MONTE CARLO SIMULATIONS OF PROTEIN MODELS: AT THE INTERFACE BETWEEN STATISTICAL PHYSICS AND BIOLOGY *

T. WÜST AND D. P. LANDAU

Center for Simulational Physics,
The University of Georgia,
Athens, GA 30602 U.S.A.
E-mail: dlandau@hal.physast.uga.edu

C. GERVAIS AND YING XU

Institute of Bioinformatics,
The University of Georgia,
Athens, GA 30622 U.S.A.

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Systems of atoms or molecules with complex free energy landscapes are common for quite diverse systems in nature ranging from magnetic "glasses" to proteins undergoing folding. Although Monte Carlo methods often represent the best approach to the study of suitable models for such systems, the complexity of the resultant rough energy landscape presents particular problems for "standard" Monte Carlo algorithms because of the long time scales that result at low temperatures where behavior is "interesting". We shall first review several inventive sampling algorithms that have proven to be useful for such systems and attempt to describe the advantages and disadvantages of each. Then we shall prevent results for wide ranges of temperature, obtained primarily using Wang-Landau sampling, for three models that are physically quite distinct. For pedagogical reasons we begin with spin glasses in condensed matter physics and then consider HP "lattice proteins" in which interest comes from disciplines as diverse as statistical mechanics, statistics, and biology. We shall then close with a "realistic" model for membrane protein dimerization in the continuum. All of these results will demonstrate advances in our understanding of the behavior of diverse systems that possess rough free energy landscapes.

*This work is partially supported by grants NIH-1R01GM075331 and NSF-DMR-0810223.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

1. Introduction

Protein folding is one of the great frontiers of the early 21st century science and methods of statistical physics are finding their way into investigation of these systems. Proteins are long linear polymers with different amino acids along the backbone and rather complicated interactions that result in rough free energy landscapes which are in many ways similar to those found in spin glass models of statistical physics. At high temperatures protein molecules are distended, but below some characteristic temperature they fold into a "native state" of low free energy. Proteins are sufficiently complicated that attempts to study them numerically rely upon simplifying the problem to one of manageable proportions yet retaining the fundamental features of the protein. Even then, investigation is non-trivial.

In this presentation we shall describe the advances that have been made using a new form of Monte Carlo simulation for two different systems: the HP (hydrophobic-polar) model, a very simplified lattice protein model that was introduced by biochemists over 20 years ago, and a continuum model for glycophorin A dimerization. In spite of its simplicity, the HP model is quite challenging to study, particularly as it attempts to fold into the native state. Because of their small size, these lattice proteins are at the intersection of biology, nanoscience, and statistical physics. In this manuscript we shall review numerical studies of lattice proteins with different sequences and introduce a new Monte Carlo method that performs a random walk in energy space that is highly efficient for such problems. We shall then consider our large scale Monte Carlo studies of several different HP proteins. Finally, we shall describe the application of Wang-Landau sampling to the problem of dimerization of glycophorin A in a membrane which is even more challenging in some ways because of the continuous degrees of freedom.

1.1. *"Wang-Landau" sampling*

While the "classic" Metropolis Monte Carlo algorithm continues to be widely used, a quite different Monte Carlo algorithm offers substantial advantages [1] in simplicity, broad applicability, and performance, particularly for systems with rough energy landscapes. (Originally termed the "random walk in energy space with a flat histogram" method, the technique is now referred to in the simulational physics community simply as "Wang-Landau sampling".) Unlike "traditional" Monte Carlo methods that generate canonical distributions at a given temperature $P(E) \sim g(E)e^{-E/k_BT}$,

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

where $g(E)$ is the density of states, this method estimates $g(E)$ directly (iteratively) and then uses it for the estimate of canonical averages for thermodynamic quantities at any temperature, e.g. the free energy

$$F(T) = -k_{\mathrm{B}}T\ln(Z) = -k_{\mathrm{B}}T\ln(\sum_{E} g(E)e^{-\beta E}). \tag{1}$$

and its derivatives.

