Frontiers in Molecular Biosciences: Hybrid Biomolecular Modeling

On the Helix Propensity in Generalized Born Solvent Descriptions of Modeling the Dark Proteome

Mark A. Olson*

Department of Cell Biology and Biochemistry, Molecular and Translational Sciences Division, USAMRIID, Fredrick, Maryland, United States of America

Abstract: Intrinsically disordered protein regions that populate the so-called "Dark Proteome" offer challenging benchmarks of conformational sampling methods and their all-atom force fields plus solvent descriptions to accurately model structural transitions on a multidimensional energy landscape. This work explores the application of parallel tempering methods with implicit solvent models as a computational framework to capture the conformational ensemble of an intrinsically disordered peptide derived from the Ebola virus protein VP35. A recent X-ray crystallographic study reported a protein-peptide interface where the VP35 peptide underwent a folding transition from a disordered form to a helix-β-turn-helix fold upon molecular association with the Ebola protein NP. An assessment is provided of the accuracy of two generalized Born solvent models (GBMV2 and GBSW2) using the CHARMM force field and applied with temperature-based replica exchange dynamics to calculate the disorder propensity of the peptide and its probability density of states in a continuum solvent. A further comparison is presented of applying an explicit/implicit solvent hybrid replica exchange simulation of the peptide.

Keywords: molecular dynamics, free-energy landscape, intrinsically disordered proteins, explicit/implicit solvent model, replica exchange simulations

*Email: mark.a.olson1.civ@mail.mil

1. INTRODUCTION

The large conformational heterogeneity and rapid dynamic transitions of intrinsically disordered peptides and proteins (IDPs) present a challenge to experimental boundaries in characterizing their functional form on rugged energy landscapes (Wright and Dyson, 1999; Wright and Dyson, 2005). From a biological perspective, the broad interest in IDPs draws principally from their fundamental role in the regulation and function of cellular protein networks. Recent experimental studies have begun to unravel the interplay between "ordered chaos" of IDPs and the kinetic transition to a topological funnel of folded states (Arai et al., 2015). The contemporary view of this dynamic transition is a process that occurs by either an "induced-fit" of the IDP upon molecular association with a protein target or by target "fly casting" of a pre-folded state in the disordered conformational ensemble of the IDP (see, e.g., Shoemaker et al., 2000; Arai et al., 2015).

Complementary to experimental studies are computer simulations which offer a powerful set of tools to understand IDPs at the all-atom level and their inherent plasticity to navigate a disordered network of microstates (see, e.g., Lee and Chen, 2016; Bhowmick et al., 2016; Chebaro et al., 2015; Zhang and Chen, 2014). Among the simulation methods, the generalized ensemble sampling method of temperature-based replica exchange (T-ReX) (Ishikawa et al., 2001; Sugitaa and Okamoto, 1999), also known as parallel tempering, has become an increasingly popular approach for exploring the energy landscape of proteins. Algorithms combined with T-ReX to generate protein configurations vary in their theoretical formulations and range from canonical molecular dynamics (MD) simulations to methods that accelerate conformational sampling. Of the latter, examples includes coarse replica-exchange molecular dynamics (Peter et al., 2016), accelerated molecular dynamics (see, e.g., Miao et al., 2015), Hamiltonian switch Metropolis Monte Carlo (Mittal et al., 2014), all-atom multicanonical molecular dynamics (Higo et al., 2011) and self-guided Langevin dynamics (SGLD) (Wu and Brooks, 2003), among others.

A computational strategy of reducing the complexity of all-atom simulations of proteins is the replacement of explicit water interactions with a continuum description of treating implicitly the bulk physical properties of solvation effects. The most common implicit solvent method for protein dynamic simulations is the generalized Born (GB) approximation. GB models are computationally faster than explicit solvent calculations and differ in their accuracy

of reproducing Poisson-Boltzmann solvation energies for single protein conformations (see, e.g., Feig et al., 2003). Application of GB solvent models to studies of IDPs has been reported by several laboratories (see, e.g., Ganguly and Chen, 2009; Click et al., 2010; Ganguly and Chen, 2015; Chebaro et al., 2015). To date the simulation results lack consensus on the accuracy of GB solvent models as a computational framework to capture the fold propensities of IDPs and their probability density of states on a conformational landscape. Particularly missing among the reported studies are comparative assessments of GB models of IDPs with those modeled by explicit all-atom solvent replica exchange simulations.

Given the current interests in characterizing the so-called "Dark Proteome" which consists of "invisible" conformational states within the human, viral and microbial genomes (Bhowmick et al., 2016; Perdigãoa et al., 2015), this work presents temperature-based replica exchange simulations of modeling an IDP derived from an Ebola virus protein. Ebola viruses are nonsegmented negative sense RNA viruses that cause severe hemorrhagic fever (Sanchez et al., 2006). An X-ray crystallographic structure was reported by Amarasinghe and coworkers (Leung et al., 2015) of the Ebola nucleoprotein NP in complex with a 28-residue peptide extracted from Ebola VP35 (designated as the NPBP peptide). The NP-VP35 viral assembly is essential for virus replication and offers a protein target for therapeutic development. Experimental data reveals the NPBP peptide binds NP with high affinity and specificity, and acts by blocking NP oligomerization. The peptide undergoes a folding transition from a disordered form free in solution to a helix-β-turn-helix fold upon molecular association with NP (Leung et al., 2015).

Two different generalized ensemble sampling methods are applied based on combining T-ReX with the SGLD simulation method (Lee and Olson, 2010) and two different GB solvent models (Lee et al., 2003; Im et al., 2003) are examined to assess their accuracy in modeling the density of states of the NPBP peptide. One of the sampling techniques is the conventional application of T-ReX with a static set of temperatures. The other technique is an adaptive T-ReX where the replica clients dynamically walk in temperature space in search of the optimal population density on a modeled energy function (Trebst et al., 2006; Katzgraber et al., 2006; Lee and Olson, 2011; Olson and Lee, 2014; Olson et al., 2016). The two GB models differ in their dielectric-boundary descriptions with one of them constructed from an analytic formulation of the molecular volume (Lee et al., 2003).

The final simulation model applied to the NPBP peptide is an explicit/implicit solvent hybrid T-ReX/MD method (Chaudhury et al., 2012). The application of this simulation model is to investigate the effect of solvent modeling resolution on the helix propensity and the search of conformational transitions. The idea behind the hybrid model is reducing the number of replica clients needed in explicit solvent simulations by replacing the contribution of explicit solvent energies in the Metropolis exchanges (Metropolis et al., 1953) with those of an accurate GB solvent approximation. The hybrid model allows the same number of replica clients to be applied as in the GB solvent T-ReX/SGLD simulations of the NPBP peptide while retaining a higher resolution in conformational sampling on an explicit solvent landscape (Chaudhury et al., 2012; Olson and Lee, 2013).

2. COMPUTATIONAL METHODS

This section provides a brief outline of the computational methods applied in this work of modeling the NPBP peptide taken from the PDB 4YPI (Figure 1). A general approach for conformational sampling is the application of T-ReX (see, e.g., Ishikawa et al., 2001). Unlike the well-established method of MD simulations at a single sampling temperature, T-ReX is a generalized ensemble method of applying multiple parallel simulations in which each replica is executed at a different temperature. In traditional applications of T-ReX, the temperatures (T_1, T_2, …, T_n), where n is the number of replica clients, are pre-determined by a static (fixed) set of values that span a desired range. It is common to model the set of temperatures by a geometrically spaced sequence (Predescu et al., 2004) using $n-1$ intervals from the minimum temperature denoted as $T_1 = T_{\min}$ to the maximum $T_n = T_{\max}$

$$T_{i+1} = T_i \left(T_{\max} / T_{\min} \right)^{\left[\frac{1}{n-1} \right]}, \qquad (1)$$

where T_i the temperature of the ith replica client in Figure 1.

