JOURNAL OF CHEMICAL INFORMATION AND MODELING

Letter

pubs.acs.org/jcim

Macrostate Identification from Biomolecular Simulations through Time Series Analysis

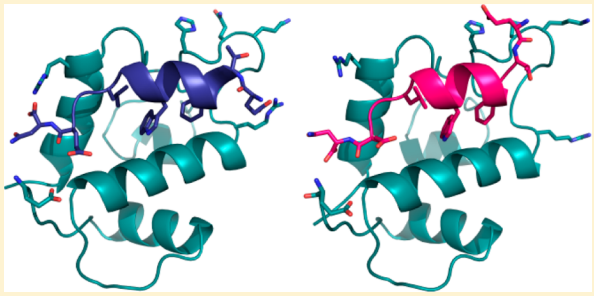
Weizhuang Zhou,[†,∥] Efthimios Motakis,[†,∥] Gloria Fuentes,*[,†] and Chandra S. Verma*[,†,‡,§]

[†]Bioinformatics Institute (A*STAR), 30 Biopolis Street, no. 07-01 Matrix, Singapore 138671

[‡]Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543

[§]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

Supporting Information



ABSTRACT: This paper builds upon the need for a more descriptive and accurate understanding of the landscape of intermolecular interactions, particularly those involving macromolecules such as proteins. For this, we need methods that move away from the single conformation description of binding events, toward a descriptive free energy landscape where different macrostates can coexist. Molecular dynamics simulations and molecular mechanics Poisson−Boltzmann surface area (MM-PBSA) methods provide an excellent approach for such a dynamic description of the binding events. An alternative to the standard method of the statistical reporting of such results is proposed.

INTRODUCTION

Free energy calculations lie at the core of using computational chemistry to understand intermolecular interactions. They serve two purposes: to provide physical insight into the molecular details that are not easily accessible by experimental techniques and to develop testable hypotheses. The development of a method to estimate free energies rapidly and accurately is still an elusive goal but remains a critical need in drug design, where it is used in virtual high throughput screening and computational lead optimization protocols. There are several methods to estimate free energies, each involving some compromise between computational efficiency/cost and accuracy.[1,2] The MM-(GB)PBSA method has gained popularity due to its fast performance and its ability to predict trends in binding affinities fairly reliably.[3−5] This approach was codeveloped more than a decade ago by David Case and Peter Kollman and takes its name from the respective components of the absolute free energy, notably the gas-phase energy or molecular mechanics internal energy (MM), the electrostatic solvation energy as determined by Generalized Born (GB) or Poisson−Boltzmann (PB) models, and the nonpolar solvation energy estimated from the exposed surface area (SA). Traditionally, one performs a molecular dynamics simulation with explicit solvent molecules and selects a set of representative snapshots from a stretch of the trajectory that is deemed to be at equilibrium (or converged). These structures are stripped of any solvent and ion molecules, and the free energy is calculated according to the following equation:

$$\Delta G_{\text{binding}} = G_{\text{complex}} - (G_{\text{receptor}} + G_{\text{ligand}}) \quad (1)$$

where the free energy of each term is calculated as a combination of MM and continuum solvent (implicit model) approaches as outlined above.

The average binding energy across the set of selected snapshots is reported as the net free energy. Large variations in the calculated free energy between the snapshots are common however, and this has been noted by many authors including the developers themselves. To estimate the statistical uncertainty in the reported average, the standard error is often used. The standard error of the mean (SEM) can be estimated[6,7] as

$$\text{SEM} = \frac{\sigma}{\sqrt{N}} = \sigma\sqrt{\frac{t_{\text{corr}}}{t_{\text{sim}}}} \quad (2)$$

where σ denotes the standard deviation of the set of snapshots,[8] N is the number of snapshots in the set which are assumed to be independent and identically distributed (i.i.d.), t_{corr} is the correlation time calculated from the autocorrelation function (acf), and t_{sim} is the total length of the simulation.

