with one prominent state located around PC1=1.5 and PC2=0 (**Fig. 4a, b**). In order to quantitatively characterize the conformational ensemble, we first divided the 2D principal subspace into 200×200 grids, and then performed cluster analysis based on the grids as described in the *Methods* section. To enhance the performance and efficiency of the cluster analysis, the cluster analysis was only performed on the grids with data populations larger than 0.1 (**Fig. S2c, d**). As shown in **Fig. 4c, d**, the conformational states are well resolved by the clustering algorithm. The population for each state was determined by summarizing the data populations of the grids the state contains. Using the last 50 ns of the REMD simulations, the populations of the most populated state are estimated to be 50.4% and 49.7% in the S1 and S2 simulations, respectively. Discarding the grids with data

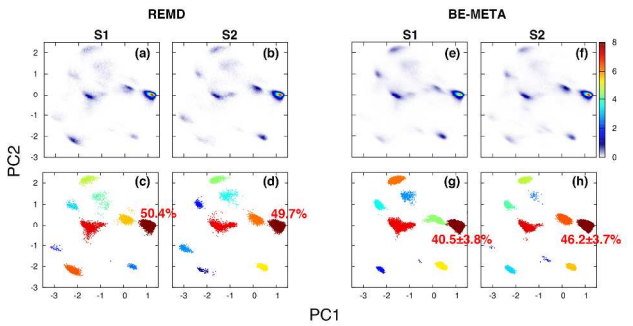

Fig. 4 Conformational density profiles as a function of the first two largest principal components of cNPF1 as obtained from the dPCA of the OPLS-AA/L simulations. The principal components were calculated using the S1/S2 combined trajectories of the last 50 ns of the REMD simulations. Left panel: Conformational density calculated from the REMD simulations using input structures S1 (a) and S2 (b), and their corresponding cluster analysis results (c and d). Clusters are colored based on their populations with the largest cluster colored in dark red and the smallest cluster colored in dark blue. Right panel: Conformational density calculated from the BE-META simulations using input structures S1 (e) and S2 (f), and their corresponding cluster analysis results (g and h).