An alternative to Eq. (1) is an adaptive replica exchange method of allowing the clients to dynamically walk in temperature space (Trebst et al., 2006; Katzgraber et al., 2006; Lee and Olson, 2011; Olson and Lee, 2014; Olson et al., 2016). In implementing the adaptive algorithm, each client is tagged as either “cold” or “hot” depending on the last temperature extreme it visited (Lee and Olson, 2011). Tracing of the clients is made by constructing histograms over

temperature space, $n_{\text{cold}}(T)$ and $n_{\text{hot}}(T)$, where each bin accumulates the number of cold and hot clients, respectively, visiting each temperature window. The fraction cold, *f*, of a client window at temperature *T* is the number of cold clients visiting that temperature divided by the total number of cold and hot client visits:

$$f(T) = \frac{n_{\text{cold}}(T)}{n_{\text{cold}}(T) + n_{\text{hot}}(T)}. \qquad (2)$$

Using the $f(T)$ term, a thermal current is defined (Lee and Olson, 2011)

$$j = D(T)\eta(T)\frac{df}{dT}, \qquad (3)$$

where $D(T)$ is the diffusivity and $\eta(T)$ is the probability that any client will reside at temperature *T*. The thermal current can be maximized by adjusting the temperatures such that $f(T_i)$ increases linearly as a function of temperature index, *i*. Here, a continuous $f(T)$ is constructed from the computed values of *f* at the current set of temperatures, T_i, and then search for the new temperatures where $f(T) = i/(N\text{-}1)$. To prevent all of the windows from clustering around the same temperature and depleting exchanges at the extremes, a constraint is applied where no neighboring temperatures can be more than two geometric spacing units apart,

$$\frac{T_{i+1}}{T} \le \left(\frac{T_{\max}}{T_{\min}}\right)^{\left[\frac{2}{N-1}\right]} \qquad (4)$$

with the lower and upper values of T_i set to $T_{\min}$ and $T_{\max}$, respectively.

The exchange of temperatures between neighboring replica clients, *a* and *b*, is determined by a probability given by the Metropolis energy criteria (Metropolis et al., 1953)

$$p(a \leftrightarrow b) = \min\left[1, e^{(\beta_a - \beta_b)(E_b - E_a)}\right], \qquad (5)$$

where $\beta_a = 1/k_B T_a$, k_B is Boltzmann's constant ,T_a is the temperature of replica client *a*, and E_a is the potential energy of client *a*.

SGLD Simulation Models

For generating trajectories of the NPBP peptide, two methods were combined with T-ReX. The first is based on the SGLD simulation method developed by Wu and Brooks (2003). The SGLD equation of motion is given by

$$\dot{\mathbf{p}}_i = \mathbf{f}_i - \gamma_i \mathbf{p}_i + \mathbf{R}_i + \lambda \mathbf{g}_i, \qquad (6)$$

where $\dot{\mathbf{p}}_i$ defines the rate of change of the momentum of particle *i*, $\mathbf{f}_i$ is the force acting on the particle, γ_i is the friction constant, $\mathbf{R}_i$ defines the random force and $\mathbf{g}_i$ is a memory function, which is scaled by an *ad hoc* guiding factor λ. The memory function $\mathbf{g}_i$ is defined by the moving average of the momentum seen by the system over an interval of time, *L*:

$$\mathbf{g}_i = \left\langle \mathbf{p}_i \right\rangle_L, \qquad (7)$$

where $\langle \ldots \rangle_L$ denotes a local average. The time interval is further defined as $L = t_L / \delta t$, where t_L is the local averaging time and δt the time step along the simulation trajectory. It should be noted that because of the *ad hoc* force in Eq. (6), the sampling algorithm deviates from a canonical ensemble (Lee and Olson, 2010; Wu and Brooks, 2011; Wu et al., 2012; Wu et al., 2016). For this work, the deviation is anticipated to be small for modeling a peptide (Lee and Olson, 2010), nevertheless the sampling distributions can be reweighted to remove the applied bias (Wu and Brooks, 2011).

In the SGLD simulations, solvent was represented by implicit solvent models GBMV2 (generalized Born molecular volume version 2) (Lee et al., 2002; Lee et al., 2003) and the GBSW2 (generalized Born smoothing window version 2) (Im et al, 2003). The most noted difference between the two models is representation of the solvent excluded volume and the treatment of the dielectric interface. The GBMV2 parameters were selected to smooth the molecular volume by setting β = –12 and P3 = 0.65 (Yeh et al., 2008). The hydrophobic cavitation term was modeled by applying a phenomenological surface tension coefficient set to a value of 0.015 kcal/mol/Å^2. For applying GBSW2, the model was parameterized to fit the Lee-Richards molecular-surface Poisson results and required w = 0.2 Å, a_0 = 1.2045 and a_1 = 0.1866. The hydrophobic cavitation-energy tension term was set to 0.030 kcal/(mol·Å^2).

The utilities and programming libraries of the Multiscale Modeling Tools for Structural Biology (MMTSB) (Feig et al., 2004) were used to carry out the T-ReX/SGLD simulations. The CHARMM simulation program (version c35b2) was applied as a modeling platform (Brooks et al., 2009). Simulations were carried out using 24 replica clients and the frequency of exchanges was set to every 1 ps of simulation. Temperatures were set where $T_{\min}$ = 300 K and $T_{\max}$ = 475 K. Because the implicit solvent models GBMV2 and GBSW2 were originally developed for

and have been extensively benchmarked with the CHARMM22 force field, this force field was applied with the CMAP backbone dihedral cross-term extension (Mackerell et al., 2004). An integration time step of 2 fs was used and parameters for SGLD consisted of the friction constant set to γ of 1 ps^{-1} for all heavy atoms, the guiding factor λ to a value of 1, and the averaging time t_L was set to 1 ps. These values were taken from our previous studies of the SGLD model (Lee and Olson, 2010; Lee and Olson, 2011; Olson and Lee, 2014). Non-bonded interaction cutoff parameters for electrostatics and vdW terms were set at a radius of 22 Å with a 2-Å potential switching function. Covalent bonds between the heavy atoms and hydrogen atoms were constrained by the SHAKE algorithm (Ryckaert et al., 1977). The NPBP peptide was modeled for 200 ns of simulation time per client, generating an ensemble of 4.8 μs.

Hybrid Simulation Model

The second method applied for generating trajectories of the NPBP peptide is an explicit/implicit solvent hybrid T-ReX/MD simulation (Chaudhury et al., 2012). In a typical explicit solvent T-ReX simulation the energies are given by

$$E_{\text{explicit}} = U_{\text{all-atom}}^{\text{prot}} + U_{\text{all-atom}}^{\text{prot-solv}} + U_{\text{all-atom}}^{\text{solv-solv}}, \qquad (8)$$

where the first term describes the protein potential energy for a CHARMM-based molecular mechanics force field, the second term is the explicit protein-solvent interactions followed by the explicit solvent-solvent interactions. The all-atom solvent-solvent energy term requires significant number of replica exchange clients to achieve adequate Metropolis updates (Chaudhury et al., 2012). In the hybrid T-ReX method, the dynamics of each replica moves on an explicit solvent landscape. During a Metropolis update, all waters are removed from a replica and the solvent energy term of the replica is calculated using the grid-based GBMV2 solvent model

$$E_{\text{implicit}} = U_{\text{all-atom}}^{\text{prot}} + \Delta G_{\text{GBMV2}}^{\text{prot-solv}}, \qquad (9)$$

where $\Delta G_{\text{GBMV2}}^{\text{prot-solv}}$ is the free-energy term due to the implicit solvent contribution.