If a random walk in energy space is performed with a probability proportional to $\frac{1}{g(E)}$, then a flat histogram will be generated for the energy distribution. This is done by modifying the estimated $g(E)$ systematically to produce a "flat" histogram over the allowed range of energy and simultaneously making it converge to the correct value from some initial estimate, e.g. $g(E) = 1$ for all E. The random walk proceeds by randomly applying a trial move (e.g. flipping spins in a spin glass model) with transition probability

$$p(E_1 \to E_2) = \min(\frac{g(E_1)}{g(E_2)}, 1). \tag{2}$$

where E_1 and E_2 are energies before and after having performed the move. Each time an energy level E is visited, $g(E)$ is updated by multiplying the existing value by a *modification factor* $f > 1$, i.e. $g(E) \to g(E) * f$. A histogram of the energies that are "visited" is also updated; and when $H(E)$ is approximately "flat", f is reduced, e.g. $f_1 = \sqrt{f_0}$, the histogram is reset to $H(E) = 0$ for all E, and a new random walk is begun. This process is iterated n times, until f_n is smaller than some predefined final value (e.g. $f_{\mathrm{final}} = \exp(10^{-8}) \simeq 1.00000001$). The final results are normalized, and if multiple walks are performed within different energy ranges, they must be matched up at the boundaries in energy. During the early stages of iteration the algorithm does not satisfy detailed balance since $g(E)$ is modified continuously; however, detailed balance is recovered to high precision after many iterations.

The final accuracy of $g(E)$ is controlled by two parameters: the final modification factor f_{final} and the flatness criterion p. Whereas Wang-Landau studies of some polymeric systems reported that $\ln(f_{\mathrm{final}}) \simeq 10^{-6}$ is sufficient, we found that for the HP model, reliable DOS estimates over the entire energy range (including **the** lowest energies) required $\ln(f_{\mathrm{final}}) \leq 10^{-7}$. We, thus, used a very stringent parameter set for our simulations of the HP model: $\ln(f_{\mathrm{final}}) = 10^{-8}$ and $p = 0.8$. Knowledge of the exact energy range is needed for the examination of the flatness of the histogram, but the energy boundaries for polymer/protein models are *a*

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

priori unknown. (For this reason, there has been substantial use of ground state search algorithms, e.g. for the HP model). To overcome this difficulty, the following procedure was used: Each time a new energy level $E_{\rm new}$ was found, it was marked as "visited" and $g(E_{\rm new})$ was set to $g_{\rm min}$, i.e. the minimum of g among all previously visited energy levels. The flatness of the histogram is checked only for those energy levels which have been visited. With this self-adaptive procedure, new regions of conformational space can be explored simultaneously as the current estimate of the density of states is further refined.

Wang-Landau sampling is a very flexible, highly efficient **and robust** Monte Carlo algorithm for the determination of the density of states of quite diverse statistical physical systems [1,2,3,4]. To demonstrate the effectiveness of the algorithm, in Fig.1 we show the dramatic variations in the canonical probabilities of states in the Edwards-Anderson (EA) spin glass as determined from a **single simulation**. (This system has been a great challenge in statistical physics for decades.)

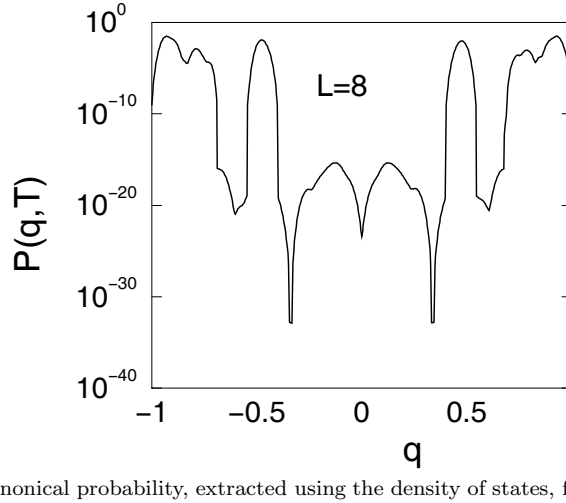

Figure 1. Canonical probability, extracted using the density of states, for the Edwards-Anderson spin glass in three dimensions. The result is for a temperature of $T/J = 0.1$ and the order parameter for this model is q.