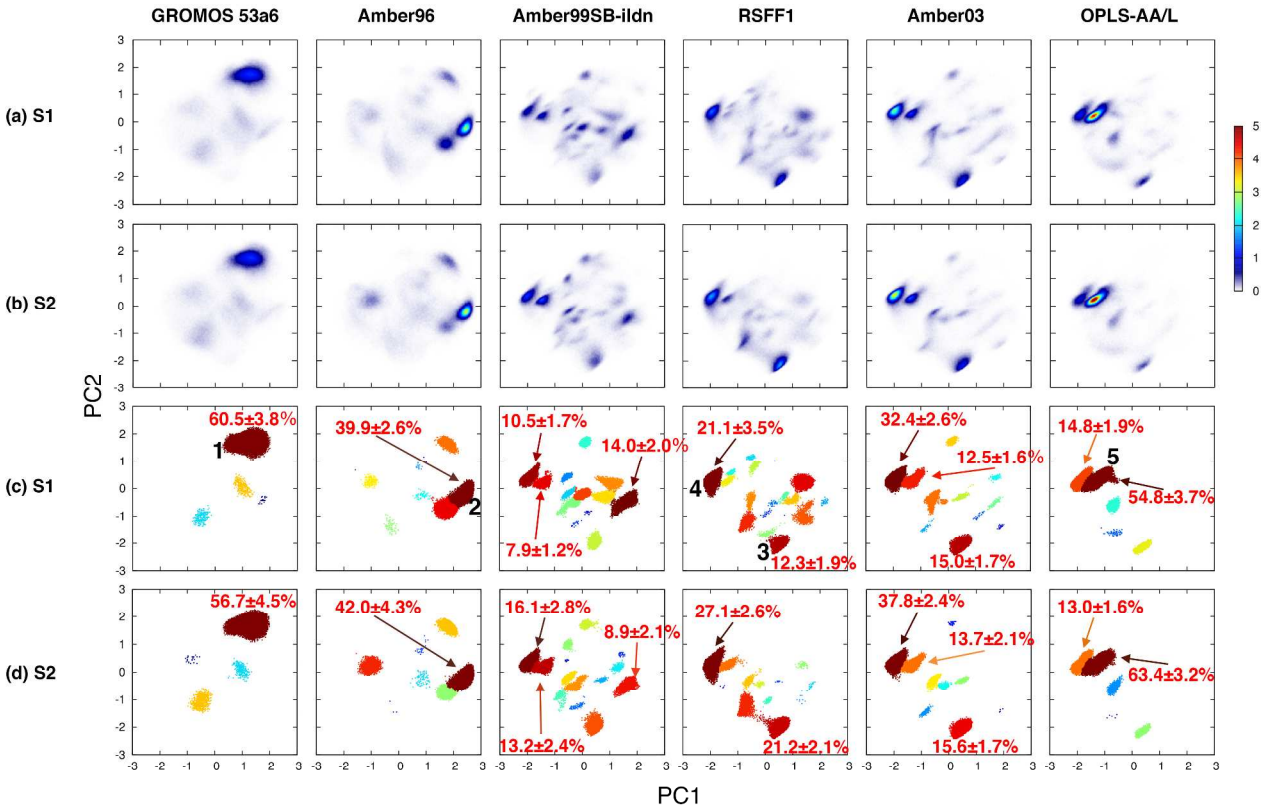

Fig. 5 Conformational density profiles as a function of the first two largest principal components of cNPF1. dPCA was performed on the last 100 ns of all twelve weighted BE-META trajectories (6 force fields × two input structures S1 and S2). Columns from left to right: Results for GROMOS 53a6, Amber96, Amber99SB-ildn, RSFF1, Amber03, and -AA/L. Rows from top to bottom: Conformational density calculated using input structures S1 (row a) and S2 (row b), and their corresponding cluster analysis results (rows c and d). Clusters in rows c and d are colored based on their populations with the largest cluster colored in dark red and the smallest cluster colored in dark blue.

populations lower than 0.1 only results in an underestimation of this population by approximately 1.5% (**Fig. S2g, h**). In **Fig. S3**, we plot the φ/ψ distributions of the 8 residues in cNPF1 for each state. The φ/ψ distributions fall in different regions for different states, indicating that the dPCA coupled with cluster analysis has successfully separated the multiple conformations of cNPF1 in the OPLS-AA/L REMD simulations.

Bias-Exchange Metadynamics Simulations of cNPF1. Given the expensive computational demands of REMD simulations and the poor convergence of the REMD simulations using the Amber99SB-ildn force field even after a relatively long REMD run, we sought an alternative enhanced sampling technique, i.e., the BE-META method. We performed 300 ns BE-META simulations for cNPF1 using the aforementioned two sets of force fields and 18 CVs ($\varphi/\psi/\chi$ dihedrals, see the *Methods* section). In **Fig. S4a, b** we plot the free energy profiles along the 18 dihedrals calculated from our BE-META simulations using the two input structures S1 (green) and S2 (blue). We observe relatively similar results between the two sets of simulations for both the Amber99SB-ildn and OPLS-AA/L force fields.

It is straightforward to perform dPCA and subsequent clustering of the principal subspace on a conventional MD or REMD simulation trajectory. For a peptide that is not intrinsically disordered, there are generally only one or several low free energy regions in its conformational space. The lower the free energy is, the more time the peptide spends in this low free energy region during the course of MD or REMD simulation. This naturally results in a data set with high signal (low free energy conformations) to noise (high free energy conformations) ratio. The dPCA and subsequent cluster analysis can therefore yield few clusters with high populations. However, performing a dPCA and cluster analysis based on BE-META simulation trajectories can be tricky. While the conformations sampled in an MD or REMD trajectory may naturally cluster around a few low free energy regions, the conformations sampled in BE-META trajectories no longer have this feature because sampling along each CV is enhanced in each BE-META trajectory. For example, in **Fig. S5a**, we plot the φ/ψ distribution for each residue sampled in each BE-META trajectory of the OPLS-AA/L S1 simulation. Enhanced sampling along each CV results in an even distribution along the corresponding CV. Such a spread in the conformational space reduces the ratio of low-free energy conformations to high-free energy conformations and consequently hinders the performance of dPCA and cluster analysis. To mitigate this problem, we applied a Boltzmann reweighting scheme to each of the original BE-META trajectories based on the free energy