Grossfield and Zuckerman suggest that the correlation-time method suffers from many shortcomings and propose the estimation of SEM by block averaging,[9,10] where the trajectory is divided into blocks of equal size and the standard error between the means of the blocks is computed. The block sizes are progressively increased until the SEM approaches an asymptote, which is considered to approximate the "true" SEM (see Figure 1). This method was subsequently applied directly to the $\Delta G_{\text{binding}}$ time series by McGillick et al.[11] to obtain a more accurate estimate of the standard error. A similar approach to obtaining the standard error of the mean is the statistical inefficiency method:[7]

Published: August 28, 2012

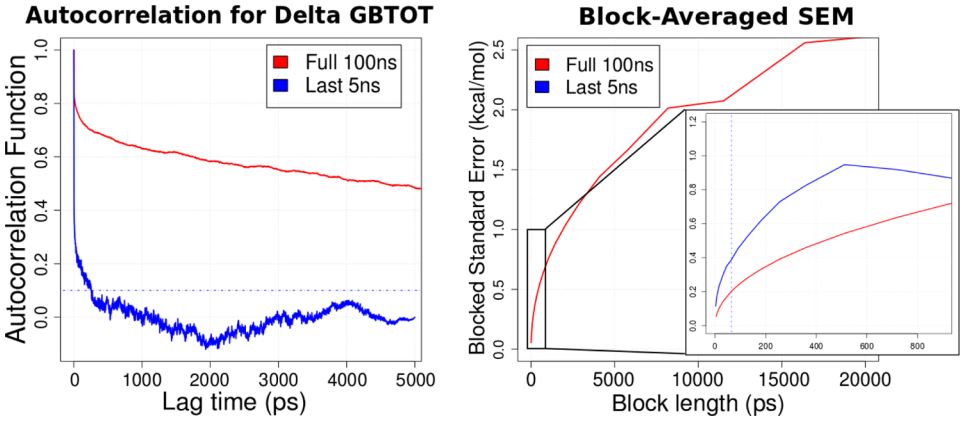


Figure 1. Effect of simulation length on correlation time. Correlation time can be approximated from the time taken for the autocorrelation function to drop below a predefined threshold (left) or from the block length at which the blocked SEM reaches a plateau (right). The data set for the blue curves is the last 5 ns of the 100 ns time series. The blue curve in the autocorrelation plot shows a much shorter correlation time than the red. The magnified view of the block-average SEM plot (inset in right figure) shows that whereas the blue curve reaches a plateau within a block length of ~800 ps, the red curve continues to rise steadily even after 20 ns.

$$s = \lim_{n_b \to \infty} \frac{n_b \sigma^2(\langle X \rangle_b)}{\sigma^2(X)} \tag{3}$$

where n_b is the number of observations within a block b, and s is the step-size that generates uncorrelated snapshots. As the true variance of the mean is approached, it would become approximately constant at some multiple of the sample variance. As such, graphs of s against n_b are plotted and the value of s is estimated by the plateau height[12] (refer to Figure S1 Supporting Information).

Both the block-averaging and statistical inefficiency methods produce correlation times that are affected by the length of simulations. In their simulations of the biotin analogue Btn6-avidin complex, Genheden and Ryde[13] noted that using the entire time series (1.6 ns) of binding energy values resulted in a correlation time of 41 ps. When the time series was broken into nonoverlapping segments of 200 ps each, the correlation time in each segment decreased drastically, with a median of 3 ps. Considering that most simulations today are of the order of tens to hundreds of nanoseconds, we performed the same statistical-inefficiency analysis on a 100 ns simulation of p53/MDM2 (refer to Supporting Information Figure S1). The correlation time was calculated to be an alarming 10 ns. If the block-average SEM method was used instead, the calculated correlation time was approximately 12.5 ns, showing good agreement between the two methods. When the analysis was done using only the last 5 ns segment of the trajectory, the correlation time decreased to ~500 ps (Figure 1) using the statistical inefficiency method. The correlation time decreases further to less than 100 ps if the sampling was instead done over a 1 ns interval, which is the typical interval used in the usual analysis of binding energy. Genheden and Ryde suggest that long-term drifts in energy are responsible for the increasing correlation times when longer simulations are performed. We suspect that such long-term drifts in energy are likely to be common for most large protein−ligand systems (around 100 residues or more), but to our knowledge, this has not been addressed in the literature concerning MM-GBSA. This is probably because papers from a decade ago commonly dealt with much smaller systems simulated over much shorter times, typically less than 5 ns. However, the increased availability of improved hardware and software are now making longer simulations the norm.