Several different methods with a computer science flavor have been applied to the HP model with varying degrees of success. In addition, the configurations of an HP polymer may be studied using the traditional Metropolis method, but the complexity of the resultant free energy surface

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

at low temperature renders the method extremely inefficient. Multicanonical Monte Carlo (i.e. sampling with a modified, fixed probability) can also be applied to the system and is more efficient than Metropolis. In our Wang-Landau sampling we have determined that traditional trial moves, e.g. end-flips, kink-flips, crankshaft, and pivots are inadequate in the collapsed state, so we have implemented a new set of trial moves that turns out to be very efficient, namely a combination between pull moves [5] and bond-rebridging moves [6]. These moves complement each other extremely well, as pull moves allow the polymer chain to naturally and fastly fold/unfold, while bond-rebridging moves change the polymer's configuration even at high densities. Our approach has been to implement Wang-Landau sampling together with these new types of trial moves to determine the density of states over the entire range of energies in a single run [3]. In addition to allowing us to find the ground state (or "native state" in biological language) we can determine thermodynamic properties as a function of temperature.

2. What have we already learned from simulations?

2.1. *The HP model*

In the hydrophobic-polar (HP) lattice model [7] the protein is represented as a self-avoiding chain of beads (i.e. coarse grained representations of the amino acid residues) on a rigid lattice. Only two types of amino acids - hydrophobic (H) and polar (P) - are included and an attractive interaction ϵ acts only between non-bonded neighboring H residues (i.e. $\epsilon_{HH} = -1, \epsilon_{HP} = \epsilon_{PP} = 0$). Different sequences of H- and P- monomers are used to "match" different proteins, and several sequences which have been deemed "benchmarks" have been studied extensively by a variety of methods. At low temperatures the HP chains fold into compact structures, but there are many different folds with different energies which are separated by free energy barriers. For this reason we have the same kinds of difficulties that one encounters with spin glass models of magnetism, and the location of the true ground state, or native state, becomes a difficult optimization problem.

We simulated various HP benchmark sequences found in the literature, emphasizing longer sequences like a 103mer in three dimensions (3D103) or a 100mer in two dimensions (2D100b). (Two dimensional models are of substantial significance within the context of statistical physics and could be physically realized if they were placed on a strongly attractive surface.) A number of different HP sequences are defined elsewhere[8]. The ground

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

states of sequence 2D100b are believed to have an energy $E = -50$ [9], and various methods have confirmed this result [10,5,8]. However, previous attempts to obtain $g(E)$ over the entire energy range $[-50, 0]$ within a single simulation have failed [9,11]. In contrast, with our approach we were able to achieve this with high accuracy and Fig. 2 shows the resulting specific heat $C_V(T)/N$, depicting a peak at $T \approx 0.48$ (coil-globule transition) and a very weak shoulder at $T \approx 0.23$ (folding transition)[3].

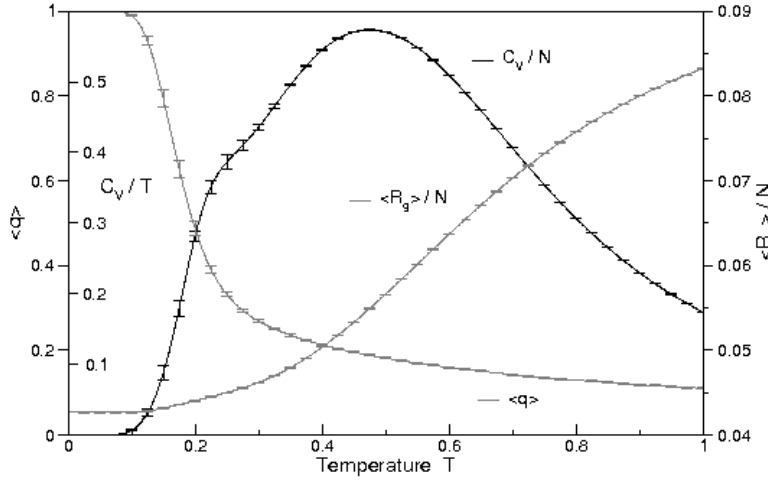

Figure 2. Specific heat C_V/N, mean radius of gyration $\langle R_g \rangle / N$ (N, chain length) and mean Jaccard index $\langle q \rangle$ as a function of temperature T for the two dimensional HP sequence 2D100b. Statistical errors were calculated by a Jackknife analysis from 15 independent Wang-Landau (C_V) and multicanonical production runs ($\langle R_g \rangle$ and $\langle q \rangle$) for each sequence.

As we shall see later, such two-step acquisition of the ground (native) state has been observed in studies of realistic protein models, e.g. glycophorin A, and is not restricted to lattice models. For different HP chains the relative locations of the peak and shoulder differ, but the general features do not.