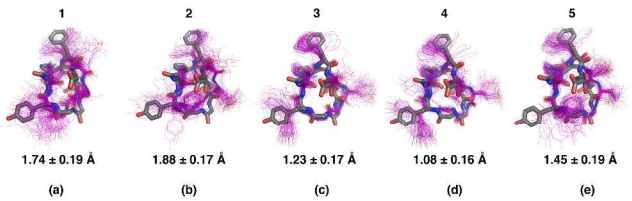

Fig. 6 NMR structure of cNPF1 (in gray) and representative conformations for the 5 most commonly populated states (in magenta). The backbone RMSD of each state with respect to the NMR structure is given below the structure.

profile along each CV. Since the resulting trajectories have thereby been properly weighed, they now represent an equilibrium structural ensemble (**Fig. S5b**), and compare well to the φ/ψ distributions from the 300K replica of the REMD simulation (**Fig. S5c**). **Fig. S6** further compares the distribution along each dihedral angle from the 300K replica of the REMD simulation (**Fig. S6, black lines**), from the raw BE-META trajectories (**Fig. S6, blue lines**), and from the Boltzmann reweighted BE-META trajectories (**Fig. S6, red lines**). The agreement between the distributions from the REMD and reweighted BE-META trajectories suggests that the trajectories generated with the Boltzmann reweighting method obey the canonical distribution.

We have demonstrated that the OPLS-AA/L REMD simulations converge after 50 ns and both the S1 and S2 simulations reveal multiple conformational states with one dominant state (**Fig. 4a-d**). These REMD simulations could serve as a "gold standard" to evaluate the effectiveness of the Boltzmann reweighting scheme. In order to compare with the REMD simulation results, we reweighted the last 100 ns of the OPLS-AA/L BE-META trajectories of the S1 and S2 simulations, and projected them onto the two PCs calculated from the REMD simulations. In **Fig. 4e, f**, the conformational density profiles of cNPF1 calculated from these processed BE-META trajectories are presented. We observe that the density profiles obtained from the reweighted BE-META trajectories are comparable to the REMD results. By dividing the 2D principal subspace into 200 × 200 grids and performing cluster analysis on the grids with data populations larger than 0.1, the populations for the dominant state are estimated to be 40.5±3.8% and 46.2±3.7% for the S1 and S2 simulations, respectively (error estimated by the standard deviation from the 18 BE-META trajectories) (**Fig. 4g, h**).

Performance of Different Force Fields on Modelling cNPF1. To evaluate the performance of different force fields on modelling cyclic peptides, we performed BE-META simulations for cNPF1 using five widely used biomolecular force fields (Amber96[42]+TIP3P[76], Amber99SB-ildn[46]+TIP3P, Amber03[44]+TIP3P, OPLS-AA/L[47]+TIP4P[76, 77], and GROMOS53a6[49]+SPC[78]) as well as a recently developed force field parameterized using the protein coil library (RSFF1[79]+TIP4P/Ew[80]). For each force field, two sets of simulations were performed, starting from the two input structures S1 and S2 (**Fig. 1b, c**). All simulations were 300 ns in length. To investigate the behaviour of the different force fields and simulation convergence, we performed dPCA on the last 100 ns of the twelve weighted BE-META trajectories (6 force fields × two input structures S1 and S2) simultaneously. Using the PC1 and PC2 thus obtained, 2D conformational density profile was constructed and cluster analysis performed for each simulation (**Fig. 5**).