Hence, it is possible that current methods of obtaining the statistical uncertainty may not be useful in evaluating long simulations that are prevalent today. A recent trend that appears to be heading in the right direction is to avoid reporting point estimators such as the mean and standard errors for MM-(GB)PBSA results, but instead show the full distribution of the results through graphical means such as histograms.[14] While this is very useful in showing the full variation in the data, it does not provide any information on the temporal relationship and presence of macrostates. In particular, the spread of data can be of the order of 50 kcal/mol or more, making it difficult to arrive at any useful conclusions. We believe that this can be further improved upon by identifying macrostates from the characterization of subpopulations estimated from time series data analysis.

Here, we propose a new method, called *MMPBSA_segmentation*, to represent the MM-(GB)PBSA results. Unlike previous approaches which assume that the whole series (from a trajectory) originates from a single population, *MMPBSA_segmentation* tests and estimates the significance of multiple, statistically distinct subpopulations (within a trajectory) in three steps: (1) data visualization and adjustment, (2) data segmentation (change-point analysis) that gives an estimate of the segments, and (3) nonparametric estimation (curve fitting) to evaluate the differences among the segments. This segmentation is optimized in C language, incurring a relatively low computational cost. A fourth step is then taken to combine the results of 1−3 and identify the final subpopulations of interest. This is done by one-dimensional pattern recognition via hierarchical clustering.

We demonstrate our method's use and practicality on the well-characterized p53/MDM2 system for which several groups, including ours, have carried out extensive simulations.[15] *MMPBSA_segmentation* is written and wrapped in a user-friendly and flexible R package that is available upon request. Our results show that *MMPBSA_segmentation* is able to identify distinct biologically meaningful subpopulations that appear to have the characteristics of ergodicity. A concise description of the algorithm's main steps is given in the Methods section.

dx.doi.org/10.1021/ci300341v | J. Chem. Inf. Model. 2012, 52, 2319−2324

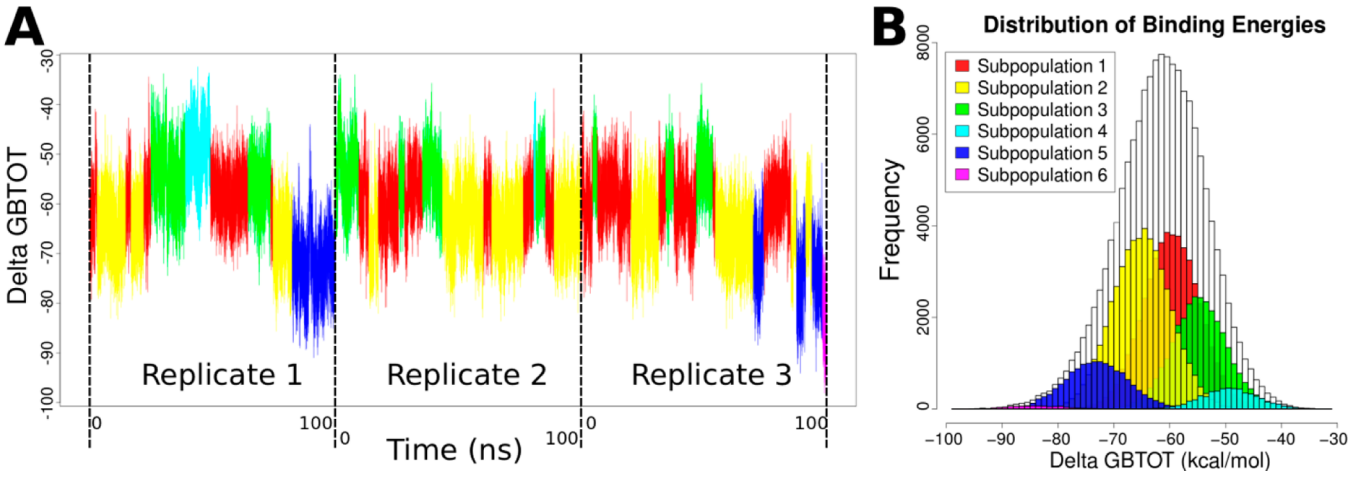


Figure 2. (A) MM-GBSA δ GBTOT time series for 3 replicates, each 100 ns in length. Colored sections represent the segments that were grouped together based on our described method. (B) Histogram of binding energies in each subpopulation. The white envelope is the histogram for binding energies from the entire trajectory.