The NAMD code (Phillips et al., 2005) was applied for the 200-ns T-ReX/MD simulation with the CHARMM22+CMAP force field. The simulation cubic box size was set to 53.19 $Å^3$ and the number of waters was 4796. For modeling the waters the TIP3P potential was

applied (Jorgensen et al., 1983). Nose´–Hoover thermostat was applied with a temperature coupling constant of 50 kcal/s^2. Since the additional computational expense of the hybrid model relative to implicit solvent calculations, the NAMD simulation parameters differ slightly from the T-ReX/SGLD simulations in that a smaller cutoff distance of 12 Å was applied with a switching distance of 8 Å. The integration time step remained identical to that used with the SGLD simulations and the SHAKE algorithm was similarly applied. Particle mesh Ewald was applied and combined with periodic boundary conditions.

Evaluation Metrics

To examine the trajectories generated by the simulations, the weighted histogram analysis method (WHAM) (Ferrenberg and Swendsen, 1989; Kumar et al., 1992; Shea et al., 1998; Gallicchio et al., 2005) was applied to the data sets. The 2D density of states, $\Omega(q_1,q_2)$, for a molecular system, where q_1 and q_2 are a set of reaction coordinates of interest, is given by

$$\Omega(q_1,q_2)=\frac{\sum_{i=1}^{R} N_i(q_1,q_2)}{\sum_{j=1}^{R} n_i \exp(f_i-\beta_i E)}, \qquad (10)$$

where n_j is the number of data points in the jth simulation and β_j and T_j are Boltzmann's constant and temperature of the jth simulation, respectively. The function $N_i(q_1,q_2)$ is the histogram of (q_1,q_2) calculated from the ith simulation, and f_j is the scaled free energy obtained by solving the following equations self-consistently,

$$P_\beta(q_1,q_2)=\frac{\sum_{i=1}^{R} N_i(q_1,q_2)\exp(-\beta E)}{\sum_{j=1}^{R} n_i \exp(f_i-\beta_i E)} \qquad (11)$$

and

$$\exp(-f_i)=\sum_{q_1,q_2}\Omega(q_1,q_2)\exp(-\beta E), \qquad (12)$$

where $P_\beta(q_1,q_2)$ is the probability density at the inverse temperature β. For calculations presented here, q_1 = fractional helicity (f_H) of the peptide determined from DSSP (Kabsch and Sanders, 1983) and q_2 = radius of gyration (R_g).

The trajectories were further analyzed by a *Q* score for the peptide. *Q* is the number of side-chain contacts in a generated conformation divided by the total number equivalent contacts in the X-ray crystal structure of NPBP. Values were computed for side-chain center-of-mass pairs (*i*,*j*), such that $j > i$ and whose distances are less than a cutoff of 4.2 Å. A sigmoidal function was applied (implemented in MMTSB) to effectively include residue pairs that are slightly further apart with a reduced weight. In addition to a *Q* score, pairwise Cα root-mean-square-deviation (RMSD) from the starting X-ray structure was computed for each peptide structure in a generated ensemble of conformations.

3. RESULTS AND DISCUSSION

Figure 2 illustrates the X-ray crystallographic structure of the NPBP peptide extracted from the Ebola virus VP35 in association with the Ebola NP protein (Leung et al., 2015). The binding of NPBP occupies a functionally critical site on NP required for RNA synthesis and the peptide conformation is stabilized by a network of electrostatic interactions dominated by NP residues Arg240, Lys248, and Asp252. Using the DSSP secondary structure algorithm, NPBP (annotated as residues 20-47) shows segments Trp28 to Thr35 and Val40 to Asp42 as distinct helical conformations. The overall f_H is 0.4 and the bound form exhibits an R_g of 10.5 Å.

Experimental characterization of the secondary structure of the NPBP peptide free in solution by circular dichroism (CD) spectroscopy is reported to show the peptide as intrinsically disordered (Leung et al., 2015). When added to a solution of 50 % trifluoroethanol (TFE), the NPBP peptide transitions from a coil to helical structures of approximately 30-40 % helicity, thus suggesting a strong underlying secondary-structure propensity. Predictions of secondary-structure without bias of the crystallographic structure estimate the NPBP peptide to encompass a consensus f_H ~0.3 with probabilities greater than 0.9 for helical formation in the sequence segment of Gly27 to Met34 (see, e.g., Kieslich et al., 2016).

To examine the accuracy of implicit solvent models to counterbalance the network of electrostatic interactions of the viral assembly interface that contribute to the stabilization of the NPBP helical fold and produce a conformational landscape with a predisposed helix propensity in bulk water, replica-exchange simulations were performed using different simulation strategies. The conformational sampling approach of SGLD was explored with two different GB solvent

models and two different temperature-based replica-exchange methods. The first simulation model result shown in Figure 2b is the SGLD-GBMV2 with a static (fixed) set of temperatures in defining the replica-exchange protocol. The 2D free-energy profile $\Omega(f_H,R_g)$ computed at 300 K using WHAM of the full ensemble shows a large manifold of conformational substates with a helix distribution of $f_H \sim 0$ to 0.5. Several representative structures extracted from the basins are illustrated in Figure 2e. The conformational density takes place in R_g space of roughly 8-11 Å and at the lower end of the population distribution non-structured states are observed to occupy a large range of R_g values and show the canonical feature of disorder.

Given the broad population distribution produced by a static set of temperatures in the T-ReX simulations, it is important to test whether the simulation model provided adequate sampling of the basins. To address this issue, an adaptive replica-exchange SGLD-GBMV2 simulation model was applied whereby the clients walk in temperature space to optimize the efficiency of exchanges between nearest-neighbor thermal windows at potential energy barriers separating conformational states (Olson et al., 2016; Olson and Lee, 2014; Lee and Olson, 2011). The 2D profile from the adaptive T-ReX is illustrated in Figure 2c and the result is shown to retain the manifold of transient states of those sampled by the static T-ReX method, yet a population shift is observed toward an $f_H \sim 0.5$ at the cost of reducing the density of unstructured conformations. The theoretical goal of the adaptive method is to enhance sampling of conformational transitions for a modeled potential energy surface. Early success of the method applied to a sharp phase transition of unfolding-folding of the protein SH3 showed better agreement with the experimental melting temperature than the traditional static approach calculated over an identical length of simulation time (Lee and Olson, 2011). The adaptive method also captured with greater accuracy the native state of SH3 extracted from the conformational ensemble. Given these prior outcomes, and while the NPBP certainly lacks the folding cooperativity of SH3, the result suggests for CHARMM22+CMAP/GBMV2 a "native" state of helix propensity near the value observed experimentally for the crystallographic bound conformation and one which is inconsistent with the CD analysis in free solution. Because the potential energy surface is the same between the static and adaptive T-ReX methods, the less-efficient sampling approach will eventually converge to find a comparable $\Omega(f_H,R_g)$.

To determine the bias of the GBMV2 solvent approximation on $\Omega(f_H,R_g)$, adaptive T-ReX simulations were performed with a different implicit solvent model based on the GBSW2

approximation. Of the GB-based solvent models developed for protein dynamics, GBMV2 is one of the most accurate models in reproducing Poisson-Boltzmann theory with a Lee-Richards molecular surface (Feig et al., 2003). The basis of GBMV2 is an analytic formulation of the molecular volume (Lee et al., 2003), whereas the less accurate but computationally much faster GBSW2 model is based on a smooth dielectric-boundary formulation constructed by applying a superposition of atomic-centered polynomials (Im et al., 2003). The dissimilarities between the two models are clearly illustrated in Figure 2d. Application of GBSW2 significantly reduces the number of high-probability conformational excursions and leads to a folding funnel at f_H ~0.5. While the "optimized" f_H from the two different implicit solvent models is surprisingly similar, the limited disorder from the GBSW2 model in its current parameterized implementation makes this solvent approximation less suitable for modeling IDPs (for an alternative parameterization of GBSW, see, e.g., Chen 2010).