As observed in **Fig. 5**, dPCA and cluster analysis results reveal multiple conformations for all the 6 force fields tested. Based on the similarity of the dPCA and cluster analysis results, we arranged the results from the 6 force fields in the order of GROMOS53a6, Amber96, Amber99SB-ildn, RSFF1, Amber03 and OPLS-AA/L from left to right in **Fig. 5**. We observe that the conformations sampled in the GROMOS53a6 simulations populate mostly the top right quadrant in the 2D principal space, and the conformations sampled gradually shift to the bottom right and middle left quadrants in a clock-wise fashion as one migrates to Amber96, Amber99SB-ildn, RSFF1, Amber03 and OPLS-AA/L force field (moving from left to right in **Fig. 5**). In **Fig. 6**, we show 100 conformations randomly selected from a number of most commonly populated states (states 1−5 as labelled in **Fig. 5c**). We first note that states 1−2 seem to form a narrower, more elongated cyclic configuration, while states 3−5 are wider, rounder and more similar to the NMR structure. State 4 has the smallest backbone RMSD to NMR (1.08 ± 0.16 Å) although noticeable deviation is observed around residues 5−8 (**Fig. 6d**); the number of violations to the experimental NOE restraints[51] is 28 ± 4, with 16 ± 4 of the violations being > 0.3 Å and 8 ± 3 of them > 1.0 Å.

To further analyze the most populated configurations sampled in these simulations, in **Fig. 7** we plot the φ/ψ distributions of the 8 residues in cNPF1 for each state. It seems that the first two PCs separate configurations mainly based on the φ/ψ of residues 5−8. All the 5 states share similar Ramachandran plots for residues 1−4 but have different Ramachandran plots for residues 5−8 (**Fig. 7**). Since the first several PCs correspond to the directions of largest variation, the separation of configurations of residues 5−8 by PC1 and PC2 implies that they are the most flexible residues in cNPF1. In **Fig. 6**, we observe that states 1−2 form a narrower, more elongated cyclic configuration, while states 3−5 are wider and rounder. This phenomenon seems to result from the different φ/ψ distribution of GLU-5 (**Fig. 7, column 5**). We observe that the φ/ψ distribution of GLU-5 for states 1−2 falls in the α-helix region (φ=-60°, ψ=-45°), while that for states 3−5 falls in the PPII/β region (φ=-75°, ψ=150°; φ=-135°, ψ=135°). The φ/ψ values of GLU-5 in the NMR structure (φ=-126°, ψ=26°, red dot in **Fig. 7**), nonetheless, are located in a high-free energy, rarely populated region in a typical Ramachandran plot, and are not captured by either states 1−2 or states 3−5. This observation suggests that the force fields tested may over-stabilize the α-helix and PPII/β regions, which are commonly populated by linear peptides. The poor performance of current peptide force fields at describing highly constrained cyclic peptides could be remedied by performing quantum calculations on a case-by-case basis.[81, 82] However, re-parameterization of a force field

that correctly describes the full Ramachandran plot would be of great interest and broad applicability. This may be achieved for example, by using the protein coil library (as adopted by the developers of the recent RSFF1 force field[79]), by including additional fitting points in the Ramachandran space from quantum chemistry calculations during force field development, or by including cyclic peptides as benchmarks during force field development.

Conclusions

Cyclic peptides are promising modulators of protein-protein interactions. The value of cyclic peptides as potential drugs would increase exponentially if we could accurately predict their conformations *de novo*. To achieve structure prediction of cyclic peptides, one must be able to sample cyclic peptide conformations efficiently. In the case of small cyclic peptides, structure sampling can be very challenging owing to their highly constrained conformations. In this paper we employed bias-exchange metadynamics simulations to enhance conformational sampling of a model cyclic peptide. Then, using a Boltzmann reweighting scheme we obtained an equilibrium structural ensemble. To characterize the structural ensemble of the cyclic peptide, we performed dihedral principal component analysis followed by density-peak based cluster analysis on the equilibrium structural ensemble.