For sequence 3D103, the lowest energy found by fragment regrowth Monte Carlo via energy-guided sequential sampling (FRESS) [8] was $E = -57$, but with our approach, we discovered an even lower state with energy -58 [3]. It was also possible to determine the density of states in the energy range $[-57, 0]$, from a single simulation, and with very high accuracy, although it was not possible to determine the relative magnitudes of the

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

ground state and 1st excited state with high precision. Figure 3 displays the specific heat for 3D103, showing a peak at $T \approx 0.51$ and a shoulder at $T \approx 0.27$[3].

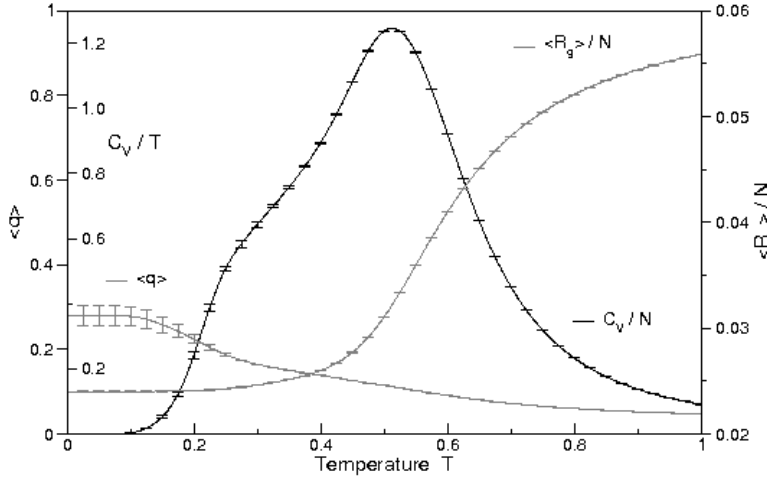

Figure 3. Specific heat C_V/N, mean radius of gyration $\langle R_g \rangle / N$ (N, chain length) and mean Jaccard index $\langle q \rangle$ as a function of temperature T for the three dimensional HP sequence 3D103. Statistical errors were calculated by a Jackknife analysis from 15 independent Wang-Landau (C_V) and multicanonical production runs ($\langle R_g \rangle$ and $\langle q \rangle$) for each sequence.

After determining $g(E)$ with high resolution we used it in a multicanonical run to measure the radius of gyration R_g [12] and the Jaccard index q [13], i.e. the similarity between a conformation and the ground state of a HP sequence:

$$q = \max \left\{ \frac{C_{s,g}}{C_{s,g} + C_s + C_g} \mid E_g = E_{\min} \right\}. \tag{3}$$

$C_{s,g}$ denotes the number of common (native) H-H contacts between a conformation s and the ground state g, and C_s, C_g are the numbers of H-H contacts found only in s and g, respectively, (the maximum stems from the degeneracy of ground states). Figures 2 and 3 also show the temperature averages $\langle R_g \rangle$ and $\langle q \rangle$ for sequences 2D100b and 3D103 and illustrates the complementary information provided by these two quantities. While $\langle R_g \rangle$ indicates the coil-to-globule collapse, $\langle q \rangle$ identifies the folding transition to the native state and may thus serve as a structural order parameter. In the

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

case of the sequence 3D103, the ground state ($E = -58$) was excluded (due to the difficulty of finding the relative density for this state) which results in only a rather small Jaccard index for $T \to 0$. This shows that there are still large structural differences between conformations with $E = -57$ and the ground state with $E = -58$.

TABLE 1 compares results obtained using various methods, and, if available, the $g(E)$ for common benchmark HP sequences[3]. We also include results from methods which were focused on the low temperature range only, i.e. FRESS [8] and the variants of PERM (pruned-enriched Rosenbluth method) [10] and do not provide the entire density of states.

Table 1. Energy minima ($E_{\min}$) found by several sophisticated numerical methods for benchmark HP sequences in 2D and 3D. The first column names the sequence (dimension and length).

	WLS	EES[11]	$MCCG$[14]	$MSOE$[9]	$FRESS$[8a]	$PERM$[10a]
$2D100a$	-48	-48	-	-	-48	-48
$2D100b$	-50	-49	-	-50[b]	-50	-50
$3D88$	-72	-	-	-	-72	-69
$3D103$	-58	-	-56	-	-57	-55

Note: [a] Ground state search only.
[b] g(E) not found.