Extensive information on the underlying methods and the R functions is given in the Supporting Information (SI).

METHODS

Simulation. The crystal structure 1YCR has been used as the initial structure in a minimization and equilibration simulation multistep protocol, as has been previously described.[16] The production run comprises three replicates of 100 ns trajectories, each started by providing a different seed to the random number generator so as to obtain different initial velocities from the Maxwell distribution. All the calculations have been carried out using the GPU code of the AMBER11 software.[17] We then applied the single trajectory MM-GBSA protocol for the whole trajectory to obtain a GB time series (Figure 2A), using the MMPBSA.py module in AMBER[5] (see the SI for parameters). The snapshots were evaluated at a spacing of 2 ps, giving rise to 50 000 snapshots per 100 ns trajectory.

MMPBSA_segmentation Algorithm. Step 1. Data Visualization and Adjustment. We denote by X_t, $t = 1, ..., T$ the (noisy) data from a single trajectory (time series). The segmentation algorithm can be applied only to trend-stationary series, i.e. series not exhibiting a stable long-run trend. This assumption is tested by the augmented Dickey−Fuller statistic.[18] If stationarity is not satisfied, we propose either detrending X_t by calculating first differences as $X_t - X_{t-1}$ or manually splitting the series into subseries $\{X_{t_{(s)}}; t_{(s)} = s - n_s, s - n_s + 1, ..., s; s = 1, n_s + 1, ...\}$ not necessarily equal in length, and analyze each $X_{t_{(s)}}$ independently. This is done by a first round of data segmentation, specifying the visually identified number of change points (described below and in the SI; see Figure S4). The choice of the best strategy depends on visualizing time series plots of the trajectories, as discussed in the SI.

Step 2. Statistical Change-Point Analysis. We estimate the locations of discontinuities in X_t by piecewise regression models and dynamic programming (PRDP).[19] In the first step, the algorithm identifies the location of S nonoverlapping segments, chosen to minimize an unbiased estimate of sum of square error (SSE) as a measure of goodness-of-fit for those segments $\nu(i, j) = 1/(j - i + 1 - l) \times \sum_{t=i}^{j}(X_t - \hat{X}_t(\theta))^2$ where θ denotes an estimated parameters vector of size l. Subsequently, it performs fast simultaneous identification of several "change-points" $t_{(c)}$, $c = 1, ..., C$, via dynamic programming and successive minimization of the SSE.[19] Each of the $C + 1$ segments $\{X_{t_{(c)}}\}$ need not be of equal length. The significance of each potential change-point is assessed by the likelihood ratio test.

Step 3. Curve Fitting via Wavelet Denoising. We assume that X_t is partitioned into $\{X_{t_1}, ..., X_{t_C}\}$, each of length l_c with the constraint $l_c > 200$ (a user-specified, adjustable threshold; choice of 200 is to ensure that segments partitioned from our trajectories are at least 400 ps in length). The data within each c are modeled as $X_{t_c} = F(\mu_{t_c}) + \varepsilon_{t_c}$, where $F(.)$ is an unknown function (nonparametric form) of the conditional expectation of X_{t_c} and $\varepsilon_{t_c} \sim N(0,\sigma_\varepsilon)$ are the independent and identically distributed residuals following zero-mean Normal distribution (a necessary assumption to separate signal from noise[20]). Typically, X_{t_c} is transformed into the Haar wavelet domain[21] and subsequently denoised by universal thresholding. The inverse wavelet transform produces the denoised signal $F(\mu_{t_c})$, which typically resembles a step function. We consider $F(\mu_{t_c})$ as the "identity" of each segment c and, at the next step, it will be used for one-dimensional pattern recognition. Typically in wavelets denoising, the estimated ε_{t_c} and $\hat{\varepsilon}_{t_c}$ are approximately $N(0,\sigma_\varepsilon)$ distributed and $\mathrm{acf}(\hat{\varepsilon}_{t_c}) = 0$ (formally tested at $\alpha = 1\%$).