To evaluate the performance of current peptide force fields on cyclic peptides, we simulated and characterized the structural ensemble for the model cyclic peptide using six popular peptide force fields (Amber96, Amber99SB-ildn, Amber03, GROMOS53a6, OPLS-AA/L, and RSFF1). Multiple conformations with significant populations were identified in all six force fields tested, in contrast to the experimental observation of a single highly populated structure by NMR spectroscopy. None of the conformations identified using the six force fields accurately recapitulates the NMR structure. Owing to the structural constraint, one or several residues in a small cyclic peptide like our model cyclic peptide likely populates a high free energy, rarely sampled region in a typical Ramachandran plot for linear peptides. All the six force fields tested seem to over-stabilize the α-helix and PPII/β regions, commonly populated by linear peptides. Our findings suggest that re-parameterization of a force field that correctly describes the full Ramachandran plot is necessary to accurately model cyclic peptides.

Acknowledgements

This research was sponsored by Tufts start-up fund and the Knez Family Faculty Investment Fund for Y.-S. L. We thank Professor Joshua Kritzer for helpful discussions, and Mr. Sean McHugh and Ms. Diana Slough for reading and providing useful comments on the manuscript.

Notes and references

[a] Department of Chemistry, Tufts University, Medford, Massachusetts 02155, United States. Email: yu-shan.lin@tufts.edu

† Electronic Supplementary Information (ESI) available: Potential energy distributions for each replica in the OPLS-AA/L REMD simulations, detailed dPCA and cluster analysis results, φ/ψ distributions for the states identified from the OPLS-AA/L REMD simulations, free energy profiles along each CV in all the BE-META simulations, and the φ/ψ distributions in the OPLS-AA/L BE-META simulations. See DOI: 10.1039/b000000x/

1. D. P. Ryan and J. M. Matthews, *Curr. Opin. Struct. Biol.*, **2005**, 15, 441-446.

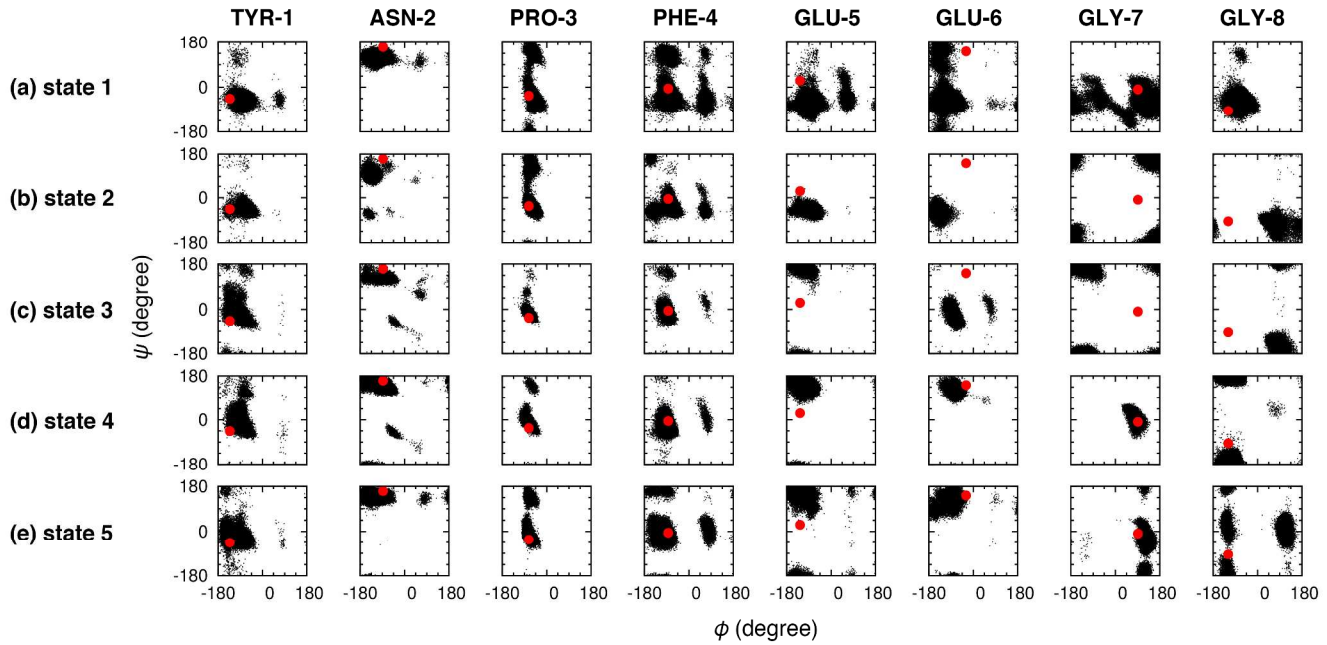

Fig. 7 φ/ψ distributions of cNPF1 residues in the 5 most commonly populated states. The φ/ψ angles in the NMR structure are shown as red dots.