Figure 3 shows the probabilities of observing R_g as a function of three ensemble sampling temperatures. The GBMV2 model produced more compact states of NPBP than the crystallographic bound form, while GBSW2 yielded R_g values less collapsed. This observation can be partially attributed to the distinction in molecular surface representations between the solvent models, where different weights are applied to the surface-tension term that describes the hydrophobic free energy. In general, MD simulations of unfolded states are more compact and tend to favor helical structures than those found experimentally (Piana et al., 2014). By example, an experimental R_g for a unfolded 28 amino acids is estimated to be 13 Å (Kohn et al., 2004).

Also shown in Figure 3 are the probability profiles of C_α-RMSD and the fraction of side-chain contacts similar to the starting conformation of NPBP. The ensemble average over contacts is denoted as $\langle Q \rangle$ and values less than 0.6 are considered unrelated to the starting structure. When combined with the analysis of the 2D profiles, the probabilities provide an interesting picture of the rare event of recognizing a peptide conformation in the ensemble that is similar to the NPBP bound form. For the GBMV2 model and considering only the last 50 ns of simulation time, the lowest RMSD is 2.9 Å with $Q = 0.6$, and is clustered in the outer periphery of the highly-populated basin labeled as III in Figure 2c. This sparse cluster of low RMSD states emerges with an f_H of 0.5 and R_g approaching 10 Å.

It is also important to understand the configurational stability of IDPs from simulations and their helix propensities. The thermal unfolding profiles for NPBP are shown in Figure 4a. Consistent with the reduced number of transient states and their populations among the GB models, GBSW2 retains helicity over a greater thermal range. The congregation of replica clients in the range of 360 K to 425 K for the adaptive method (GBMV2 and GBSW2) is the effect of enhanced sampling of transition points that stabilize helix formation. The statistical errors in the histograms for all model simulations are approximately $f_H \pm 0.1$ along the temperature contour. Simulation convergence and the dominance of helix formation in NPBP can be further tested by conducting T-ReX simulations starting from a random coil state rather than the bound helix-β-turn-helix conformation. Although these latter simulations were executed only to 100 ns using the adaptive method, Figure 4b shows convergence to a folded state of helical conformations and establishes the strong helix propensity of applying the GB solvent descriptions.

The overweighting of secondary structure biases from the GBMV2 and GBSW2 solvent models is comparable to other studies of using different GB solvent models and parameterizations (Chebaro et al., 2015; Ganguly and Chen, 2009; Click et al., 2010). As a further test of the impact of the GBMV2 solvent model and its mean-field resolution of smearing out the details of the solvent on sampling conformational transitions of NPBP, the final simulation model tested is an explicit/implicit solvent hybrid T-ReX/MD method. This model generates peptide configurations on an explicit solvent (TIP3P) landscape while using the same number of replica clients as in the implicit solvent calculations. The latter is achieved by using GBMV2 in the Metropolis exchanges rather than explicit solvent. It is worth noting that, while it is not the goal to determine unconstrained folding free energies to high accuracy, replacement of energies from an all-atom representation to a mean-field approximation can produce errors in the detailed balance required of a canonical ensemble (Chaudhury et al., 2012).

Figure 5 shows $\Omega(f_H,R_g)$ at 300 K from the WHAM calculation of the hybrid simulation model ensemble and the thermal unfolding profile. Several important observations can be made in comparison to the static GBMV2 model which best corresponds to the non-adaptive hybrid model. The most important distinction between the results is the striking difference in the favorable free energies and their network that shuttle conformations among the helical basins. While the hybrid and GBMV2 model show sufficient plasticity among the states, the hybrid

model shows a free-energy minimum at a slightly lower f_H = 0.26 *vs.* 0.37, and yields good agreement with secondary-structure predictions. The distinction in the potentials of mean force among the models can be illustrated by considering a transition between an unstructured state and the free-energy minimum. For the static GBMV2, the transition (f_H = 0; R_g = 11 Å) → (f_H = 0.37; R_g = 8 Å) yields ΔΔG = 0.1 kcal/mol, whereas for the adaptive model the transition from the same disordered state → (f_H = 0.47; R_g = 9 Å) ΔΔG = 1.0 kcal/mol, and for the hybrid model the transition → (f_H = 0.26; R_g = 9 Å) ΔΔG = 1.7 kcal/mol. While the static model exhibits a reversible transition to unstructured states and would appear to be in better agreement with the CD experiments (Leung et al., 2015), enhanced sampling of $\Omega(f_H,R_g)$ by the adaptive method for this solvent description revealed a more costly transition to the densely populated f_H ~0.5.

The lowest RMSD conformer for the hybrid model via the last 50 ns is 3.3 Å with Q = 0.6 and R_g = 9.4 Å. This conformer is illustrated in Figure 5b as the first structure depicted for the basin labeled III. The conformation is formed from a helical hairpin of residues Ser26-Met34 and Val40-Phe44. The top-rank conformer based on potential energies for the free-energy minimum at f_H = 0.26 is illustrated as the first structure for basin I. This structure shows a 5-residue helix of Trp28-Met34. Among the highly populated sampled basins, a distinction between the simulation models is the cluster at f_H = ~0.6, where the hybrid model shows an enhanced free energy of population. Unlike the other basins, this basin lacks a direct low-energy pathway along the manifold of states.

A statistical average of the ensemble for the hybrid model computed from the multiple temperatures of the T-ReX simulation is illustrated in Figure 5c along with a comparison with the static GBMV2 model. Despite the differences in the $\Omega(f_H,R_g)$ profiles between the models, a simple statistical average without reweighting based on free energies shows remarkably similar f_H values at 300 K. Because of the lack of instantaneous relaxation of the explicit waters in contrast to GB approximations, the hybrid model shows more confined excursions of unfolded states at the upper R_g boundaries. Like many MD simulations of unfolded states with explicit solvent (Piana et al., 2014), a residual secondary-structure propensity is observed at 475 K.

The more compact favorable states observed in the explicit/implicit solvent hybrid model than that corresponding to the bound NPBP conformation is unlikely due to the GB model, but rather the additive force field (Piana et al., 2014). As noted above, the CHARMM22+CMAP force field was selected because of extensive benchmarks in reported studies of the GBMV2 and

GBSW2 solvent descriptions to successfully model natively folded structures of proteins (see, e.g., Yeh et al., 2008). While there are no reported studies of applying either GBMV2 or GBSW2 with the more refined CHARMM36m force field and its parameterization for TIP4P-based explicit solvent simulations (Huang et al., 2016), switching to this description may help reconcile the underestimated R_g values with those experimentally determined for unfolded states and reduce the overall weight and stabilization of helix propensities.

4. CONCLUSIONS

The current initiative to develop an atomistic understanding of "invisible" conformational states of the human/viral/bacterial proteomes requires an accurate computational framework for modeling conformational transitions within a disordered ensemble and their population density. The work presented here examined the application of temperature-based replica exchange simulations with different sampling methods and solvent descriptions of modeling an intrinsically disorder 28-residue peptide from the Ebola virus protein VP35. The X-ray crystallographic conformation of the VP35 peptide bound to Ebola NP reports a helix-β-turn-helix fold of roughly 40 % helical structure, whereas in free solution the peptide is unstructured. The simulations of the unbound peptide showed the selection of a GB solvent model combined with a replica-exchange sampling protocol can have a significant effect on the sampling populations. Overall, the tested GB models tend to favor a free-energy minimum of roughly 50 % helical content for the peptide. The effect of an adaptive temperature-based replica exchange protocol compared to a traditional approach of a static set of temperatures was found to reduce the amount of population disorder and shifted the ensemble to helical conformations with an extended peptide folding stabilization. A comparison with an explicit/implicit solvent hybrid MD-based replica exchange simulation showed that conformational sampling on an explicit solvent landscape leads to a free-energy minimum of approximately 20 % helicity, yet the overall conformational network underlying transient states resembles more of a helix-fold propensity in a solvent mixture of TFE-water rather than bulk water. The simulation results can be summarized as a benchmark for the testing of more refined CHARMM-based force fields and different GB model parameterizations. The ultimate goal is to capture greater heterogeneity in conformational probabilities and reduce the overstabilization of helix propensities in modeling

intrinsically disordered peptides.