Step 4. One-Dimensional Pattern Recognition by Hierarchical Clustering. We merge the segmented and denoised data of the multiple trajectories and estimate the distinct subpopulations by hierarchical clustering. Note that the time information is not considered here because the multiple independent trajectories are joined and analyzed in one vector. Thus, we wish to cluster the segments based on the time-free characteristics of the denoised signal $F(\mu_{t_c})$. For computational reasons, we describe each $F(\mu_{t_c})$ by the 21 quantiles of its distribution

$$\vec{q}_c = \{\min(F(\mu_{t_c})), q_{5\%}^{F(\mu_{tc})}, q_{10\%}^{F(\mu_{tc})}, q_{15\%}^{F(\mu_{tc})}, ..., \widetilde{F(\mu_{t_c})}, ..., \max(F(\mu_{t_c}))\}$$

dx.doi.org/10.1021/ci300341v | J. Chem. Inf. Model. 2012, 52, 2319−2324

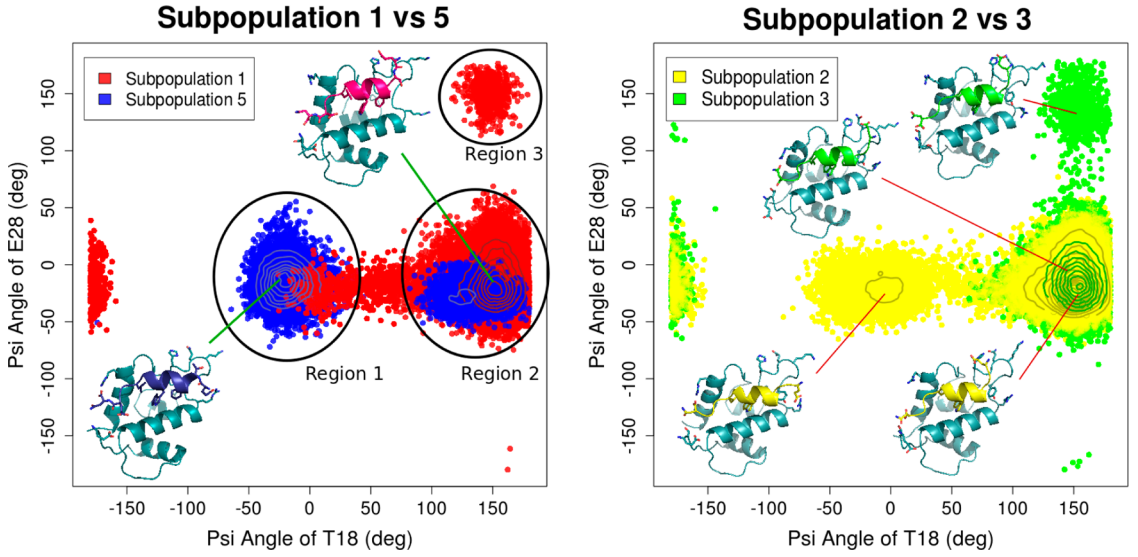


Figure 3. Structural comparisons between subpopulations 1 and 5 and subpopulations 2 and 3. These subpopulations have been characterized based on the psi angles of T18 and E28 of p53 peptide. The different rotameric states introduce markedly distinct structural rearrangements in the two terminals of the peptide, affecting its interactions with MDM2. Although subpopulations may be found in both regions 1 and 2, the relative density (seen from the contour lines) in the regions affects the overall binding affinity. Individual plots can be found in the SI.

where $\widetilde{F(\mu_{t_c})}$ denotes the median denoised signal and $q_{i\%}^{F(\mu_{tc})}$, the ith quantile of $F(\mu_{t_c})$. Next, the mean Euclidean distances across all elements of the 21-dimensional vectors $\vec{q}_c$, $\vec{q}_{c'}$ are subsequently calculated as $D_{c,c'} = \sum_{i=1}^{20}\sum_{j=(i+1)}^{20}(\{q_{i\%}^{F(\mu_{tc})} - q_{i\%}^{F(\mu_{tc})}\}^2)^{1/2}/(0.5 \times 21 \times 20)$ where $c \neq c'$. Typically, the input distance matrix has dimension $(C + 1) \times (C + 1)$ and is used for hierarchical clustering. Hierarchical clustering leads to numerical and visual evidence of similar segments via tree plots. The cutoff in the tree can be automatically estimated as the median(diff(l$D_{c,c'}$l)), i.e. the median of the first differences of the absolute $D_{c,c'}$ values. Alternatively, a user-defined cutoff can be used to cluster similar segments. These clusters constitute the different subpopulations of our study.