Figure 4 shows the ground state structure for 3D103. The hydrophobic core is easy to see and the overall structure is quite symmetric. Observation shows that "antiphase" domains form near the shoulder so that portions of the native state are shifted with respect to other portions, and at the higher temperature peak the protein structure becomes distended.

2.2. *Glycophorin A*

We have used Wang-Landau sampling to investigate the behavior of glycophorin A [15]. The system contains two identical α-helices, A and B, of 22 residues each, running from $E72$ to $Y93$ (`EITLIIFGVMAGVIGTILLISY`). In order to decrease the large number of degrees of freedom inherent to this system, the backbone was kept fixed during the simulations and a unified atom representation was employed, in which only the heavy atoms and those polar hydrogen atoms susceptible to being involved in hydrogen bonding are explicitly modelled (a total of 378 atoms). Besides, the

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

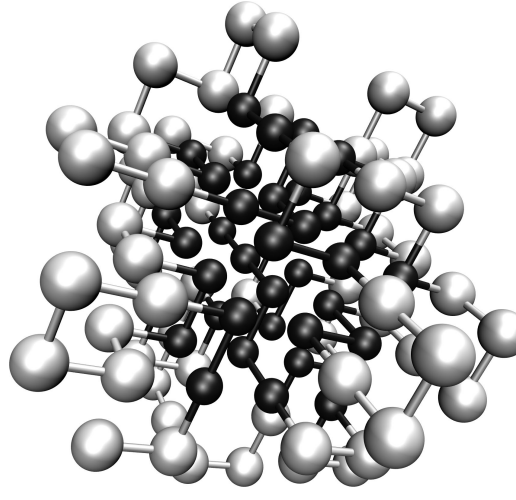

Figure 4. Native state structure for the 3D103 HP model. Hydrophobic residues are the small, dark balls and the large grey balls represent polar residues.

membrane was represented implicitly, i. e. the interaction between membrane and protein was treated by a mean-field (see E_{lipid} below). Like for the HP model, designing appropriate moves that allow the algorithm to search the entire energy landscape efficiently is crucial for the success of the simulation. Here, a total of seven trial moves were employed, designed to allow either global modifications of the protein or local changes in the conformation[16].

The energy is based on the CHARMM19 force-field [17] and a knowledge-based potential designed to take into account the membrane environment implicitly [18,19]. The energy is then given by

$$E = E^{A,B}_{inter} + E^{A}_{intra} + E^{B}_{intra} + E^{A}_{lipid} + E^{B}_{lipid}, \tag{4}$$

where $E^{A,B}_{inter}$ is the sum of the van der Waals and electrostatic energies between atoms of the helix A and those in helix B; E^{A}_{intra} and E^{B}_{intra} define the sum of the van der Waals, electrostatic and dihedral energies within the helix A and B, respectively, see [17]. Finally, E^{A}_{lipid} and E^{B}_{lipid} are the sums of the lipid-residue interactions of helix A and B, respectively. The lipid energy is a function of the z-coordinates of the C_α atoms of the residues,

$$E_{lipid} = \sum_{residues} \mu_{type}(|z|). \tag{5}$$

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

It reflects the propensity of an amino-acid to be located in different regions of the membrane. Because of this simplified representation of the membrane-protein interaction the values for transition temperatures obtained below should not be taken too seriously as the flexibility and deformation of the membrane at high temperature are not included in our model. We can, however, still draw conclusions about the characteristic behavior of the dimer, especially at low and medium temperatures where membrane-protein interactions are likely to play a limited role on the thermodynamics of the system [20].

The computed density of states illustrates the complexity of the system as a large range of $g(E)$, spanning nearly 110 orders of magnitude, was obtained. Such complexity requires making some compromises to reduce the computation time while keeping good thermodynamic resolution. In our case, one way to achieve this is to neglect the few bins with a very low energy. Indeed, the density of states for these bins is likely to have only a negligible influence on the thermodynamic quantities at temperatures of interest (i.e. $T > 200$K). For the glycophorin A system, the lowest energy found during a simulation was -666.7 kcal/mol. Restricting the energy range from -665 to -300 kcal/mol with a bin width $\Delta E = 1$kcal/mol, allowed all runs to converge within about 100 CPU hours per run (AMD Opteron processors) yet still provided good thermodynamic resolution. Once the density of states is obtained, one can also gather thermodynamic information for all types of observables, both energetic and structural. For that purpose, it suffices to run a second simulation with Wang-Landau sampling, however without updating $g(E)$ anymore. This so-called production run enables efficient sampling of the observables, including in conformational regions with low energies[16]. In our case, we sampled the inter-helix interaction ($E_{inter}^{A,B}$) and studied global structural changes by looking at the distance between the center of mass of the two helices ($d^{A,B}$) and the root-mean square deviation ($RMSD$) of the C_{alpha} atoms with respect to the experimental reference structure (model 1 of the NMR structure with PDB code 1AFO [15]). All results presented below were obtained by running and averaging five production runs of 8×10^8 MC moves each (about 100 CPU hours per run), a simulation length found sufficient to ensure the reliability of our results.