Funding

Financial support for this work comes from US Department of Defense Threat Reduction Agency grant (DTRA 4.10011_07_RD_B). The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting the views of the US Army or of the US Department of Defense. This article has been approved for public release with unlimited distribution.

Conflict of Interest Statement

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The author thanks Michael Lee for earlier collaborations on the adaptive temperature-based replica exchange method and assistance from Evan Olson on the graphics of 2D profiles.

References

Arai, M., Sugase, K., Dyson, H. J., and Wright P. E. (2015). Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proc. Natl. Acad. Sci. U.S.A.* 112, 9614-9.

Bhowmick, A., Brookes, D. H., Yost, S. R., Dyson, H. J., Forman-Kay, J. D., Gunter, D., Head-Gordon, M., Hura, G. L., Pande, V. S., Wemmer, D. E., Wright, P. E., and Head-Gordon, T. (2016). Finding Our Way in the Dark Proteome. *J. Am. Chem. Soc*. 138, 9730-42.

Brooks, B. R., Brooks, C. L. 3rd., Mackerell, A. D. Jr., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G. Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post ,C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu ,X., Yang, W., York, D. M., and Karplus, M. (2009). CHARMM: The biomolecular simulation program. *J. Comput. Chem*. 30, 1545–1614.

Chaudhury, S., Olson, M. A., Tawa, G., Wallqvist, A., and Lee, M. S. (2012). Efficient conformational sampling in explicit solvent using a hybrid replica exchange molecular dynamics method. *J. Chem. Theory Comput*. 8, 677-687.

Chebaro, Y., Ballard, A. J., Chakraborty, D., and Wales, D. J. (2015). Intrinsically disordered energy landscapes. *Sci. Rep*. 5, 10386.

Chen, J. (2010). Effective Approximation of Molecular Volume Using Atom-Centered Dielectric Functions in Generalized Born Models. *J. Chem. Theory Comput*. 6, 2790-803.

Click, T. H., Ganguly, D., and Chen, J. (2010). Intrinsically disordered proteins in a physics-based world. *Int. J. Mol. Sci*. 11, 5292-309.

Feig, M., Karanicolas, J., and Brooks, C. L 3rd. (2004). MMTSB Tool Set: Enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model*. 22, 377-395.

Ganguly, D., and Chen. J. (2009). Atomistic details of the disordered states of KID and pKID. Implications in coupled binding and folding. *J. Am. Chem. Soc*. 131, 5214-23.

Ganguly, D., and Chen. J. (2015). Modulation of the disordered conformational ensembles of the p53 transactivation domain by cancer-associated mutations. *PLoS Comput. Biol*. 11, e1004247.

Higo, J., Nishimura, Y., and Nakamura, H. (2011). A free-energy landscape for coupled folding and binding of an intrinsically disordered protein in explicit solvent from detailed all-atom computations. *J. Am. Chem. Soc*. 133, 10448-58.

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., and MacKerell, A. D. Jr. (2016). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods*. doi: 10.1038/nmeth.4067. [Epub ahead of print].

Im, W., Lee, M. S., and Brooks, C. L. III. (2003). Generalized born model with a simple smoothing function. *J. Comput. Chem*. 24, 1691-702.

Ishikawa, Y., Sugita, Y., Nishikawa, T., and Okamoto, Y. (2001). Ab initio replica-exchange Monte Carlo method for cluster studies. *Chem. Phys. Lett*. 33, 199–206.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*.79, 926-935.

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

Kohn, J.E., Millett, I. S., Jacob, J., Zagrovic, B., Dillon, T. M., Cingel, N., Dothager, R. S., Seifert, S., Thiyagarajan, P., Sosnick, T. R., Hasan, M. Z., Pande, V. S., Ruczinski, I., Doniach, S., and Plaxco, K. W. (2004). Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A*. 101, 12491-6.

Lee, K. H., and Chen, J. (2016). Multiscale enhanced sampling of intrinsically disordered protein conformations. *J. Comput. Chem*. 37, 550-7.

Lee, M. S., Feig, M., Salsbury, F. R. Jr., and Brooks, C.L. 3rd. (2003). New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem*. 24, 1348-1356.

Lee, M. S., and Olson, M. A. (2010). Protein folding simulations combining self-guided Langevin dynamics and temperature-based replica exchange. *J. Chem. Theory Comput*. 6, 2477-2487.

Lee, M. S., and Olson, M. A. (2011). Comparison of two adaptive temperature-based replica exchange methods applied to a sharp phase transition of protein unfolding-folding. *J. Chem. Phys.* 134, 244111-24417.

Leung, D. W., Borek, D., Luthra, P., Binning, J. M., Anantpadma, M., Liu, G., Harvey, I. B., Su, Z., Endlich-Frazier, A., Pan, J., Shabman, R. S., Chiu, W., Davey, R. A., Otwinowski, Z., Basler, C.F., and Amarasinghe G.K. (2015). An intrinsically disordered peptide from Ebola virus VP35 controls viral RNA synthesis by modulating nucleoprotein-RNA interactions. *Cell Rep*. 11, 376-89.

Mackerell, A.D. Jr., Feig, M., and Brooks, C.L. 3rd. (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* 25, 1400–1415.

Metropolis, N,, Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys*. 21, 1087–1092.

Miao, Y., Feixas, F., Eun, C., and McCammon, J. A. (2015). Accelerated molecular dynamics simulations of protein folding. *J. Comput. Chem*. 36, 1536-49.

Mittal, A., Lyle, N., Harmon, T. S., and Pappu, R. V. (2014). Hamiltonian switch Metropolis Monte Carlo simulations for improved conformational sampling of intrinsically disordered regions tethered to ordered domains of proteins. *J. Chem. Theory Comput*. 10, 3550-3562.

Olson, M. A., and Lee, M. S. (2014). Evaluation of unrestrained replica-exchange simulations using dynamic walkers in temperature space for protein structure refinement. *PLoS One*. 9, e96638.

Olson, M. A., Legler, P. M., and Goldman, E. R. (2016). Comparison of replica exchange simulations of a kinetically trapped protein conformational state and its native form. *J. Phys. Chem. B.* 120, 2234-40.

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., Schafferhans, A., and O'Donoghue, S. I. (2015). Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U.S.A*. 112, 15898-903.

Peter, E. K., Shea, J. E., and Pivkin, I. V. (2016). Coarse kMC-based replica exchange algorithms for the accelerated simulation of protein folding in explicit solvent. *Phys. Chem. Chem. Phys.* 18, 13052-65.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kalé, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem*. 26, 1781-1802.

Piana, S., Klepeis, J. L., and Shaw, D. E. (2014). Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol*. 24, 98-105.

Predescu, C., Predescu, M., and Ciobanu, C. V. (2004). The incomplete beta function law for parallel tempering sampling of classical canonical systems. *Chem. Phys*. 120, 4119-28.

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

Sanchez, A., Geisbert, T.W., and Feldmann, H. (2006). Filoviridae: Marburg and Ebola viruses. In Fields Virology, D.M. Knipe, P.M. Howley, R.A. Griffin, M.A. Martin, B. Roizman, and S.E. Straus, eds. (Lippincott Williams & Wilkins), pp. 1409–1448.

Shoemaker, B. A., Portman, J. J., and Wolynes, P. G. (2000). Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. U.S.A.* 97, 8868-73.