Here, we argue that simple hierarchical clustering is preferable to a formal statistical algorithm (based on hypothesis testing between nonadjacent segments), since the latter might separate two segments c and $c + i$, where $i > 1$, even when the segments are biologically similar. We allow the researcher to judge and decide the number of subpopulations based on previous assumptions and data visualization. The visualization is aided by user-friendly *R* interactive plots.

The final result of this method graphically reports a number of biologically relevant subpopulations and the ergodicity of the system. According to the energy landscape for binding,[22,23] the energetically different subpopulations correspond to metastable conformations of the complexes and can be treated as macrostates. The binding landscape for biological systems shows a degree of ruggedness that is similar to folding landscapes, and this corresponds to the different binding modes that MD simulations explore. Indeed, it has been hypothesized that this ruggedness also modulates the kinetics of binding.[24]

RESULTS

The method developed here has been applied to the p53-MDM2 system. The identification of the different subpopulations is shown in Figure 2A. Histograms of binding energies for each of the identified subpopulations are then plotted on the same graph, with the overall histogram in the background (Figure 2B). This allows one to immediately see the mean binding energies of those macrostates and identify those that are more prevalent in the overall composition of the trajectory (and hence contribute more to the overall mean). The proposed method also allows one to check for reproducibility and ergodicity in replicates by comparing the presence of macrostates. We note that there is potential to perform a wider range of characterizations of the macrostates that may reveal more information regarding the binding modes. In the present study however, we have limited the analysis to discern structural features that characterized the subpopulations.

Identified Macrostates are Structurally Different. As a proof of concept, we investigate the relevance of the current suggestions to the conformational flexibility of the N- and C-termini of the p53 peptide complexed to MDM2.[25,15,16] We look at the variations in the backbone psi angle of residues T18 and E28 (found near the N- and C- termini respectively) in the p53 peptide across the different subpopulations and perform a pairwise comparison of two pairs of dissimilar populations.

Three main structural regions are observed in the 2D density plots (Figure 3), and the relative density in these regions is found to be correlated with the binding affinity of a subpopulation. Subpopulation 1 (red) is characterized by the stabilization of E17 in a cationic field formed by residues K70, Q71, and H73, a conformation similar to the crystal structure (see SI Figure S3 for residue labels). At the C-terminus of the p53 peptide in the same subpopulation, N29 interacts mainly with the backbone and side chains of E25 and T26 of MDM2. The salt bridge interaction between the carboxylate terminus of p53 and R97 of MDM2 (corresponding to region 3 in Figure 3) is found to be relatively infrequent in subpopulation 1 (~1% of the snapshots). As such, up to 97.3% of the snapshots in subpopulation 1 are found to reside in region 2, defined by $\psi(T18) \in [100°, 180°]$ and $\psi(E28) \in [-50°, 50°]$. In contrast, only 14.4% of subpopulation 5 (blue), which corresponds to the second highest binding affinity of all six identified macrostates, is found in region 2. Instead, 85.6% of subpopulation 5 is found in region 1, defined by $\psi(T18) \in$

dx.doi.org/10.1021/ci300341v | J. Chem. Inf. Model. 2012, 52, 2319−2324

[−60°, 40°] and $\psi(E28) \in [-50°, 50°]$. The N-terminus of the peptide in subpopulation 5 adopts an extra turn of a helix with E17 stabilized by R65 and Y67. This appears to allow for tighter interactions between the peptide and MDM2 than would be gauged from the original crystal structure. Since the extra turn of the helix is effected by the psi angle of T18 at the N-terminus, the regions 1 and 2, which demarcate the two allowable ranges of psi angles for T18 in the peptide, have a direct impact on the binding energy. In particular, region 1 is more favorable than region 2 for binding (see Table S1 in the SI).