2. S. Jones and J. M. Thornton, *Proc. Natl. Acad. Sci. USA*, **1996**, 93, 13-20.
3. L. L. Conte, C. Chothia and J. Janin, *J. Mol. Biol.*, **1999**, 285, 2177-2198.
4. J. A. Wells and C. L. McClendon, *Nature*, **2007**, 450, 1001-1009.
5. E. Biron, J. Chatterjee, O. Ovadia, D. Langenegger, J. Brueggen, D. Hoyer, H. A. Schmid, R. Jelinek, C. Gilon, A. Hoffman and H. Kessler, *Angew. Chem. Int. Ed.*, **2008**, 47, 2595-2599.
6. T. R. White, C. M. Renzelman, A. C. Rand, T. Rezai, C. M. McEwen, V. M. Gelev, R. A. Turner, R. G. Linington, S. S. Leung, A. S. Kalgutkar, J. N. Bauman, Y. Zhang, S. Liras, D. A. Price, A. M. Mathiowetz, M. P. Jacobson and R. S. Lokey, *Nat. Chem. Biol.*, **2011**, 7, 810-817.
7. J. G. Beck, J. Chatterjee, B. Laufer, M. U. Kiran, A. O. Frank, S. Neubauer, O. Ovadia, S. Greenberg, C. Gilon, A. Hoffman and H. Kessler, *J. Am. Chem. Soc.*, **2012**, 134, 12125-12133.
8. B. C. Buer, B. J. Levin and E. N. Marsh, *J. Am. Chem. Soc.*, **2012**, 134, 13027-13034.
9. L. Jin and S. C. Harrison, *Proc. Natl. Acad. Sci. USA*, **2002**, 99, 13522-13526.
10. G. F. Gause and M. G. Brazhnikova, *Nature*, **1944**, 154, 703.
11. P. J. Loll and P. H. Axelsen, *Annu. Rev. Biophys. Biomol. Struct.*, **2000**, 29, 265-289.
12. P. Sandhu, X. Xu, P. J. Bondiskey, S. K. Balani, M. L. Morris, Y. S. Tang, A. R. Miller and P. G. Pearson, *Antimicrob. Agents Chemother.*, **2004**, 48, 1272-1280.
13. V. Dewan, T. Liu, K. M. Chen, Z. Qian, Y. Xiao, L. Kleiman, K. V. Mahasenan, C. Li, H. Matsuo, D. Pei and K. Musier-Forsyth, *ACS Chem. Biol.*, **2012**, 7, 761-769.
14. S. Liu, W. Gu, D. Lo, X. Z. Ding, M. Ujiki, T. E. Adrian, G. A. Soff and R. B. Silverman, *J. Med. Chem.*, **2005**, 48, 3630-3638.
15. E. M. Driggers, S. P. Hale, J. Lee and N. K. Terrett, *Nat. Rev. Drug Discov.*, **2008**, 7, 608-624.
16. J. Chatterjee, D. Mierke and H. Kessler, *J. Am. Chem. Soc.*, **2006**, 128, 15164-15172.
17. T. Congdon, R. Notman and M. I. Gibson, *Biomacromolecules*, **2013**, 14, 1578-1586.
18. T. Rezai, J. E. Bock, M. V. Zhou, C. Kalyanaraman, R. S. Lokey and M. P. Jacobson, *J. Am. Chem. Soc.*, **2006**, 128, 14073-14080.
19. L. Corcilius, G. Santhakumar, R. S. Stone, C. J. Capicciotti, S. Joseph, J. M. Matthews, R. N. Ben and R. J. Payne, *Bioorg. Med. Chem.*, **2013**, 21, 3569-3581.
20. J. Beaufays, L. Lins, A. Thomas and R. Brasseur, *J. Pept. Sci.*, **2012**, 18, 17-24.
21. J. M. Damas, L. C. S. Filipe, S. R. R. Campos, D. Lousa, B. L. Victor, A. M. Baptista and C. M. Soares, *J. Chem. Theory Comput.*, **2013**, 9, 5148-5157.
22. A. M. Razavi, W. M. Wuest and V. A. Voelz, *J. Chem. Inf. Model.*, **2014**, 54, 1425-1432.
23. D. J. Drucker and M. A. Nauck, *The Lancet*, **2006**, 368, 1696-1705.
24. D. J. Earl and M. W. Deem, *Phys. Chem. Chem. Phys.*, **2005**, 7, 3910-3916.