Sugitaa, Y. and Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett*. 314: 141–151.

Trebst, S., Troyer, M., and Hansmann, U. H. (2006). Optimized parallel tempering simulations of proteins. *J. Chem. Phys*. 124: 174903-09.

Wright, P. E., and Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol*. 293, 321-31.

Wright, P. E., and Dyson, H. J. (2005). Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol*. 6, 197-208.

Wu, X., and Brooks, B.R. (2003). Self-guided Langevin dynamics simulation method. *Chem. Phys. Letters*. 381, 512-518.

Wu, X., and Brooks, B. R. (2011). Toward canonical ensemble distribution from self-guided Langevin dynamics simulation. *J. Chem. Phys.* 134, 134108-134119.

Wu, X., Brooks, B. R., and Vanden-Eijnden, E. (2016). Self-guided Langevin dynamics via generalized Langevin equation. *J. Comput. Chem*. 37, 595-601.

Yeh, I.C., Lee, M. S., and Olson, M.A (2008). Calculation of protein heat capacity from replica-exchange molecular dynamics simulations with different implicit solvent models. *J. Phys Chem. B*. 112, 15064-73.

Kieslich, C. A., Smadbeck, J., Khoury, G. A., and Floudas, C. A. (2016). conSSert: Consensus SVM model for accurate prediction of ordered secondary structure. *J. Chem. Inf. Model*. 56, 455-61.

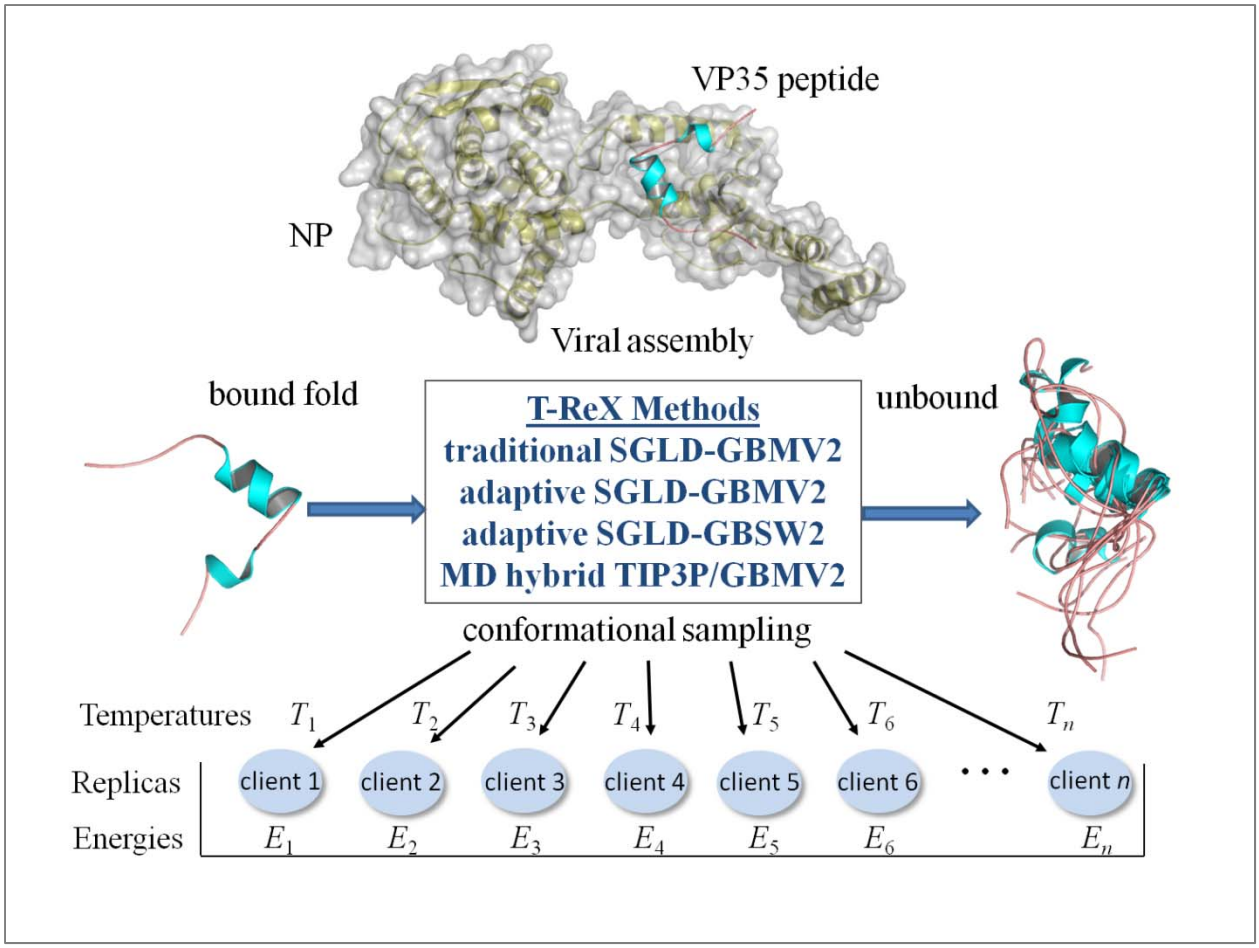

Figure 1 | Computational strategies of modeling the Ebola virus VP35 peptide (PDB: 4YPI) in its unbound form using temperature-based replica exchange (T-ReX) simulation methods. The methods include: (1) GBMV2 solvent model applied with a traditional (static) set of temperatures spanning a range from a minimum temperature (T_1) to the upper extreme (T_n), where n is the number thermal windows (ensemble computing clients); (2) GBMV2 using an adaptive (dynamically walking) set of temperatures between T_1 and T_n; (3) GBSW2 solvent model applied by with adaptive sampling; and (4) TIP3P/GBMV2 hybrid replica exchange method. Energies (E_i) used in the replica exchanges by the different strategies are described in the text. Molecular figures were drawn with PyMOL (www.pymol.org).