Not surprisingly, region 1 is not typically occupied by the other subpopulations with lower binding affinity. Instead, those subpopulations are found to mainly occupy region 2. For subpopulations 2 and 3, in which there was some overlap in the energy distribution (see Figure 2B), we also found that there was overlap in the structural distribution when the subpopulations were mapped along the same two structural parameters examined earlier (Figure 3). In subpopulation 2, E17 toggles between the K70/Q71/H73 and the R65/Y67 surfaces while N29 shows an interaction pattern similar to the one seen in subpopulation 5. In subpopulation 3, only the former interaction of E17 is seen while the C-terminus of p53 peptide is characterized by interactions similar to those in subpopulation 1, i.e. N29 engaging residues E25/T26 and R97 (hence the occupancy of region 3 in Figure 3). Although the subpopulations may occupy more than one region, it is their relative occupancy of regions 1 and 2 that affects their overall binding affinity. For instance, 12.9% of subpopulation 2 is found to reside in region 1 as compared to only 1.8% of subpopulation 1 and less than 1.0% of subpopulation 3. This is enough to make subpopulation 2 a better binder than subpopulations 1 and 3, with the order of binding affinity correlating with the degree of occupancy of region 1 (binding affinity of subpopulations: 2 > 1 > 3).

Apart from structural differences in the conformations that the p53 peptide adopts, we also find indications of different dynamical behaviors of MDM2 within the subpopulations identified. Preliminary results suggest that in the higher affinity binding subpopulation, there is an increased correlated motion of MDM2 around the Y100 region (refer to SI Figure S2), which has been hypothesized to be involved in downstream signaling.[26]

CONCLUSION

In conclusion, we believe that the time series of binding free energy generated using the MM-(GB)PBSA method is a very useful data set that could be used to identify distinct conformational populations. MM-(GB)PBSA is well suited for this in that no additional simulations need to be done and the calculations are performed relatively quickly. It is increasingly being recognized that binding is modulated through an ensemble of states.[27] Identifying these macrostates as has been demonstrated here should be a very useful technique in the field of computational biology.

Our results indicate that while moderately long segments (10 ns) in a trajectory can produce converged MM-(GB)PBSA results, these results may not be representative of a longer trajectory, which typically contain many other segments of different MM-(GB)PBSA profiles. More importantly, we note that the traditional way of reporting point estimators should be discouraged in favor of graphical means that more accurately depict the range of data values. Where graphs are inappropriate, both the standard deviation and standard error of the mean should be reported in full.

We also note that there is a possibility for extending our method to analyze time series generated by other free energy estimators, or to simply analyze an energetic component of the MM-(GB)PBSA method such as electrostatics.

ASSOCIATED CONTENT

Supporting Information

Specific details of the *MMPBSA_segmentation* method and the specific parameters used in the MM-GBSA calculation; individual $\psi(T18)/\psi(E28)$ plots for each identified subpopulation, graphical representation of the statistical inefficiency method as applied on the data set, and other plots/tables regarding the structural properties of the subpopulations. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: gfuentes@bii.a-star.edu.sg (G.F.); chandra@bii.a-star.edu.sg (C.S.V.).

Author Contributions

‖These authors contributed equally to this work

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was done with the support of the Bioinformatics Institute (A*STAR) in Singapore. The Biomedical Sciences Institutes (BMSI), Singapore, are acknowledged as a funding source.

REFERENCES

(1) Bash, P.; Singh, U.; Langridge, R.; Kollman, P. Free energy calculations by computer simulation. *Science* **1987**, *236* (4801), 564−568.

(2) Shirts, M. R.; Mobley, D. L.; Chodera, J. D. Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? *Annu. Rep. Comput. Chem.* **2007**, *3*, 41−59.

(3) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33* (12), 889−897.

(4) Wang, J.; Hou, T.; Xu, X. Recent Advances in Free Energy Calculations with a Combination of Molecular Mechanics and Continuum Models. *Curr. Comput.-Aided Drug Des.* **2006**, *2* (3), 287−306.

(5) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, (in press).

(6) Basdevant, N.; Weinstein, H.; Ceruso, M. Thermodynamic Basis for Promiscuity and Selectivity in Protein−Protein Interactions: PDZ Domains, a Case Study. *J. Am. Chem. Soc.* **2006**, *128* (39), 12766−12777.