25. J. Kästner, *WIREs Comput. Mol. Sci.*, **2011**, 1, 932-942.
26. S. K. Ramadugu, Y.-H. Chung, J. Xia and C. J. Margulis, *J. Phys. Chem. B*, **2009**, 113, 11003-11015.
27. M. M. Mackeen, A. Almond, M. Deschamps, I. Cumpstey, A. J. Fairbanks, C. Tsang, P. M. Rudd, T. D. Butters, R. A. Dwek and M. R. Wormald, *J. Mol. Biol.*, **2009**, 387, 335-347.
28. H. L. Malaby and W. R. Kobertz, *Biochemistry*, **2014**, 53, 4884-4893.
29. P. Mark, M. J. Baumann, J. M. Eklof, F. Gullfot, G. Michel, A. M. Kallas, T. T. Teeri, H. Brumer and M. Czjzek, *Proteins*, **2009**, 75, 820-836.
30. E. A. Villar, D. Beglov, S. Chennamadhavuni, J. A. Porco, Jr., D. Kozakov, S. Vajda and A. Whitty, *Nat. Chem. Biol.*, **2014**, 10, 723-731.
31. D. L. Rubin, N. H. Shah and N. F. Noy, *Brief. Bioinform.*, **2008**, 9, 75-90.
32. G. K. Nguyen, S. Wang, Y. Qiu, X. Hemu, Y. Lian and J. P. Tam, *Nat. Chem. Biol.*, **2014**, 10, 732-738.
33. A. P. Willard and D. Chandler, **2014**.
34. D. Xu, E. I. Newhouse, R. E. Amaro, H. C. Pao, L. S. Cheng, P. R. Markwick, J. A. McCammon, W. W. Li and P. W. Arzberger, *J. Mol. Biol.*, **2009**, 387, 465-491.
35. S. Piana and A. Laio, *J. Phys. Chem. B*, **2007**, 111, 4553-4559.
36. Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, **1999**, 314, 141-151.
37. A. Patriksson and D. van der Spoel, *Phys. Chem. Chem. Phys.*, **2008**, 10, 2073-2077.
38. A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. USA*, **2002**, 99, 12562-12566.
39. N. Todorova, F. Marinelli, S. Piana and I. Yarovsky, *J. Phys. Chem. B*, **2009**, 113, 3556-3564.
40. F. Baftizadeh, P. Cossio, F. Pietrucci and A. Laio, *Curr. Phys. Chem.*, **2012**, 2, 79-91.
41. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, **1995**, 117, 5179-5197.
42. P. A. Kollman, *Acc. Chem. Res.*, **1996**, 29, 461-469.
43. J. Wang, P. Cieplak and P. A. Kollman, *J. Comput. Chem.*, **2000**, 21, 1049-1074.
44. Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang and P. A. Kollman, *J. Comput. Chem.*, **2003**, 24, 1999-2012.
45. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, **2006**, 65, 712-725.
46. K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins*, **2010**, 78, 1950-1958.
47. G. A. Kaminski and R. A. Friesner, *J. Phys. Chem. B*, **2001**, 105, 6474-6487.
48. L. D. Schuler, X. Daura and W. F. van Gunsteren, *J. Comput. Chem.*, **2001**, 22, 1205-1218.
49. C. Oostenbrink, A. Villa, A. E. Mark and W. F. van Gunsteren, *J. Comput. Chem.*, **2004**, 25, 1656-1676.
50. N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark and W. F. van Gunsteren, *Eur. Biophys. J.*, **2011**, 40, 843-856.
51. A. J. Kamens, R. J. Eisert, T. Corlin, J. D. Baleja and J. A. Kritzer, *Biochemistry*, **2014**, 53, 4758-4760.
52. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, **2004**, 25, 1605-1612.