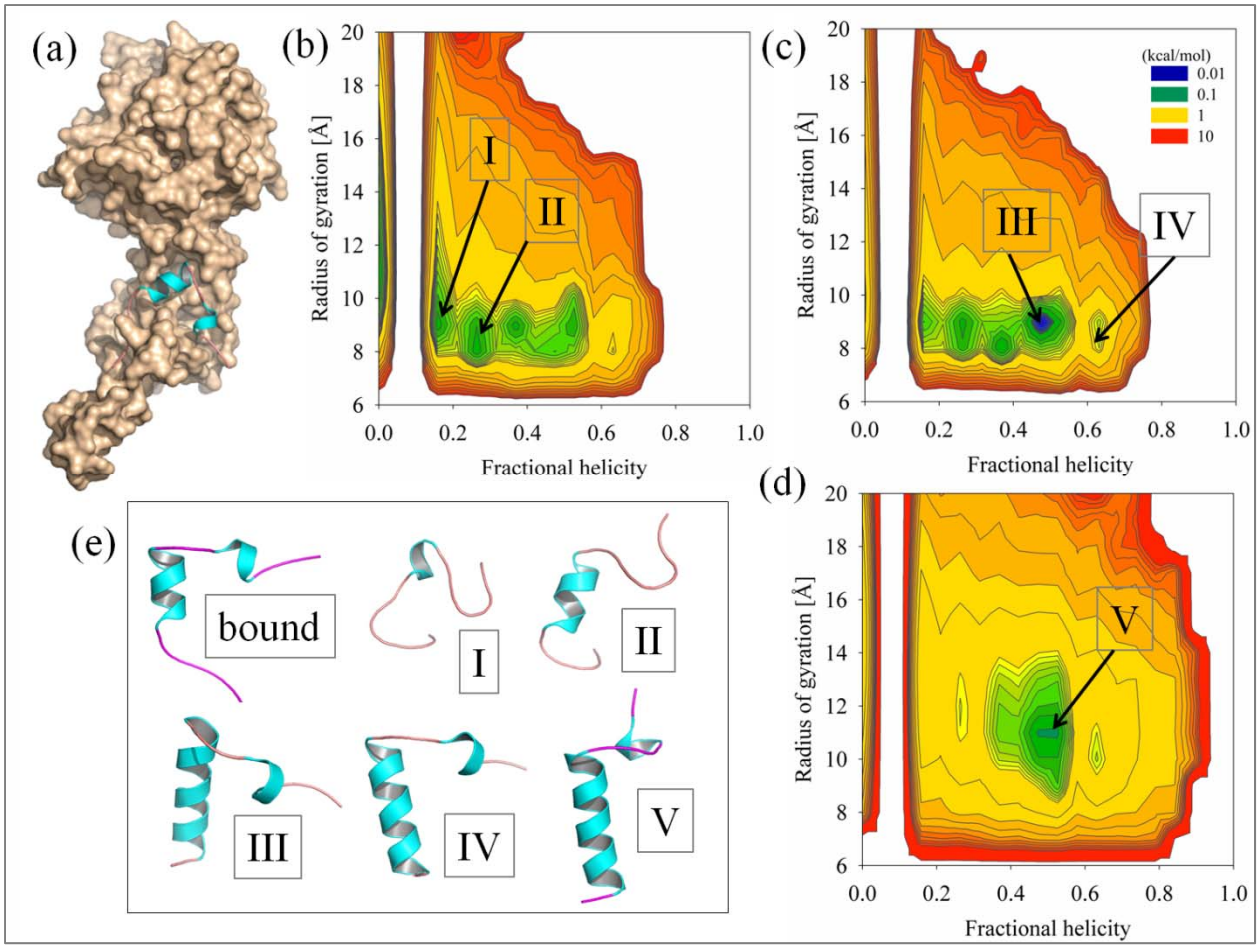

Figure 2 | Simulation results of sampling the Ebola virus VP35 NPBP peptide using GB-based solvent models and replica exchange methods. (a) X-ray crystallographic structure of the NPBP peptide bound to Ebola NP (depicted as a molecular surface). (b) Probability density profile $\Omega(f_H,R_g)$ computed from the static T-ReX simulation method with order parameters of fractional helicity and radius of gyration. (c) Probability density profile results from the adaptive T-ReX method with the GBMV2 solvent model. (d) Adaptive T-ReX with GBSW2 solvent model. (e) Representative conformations extracted from the simulations at indicated basins.

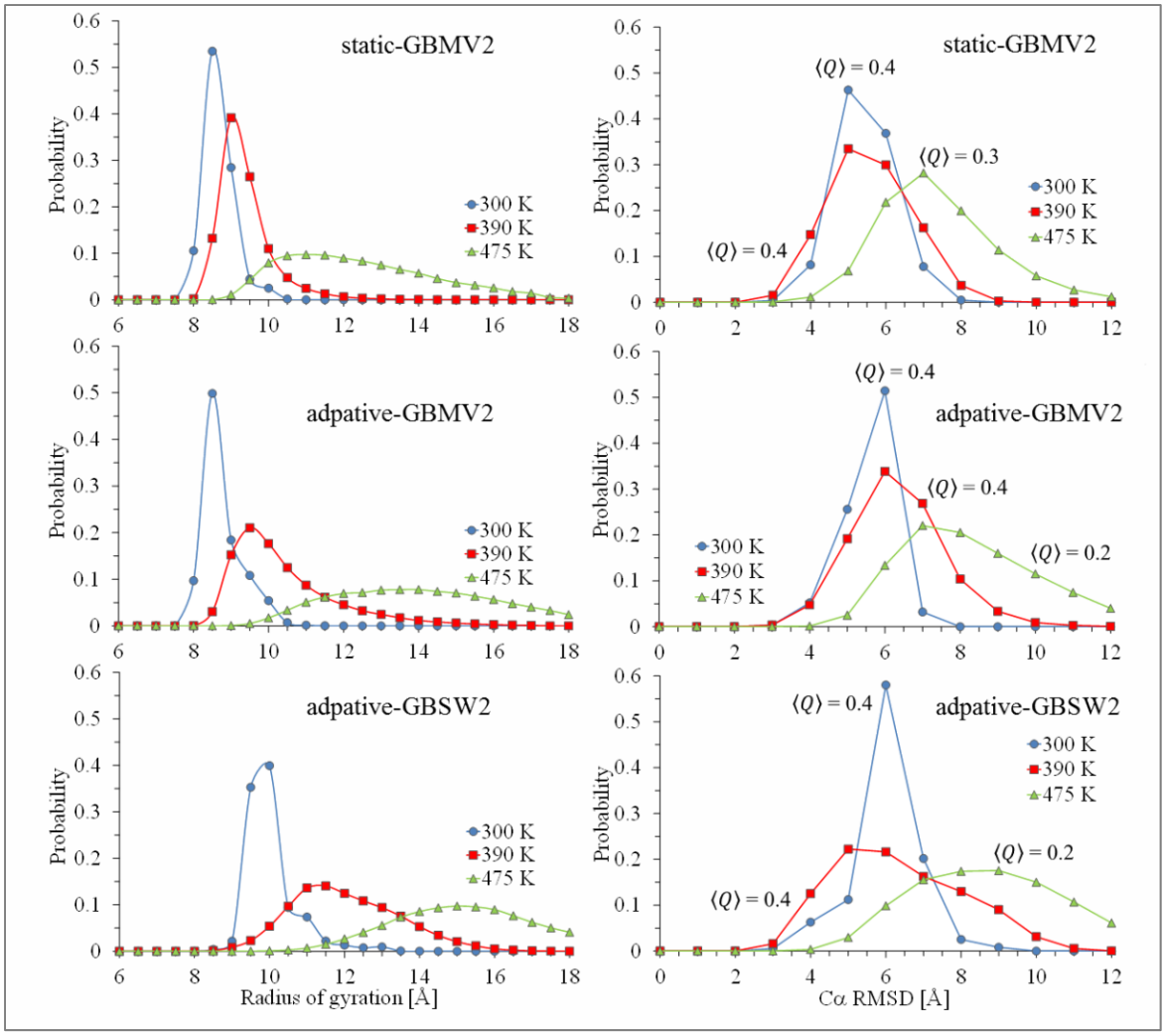

Figure 3 | Calculated probability profiles for sampling values of radius of gyration and Cα-RMSD from the starting bound conformation of the NPBP peptide. Plot lines colored blue represent quantities extracted at 300 K from the generated conformational ensembles, red represent values at 390 K and green at 475 K. From the top figure to bottom, simulation results are static T-ReX/GBMV2, adaptive T-ReX/GBMV2 and adaptive T-ReX/GBSW2.