(7) Genheden, S.; Ryde, U. How to obtain statistically converged MM/GBSA results. *J. Comput. Chem.* **2010**, *31* (4), 837−846.

(8) Altman, D. G.; Bland, J. M. Standard deviations and standard errors. *BMJ.* **2005**, *331* (7521), 903.

(9) Flyvbjerg, H.; Petersen, H. G. Error estimates on averages of correlated data. *J. Chem. Phys.* **1989**, *91* (1), 461−466.

(10) Grossfield, A.; Zuckerman, D. M. Chapter 2 Quantifying Uncertainty and Sampling Quality in Biomolecular Simulations. *Annu. Rep. Comput. Chem.* **2009**, *5*, 23−48.
(11) McGillick, B. E.; Balius, T. E.; Mukherjee, S.; Rizzo, R. C. Origins of Resistance to the HIVgp41 Viral Entry Inhibitor T20. *Biochemistry* **2010**, *49* (17), 3575−3592.
(12) Bishop, M.; Frinks, S. Error analysis in computer simulations. *J. Chem. Phys.* **1987**, *87* (6), 3675−3676.
(13) Genheden, S.; Ryde, U. Comparison of the Efficiency of the LIE and MM/GBSA Methods to Calculate Ligand-Binding Energies. *J. Chem. Theory Comput.* **2011**, *7* (11), 3768−3778.
(14) Stoica, I.; Sadiq, S. K.; Coveney, P. V. Rapid and Accurate Prediction of Binding Free Energies for Saquinavir-Bound HIV-1 Proteases. *J. Am. Chem. Soc.* **2008**, *130* (8), 2639−2648.
(15) Brown, C. J.; Dastidar, S. G.; Quah, S. T.; Lim, A.; Chia, B.; Verma, C. S. C-Terminal Substitution of MDM2 Interacting Peptides Modulates Binding Affinity by Distinctive Mechanisms. *PLoS ONE* [Online] **2011**, *6* (8), e24122. doi:10.1371/journal.pone.0024122
(16) Liu, Y.; Lane, D.; Verma, C. Systematic mutational analysis of an ubiquitin ligase (MDM2)-binding peptide: computational studies. *Theor. Chem. Acc.* **2011**, *130* (4), 1145−1154.
(17) Case, D. A.; Darden, T. A.; Cheatham, T. E. III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A., *AMBER*, 11; University of California, San Francisco, 2010.
(18) Said, S. E.; Dickey, D. A. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* **1984**, *71* (3), 599−607.
(19) Auger, I. E.; Lawrence, C. E. Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **1989**, *51* (1), 39−54.
(20) Donoho, D. L.; Johnstone, I. M. Adapting to Unknown Smoothness via Wavelet Shrinkage. *J. Am. Stat. Assoc.* **1995**, *90* (432), 1200−1224.
(21) Haar, A. Zur theorie der orthogonalen funktionensysteme. *Math. Annal.* **1910**, *69* (3), 331−371.
(22) Miller, D. W.; Dill, K. A. Ligand binding to proteins: The binding landscape model. *Protein Sci.* **1997**, *6* (10), 2166−2179.
(23) Tsai, C.-J.; Kumar, S.; Ma, B.; Nussinov, R. Folding funnels, binding funnels, and protein function. *Protein Sci.* **1999**, *8* (6), 1181−1190.
(24) ElSawy, K. M.; Twarock, R.; Verma, C. S.; Caves, L. S. D. Peptide Inhibitors of Viral Assembly: A Novel Route to Broad-Spectrum Antivirals. *J. Chem. Inf. Model.* **2012**, *52* (3), 770−776.
(25) Joseph, T. L.; Madhumalar, A.; Brown, C. J.; Lane, D. P.; Verma, C. S. Differential binding of p53 and nutlin to MDM2 and MDMX: Computational studies. *Cell Cycle* **2010**, *9* (6), 1167−1181.
(26) Dastidar, S. G.; Lane, D. P.; Verma, C. S. Why is F19Ap53 unable to bind MDM2? Simulations suggest crack propagation modulates binding. *Cell Cycle* **2012**, *11* (12), 2239−2247.
(27) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure* **2009**, *17* (4), 489−498.