53. B. Hess, C. Kutzner, D. Van Der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, **2008**, 4, 435-447.
54. G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, **2007**, 126, 014101.
55. A. Cheng and K. M. J. Merz, *J. Phys. Chem.*, **1996**, 100, 1927-1937.
56. M. Lingenheil, R. Denschlag, R. Reichold and P. Tavan, *J. Chem. Theory Comput.*, **2008**, 4, 1293-1306.
57. A. Mor, G. Ziv and Y. Levy, *J. Comput. Chem.*, **2008**, 29, 1992-1998.
58. M. Parrinello and A. Rahman, *J. Appl. Phys.*, **1981**, 52, 7182-7190.
59. R. W. Hockney, S. P. Goel and J. W. Eastwood, *J. Comput. Phys.*, **1974**, 14, 148-158.
60. B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, **1997**, 18, 1463-1472.
61. U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J. Chem. Phys.*, **1995**, 103, 8577-8593.
62. M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*, Clarendon press, Oxford, 1987.
63. G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni and G. Bussi, *Comput. Phys. Commun.*, **2014**, 185, 604-613.
64. F. Marinelli, F. Pietrucci, A. Laio and S. Piana, *PLoS Comput. Biol.*, **2009**, 5, e1000452.
65. T. Ichiye and M. Karplus, *Proteins*, **1991**, 11, 205-217.
66. A. E. Garcia, *Phys. Rev. Lett.*, **1992**, 68, 2696-2699.
67. A. Amadei, A. B. Linssen and H. J. Berendsen, *Proteins*, **1993**, 17, 412-425.
68. A. Kitao and N. Go, *Curr. Opin. Struct. Biol.*, **1999**, 9, 164-169.
69. B. L. de Groot, X. Daura, A. E. Mark and H. Grubmuller, *J. Mol. Biol.*, **2001**, 309, 299-313.
70. Y. Mu, P. H. Nguyen and G. Stock, *Proteins*, **2005**, 58, 45-52.
71. A. Jain, R. Hegger and G. Stock, *J. Phys. Chem. Lett.*, **2010**, 1, 2769-2773.
72. S. R. R. Campos, M. Machuqueiro and A. M. Baptista, *J. Phys. Chem. B*, **2010**, 114, 12692-12700.
73. S. R. R. Campos and A. M. Baptista, *J. Phys. Chem. B*, **2009**, 113, 15989-16001.
74. F. Sittel, A. Jain and G. Stock, *J. Chem. Phys.*, **2014**, 141, 014111.
75. A. Rodriguez and A. Laio, *Science*, **2014**, 344, 1492-1496.
76. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, **1983**, 79, 926-935.
77. H. C. Tang and C. Y. Chen, *Evid. Based Complement. Alternat. Med.*, **2014**, 2014, 385120.
78. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren and J. Hermans, in *Intermolecular forces*, ed. B. Pullman, 1981, p. 331.
79. F. Jiang, C. Y. Zhou and Y. D. Wu, *J. Phys. Chem. B*, **2014**, 118, 6983-6998.
80. H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, *J. Chem. Phys.*, **2004**, 120, 9665-9678.
81. G. L. Butterfoss, B. Yoo, J. N. Jaworski, I. Chorny, K. A. Dill, R. N. Zuckermann, R. Bonneau, K. Kirshenbaum and V. A. Voelz, *Proc. Natl. Acad. Sci. USA*, **2012**, 109, 14320-14325.
82. V. A. Voelz and G. Zhou, *J. Comput. Chem.*, **2014**, 35, 2215-2224.