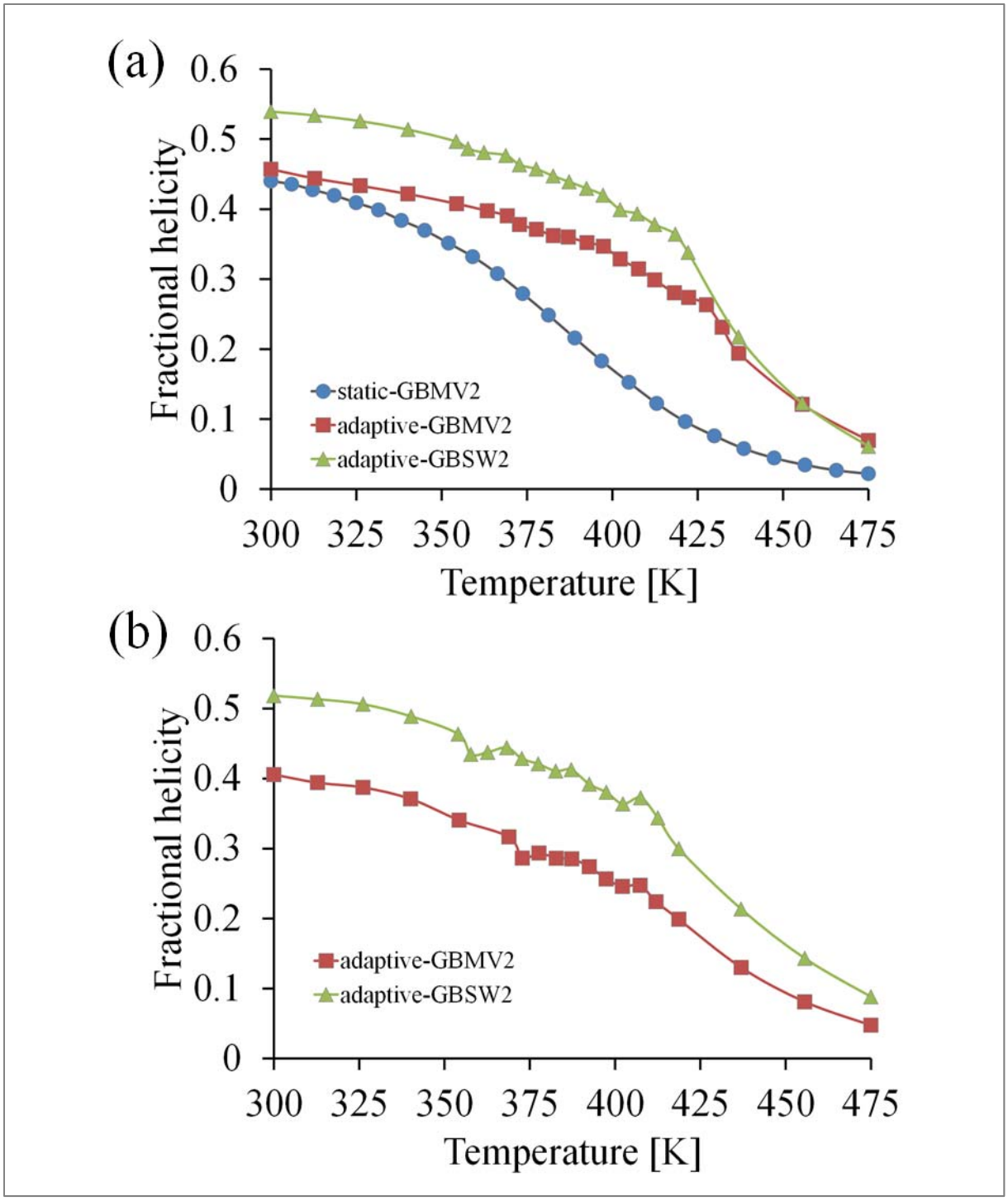

Figure 4 | Thermal unfolding profiles computed from the simulations of the Ebola NPBP peptide. (a) Profiles calculated from the starting folded peptide conformation using the three simulation models of the static T-ReX/GBMV2 (blue colored line), adaptive T-ReX/GBMV2 (red colored line) and adaptive T-ReX/GBSW2 (green colored line). (b) Profiles calculated from the adaptive T-ReX simulations of starting from an unstructured (coil) peptide fold.

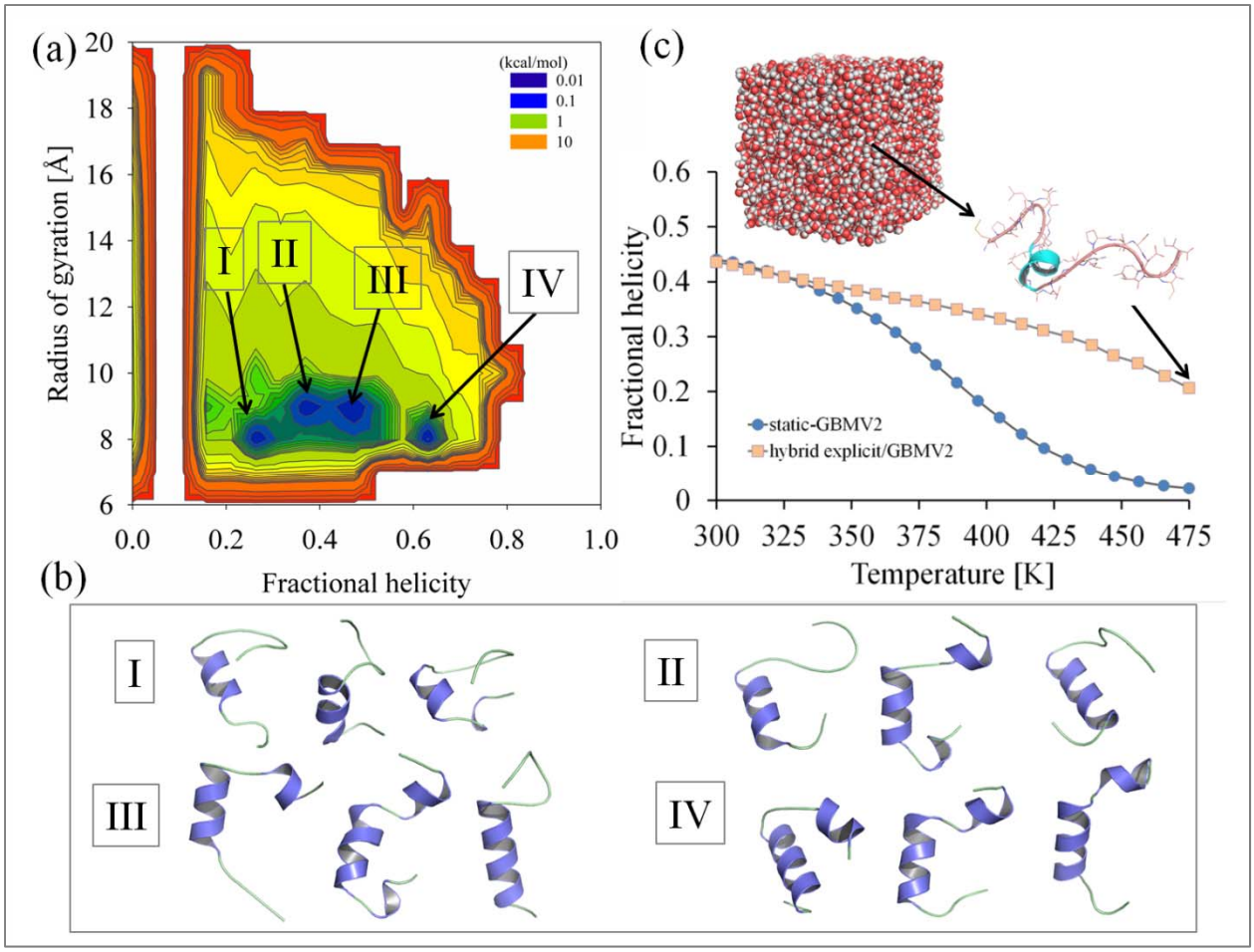

Figure 5 | Simulation results of sampling the Ebola virus VP35 NPBP peptide using the explicit/implicit solvent hybrid T-ReX/MD method. (a) Probability density profile $\Omega(f_H,R_g)$ computed from sampling fractional helicity and radius of gyration. (b) Representative conformations extracted from the simulations are illustrated for selected basins. (c) Thermal unfolding profiles of the peptide computed using the explicit/implicit solvent hybrid T-ReX/MD method (light colored symbols) compared to the static T-ReX/SGLD method using GBMV2 (blue colored symbols). A representative structure is shown from the explicit solvent calculation.