Figure 5 shows the specific heat of the glycophorin A (C_V), the thermodynamic average of $\langle E_{inter}^{A,B} \rangle$ and its derivative as a function of temperature. The specific heat shows a clear, rounded peak at $\sim$ 800K followed by a shoulder around $\sim$ 300K. While there is little effective interaction

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

between the helices at 1100K, the system undergoes a dimerization transition at $\sim$ 800K characterized by a significant peak in the derivative. However, $d\langle E_{inter}^{A,B}\rangle/dT$ does not show another peak at lower temperatures, indicating that the inter-helix interaction is not responsible for the shoulder observed in the specific heat at 300K. The two structural observables

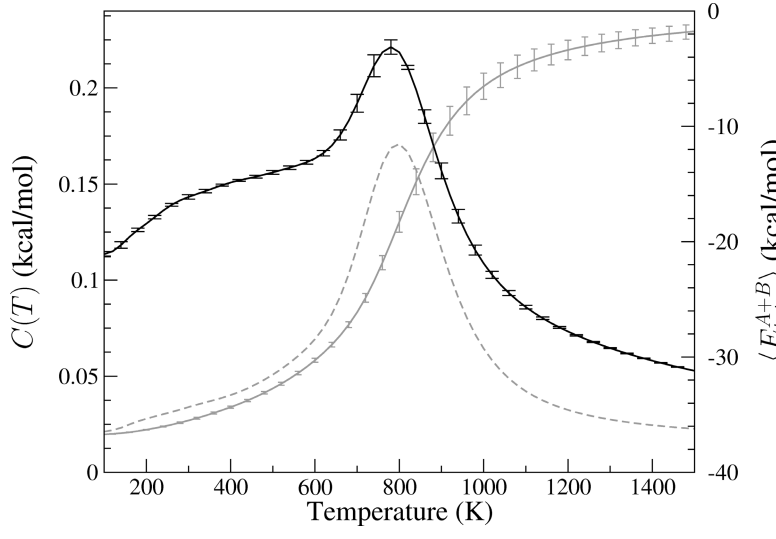

Figure 5. Black: Specific heat $C(T)$ of glycophorin A. Statistical errors were estimated using a Jackknife analysis from ten runs. Note that only the temperature range for which results are reliable (i.e. T > 100K) is shown. Gray: Variation of the inter-helix energy $\langle E_{inter}^{A,B}\rangle$ (solid line) and its derivative (dashed line) as a function of temperature.

$\langle d^{A,B}\rangle$ and $\langle RMSD\rangle$ show a similar behavior at high temperatures (results not shown). A large structural transition slightly above 800K is observed, indicating that close contact between the helices has established. However, contrary to $\langle E_{inter}^{A,B}\rangle$, an additional peak is observed around 300K. At this temperature, the RMSD has fallen below 2Å and the distribution of $d^{A,B}$ shows a preference for $d^{A,B} \approx 6.7$Å, a value close to the native distance of 6.64Å[16]. These observations clearly suggest that this second peak corresponds to the association of the helices into a native-like conformation. Indeed, representative structures at this transition temperature (300K) were found to be similar to the experimental structure (see Fig. 6).

The motif assumed to be responsible for the dimerization of the homodimer glycophorin A is composed of 7 residues `LIxxGVxxGVxxT` [21]. The

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

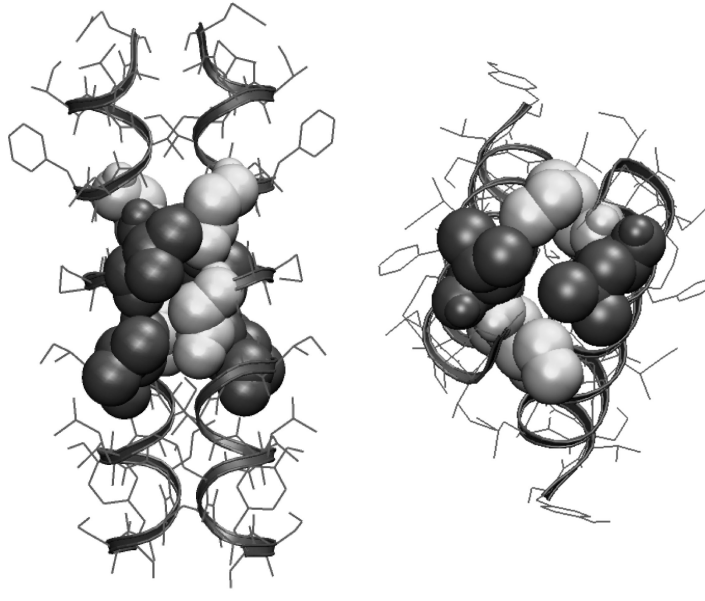

Figure 6. Typical structure for glycophorin A at 300K: *(Left)* Side view. Glycine residues G79 and G83 (white) pack densely with valine residues V80 and V84 (gray); *(Right)* Top view. Isoleucine I76 (white) and leucine L75 (gray) form a hook which stabilizes the dimer interface via hydrophobic interactions and close contact packing between branched residues.

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

motif `GxxxG`, well known to promote dimerization in membrane proteins and to favor helix-helix interactions in soluble proteins [22], acts as an anchor point (Fig. 6, left). The glycine residues G79 and G83 facilitate the approach of the two helices because of their small size and their minimal entropic contribution (absence of side chain). Hence, they allow a dense packing in a "groove and ridge" fashion with the two neighboring valines V80 and V84 [15]. Above this anchor point, leucine L75 and isoleucine I76 are necessary to stabilize the dimer. At the other end of the α-helix, threonine T87 stabilizes the dimer by forming an inter-helical hydrogen bonding [23,24].

To investigate the relative importance of these three strategic points at the dimer interface, during a production run we sampled the temperature dependence of the average energies $\langle E_{\text{motif}} \rangle$ of the three following *motifs*: the motif *leucine*, composed of the interactions of residues L75 and I76; the motif *glycine*, composed of the interactions of residues G79, V80 and

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

G83; and the motif *threonine*, composed of the interactions of the residue T87. All three motifs show qualitative thermodynamic behavior similar to that observed above, i.e. their derivatives (dE/dT) exhibit two peaks near $\sim$ 800K and $\sim$ 300K, respectively. However, the significantly different magnitudes of the peaks at the two transitions clearly indicate the difference in thermal stability of the three motifs. Whereas $E_{leucine}$ undergoes mainly a transition at 800K, $E_{glycine}$ shows similar peak ratios at both 300K and 800K, and $E_{threonine}$ has a major transition at 300K. This suggests the following order of stability: *leucine* $>$ *glycine* $>$ *threonine*.

From the results for both global and motif-based observables, we can describe the structural and energetic changes taking place during the dimerization process of glycophorin A. At 800K, the two helices come into contact and interact with a significant inter-helix energy. Structurally, the majority of dimers found at this temperature feature some characteristics of the native structure. Indeed, the motif *leucine*, and to a lesser extent, the motif *glycine*, already exhibit significant contributions to the overall ground state energy. This observation is not surprising considering that the GxxxG motif is known to promote dimerization of many membrane proteins [21,25]. Besides, leucine L75 and isoleucine I76 form a "hook" in the native structure, which stabilizes the dimer interface via hydrophobic interactions and close contact packing between branched residues, (see Fig. 6, right). Another transition, at about 300K, corresponds to the convergence towards the native structure. The average RMSD falls below 2Å and the motif *glycine*, and especially the motif *threonine*, undergo a transition towards the native energies. Stabilization of the dimer is affected via the formation of inter-helical hydrogen bonding.

Our findings on the two-stage dimerization process of glycophorin A agree very well with the hypothesis proposed by Schneider [26] who suggested decomposing the oligomerization into two stages. First, the contact between helices is promoted by a detailed fit between the helical surfaces, leading to close packing and van der Waals interactions. In a second stage, stabilization of the preformed dimer is obtained by electrostatic interactions, i.e. hydrogen bonding, or binding of a cofactor. We found indeed a first dimerization step governed by dispersive interactions (motif *leucine*) and close packing (motif *glycine*), while the second transition involved the formation of hydrogen bonds within the motif *threonine*.

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

3. Conclusions

Wang-Landau sampling has proven to be effective for studying both folding of lattice HP model proteins and dimerization of glycophorin A. The same basic techniques that are used for models in statistical physics work well, and in both cases a two-step process leads to the low temperature native state. However, it is interesting to see the different level of accuracy needed by the two models to obtain significant results. The HP model, simple in appearance, requires accurate estimation of the density of states to obtain sufficient details about the folding transition. To the contrary, information on the dimerization of glycophorin A can be well tackled by using a rough estimate of the density of states followed by multicanonical sampling of structural and energetical observables.

Acknowledgements

We wish to thank S.-H. Tsai and D. Seaton for valuable discussions.

References

1. F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001); Phys. Rev. E **64**, 056101 (2001); Comput. Phys. Commun. **147**, 674 (2002).
2. D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, UK, 2005), 2nd ed.
3. T. Wüst and D. P. Landau, Phys. Rev. Lett. **102**, 178101 (2009).
4. T. Wüst and D. P. Landau, Comput. Phys. Commun. **179**, 124 (2008).
5. N. Lesh, M. Mitzenmacher, and S. Whitesides, in *RECOMB* (2003), p. 188.
6. J. M. Deutsch, J. Chem. Phys. **106**, 8849 (1997).
7. K. A. Dill, Biochemistry **24**, 1501 (1985); K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
8. J. Zhang, S. C. Kou, and J. S. Liu, J. Chem. Phys. **126**, 225101 (2007).
9. Y. Iba, G. Chikenji, and M. Kikuchi, J. Phys. Soc. Jpn. **67**, 3327 (1998); Phys. Rev. Lett. **83**, 1886 (1999).
10. H. Frauenkron *et al.*, Phys. Rev. Lett. **80**, 3149 (1998); U. Bastolla *et al.*, Proteins **32**, 52 (1998); H.-P. Hsu *et al.*, Phys. Rev. E **68**, 021113 (2003).
11. S. C. Kou, J. Oh, and W. H. Wong, J. Chem. Phys. **124**, 244903 (2006).
12. A. D. Sokal, in *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, edited by K. Binder (Oxford University Press, New York, Oxford, 1995), p. 47.
13. R. Fraser and J. I. Glasgow, in *ICANNGA (1)* (2007), p. 758.
14. M. Bachmann and W. Janke, Phys. Rev. Lett. **91**, 208105 (2003); J. Chem. Phys. **120**, 6779 (2004).
15. K. R. MacKenzie, J. H. Prestegard, and D. M. Engelman, Science **276**, 131 (1997).

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.

16. C. Gervais, T. Wüst, D. P. Landau, and Y. Xu, J. Chem. Phys. **130**, 215106 (2009).
17. E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).
18. Z. Chen and Y. Xu, Proteins **62**, 539 (2006).
19. Z. Chen and Y. Xu, J. Bioinform. Comput. Biol. **4**, 317 (2006).
20. J. U. Bowie, Curr. Opin. Struct. Biol. **11**, 397 (2001).
21. B. Brosig and D. Langosch, Protein Sci. **7**, 1052 (1998).
22. G. Kleiger, R. Grothe, P. Mallick, and D. Eisenberg, Biochem. **41**, 5990 (2002).
23. S. O. Smith, M. Eilers, D. Song, E. Crocker, W. Ying, M. Groesbeek, G. Metz, M.Ziliox, and S. Aimoto, Biophys. J. **82**, 2476 (2002).
24. J. M. Cuthbertson, P. J. Bond, and M. S. P. Sansom, Biochem. **45**, 14298 (2006).
25. W. P. Russ and D. M. Engelman, J. Mol. Biol. **296**, 911 (2000).
26. D. Schneider, FEBS Lett. **577**, 5 (2004).

Copyright © 2010. World Scientific Publishing Co Pte Ltd. All rights reserved.

Mondaini, Rubem P. Biomat 2009 : International Symposium on Mathematical and Computational Biology, World Scientific Publishing Co Pte Ltd, 2010. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/eth/detail.action?docID=731307.
Created from eth on 2019-01-05 12:33:09.