A New Approach In Protein Folding Studies Revealed The Potential Site For Nucleation Center

Nurul Bahiyah Ahmad Khairudin, Habibah A Wahab

***Abstract*—**A new approach to predict the 3D structures of proteins by combining the knowledge-based method and Molecular Dynamics Simulation is presented on the chicken villin headpiece subdomain (HP-36). Comparative modeling is employed as the knowledge-based method to predict the core region (Ala9-Asn28) of the protein while the remaining residues are built as extended regions (Met1-Lys8; Leu29-Phe36) which then further refined using Molecular Dynamics Simulation for 120 ns. Since the core region is built based on a high sequence identity to the template (65%) resulting in RMSD of 1.39 Å from the native, it is believed that this well-developed core region can act as a 'nucleation center' for subsequent rapid downhill folding. Results also demonstrate that the formation of the non-native contact which tends to hamper folding rate can be avoided. The best 3D model that exhibits most of the native characteristics is identified using clustering method which then further ranked based on the conformational free energies. It is found that the backbone RMSD of the best model compared to the NMR-MDavg is 1.01 Å and 3.53 Å, for the core region and the complete protein, respectively. In addition to this, the conformational free energy of the best model is lower by 5.85 kcal/mol as compared to the NMR-MDavg. This structure prediction protocol is shown to be effective in predicting the 3D structure of small globular protein with a considerable accuracy in much shorter time compared to the conventional Molecular Dynamics simulation alone.

***Keywords*—**3D model, Chicken villin headpiece subdomain, Molecular dynamic simulation NMR-MDavg, RMSD.

I. Introduction

THE available approaches to predict the protein structure rely on two distinct sets of principles; the laws of physics mainly employing Molecular Dynamics (MD) Simulation and the theory of evolution which gives rise to comparative modeling. To date, comparative modeling remains the only accurate knowledge-based prediction method. However, it is limited to proteins that share a certain degree of sequence similarity with other protein templates [1, 2]. On the other hand, all-atom MD folding simulations do not yet seem to be able to provide high-resolution information for the majority of proteins. Furthermore, it is extremely expensive as it needs to simulate beyond the microsecond time regime, which is the minimum bound for proteins to fold. Thus far, all-atom MD folding simulation is currently limited to small proteins and peptides [3-6]. This study intends to combine the knowledge-based and the physics-based MD folding simulation. The general idea of the present work is to model the core region of the protein using information from the template structure whereas leaving the end-terminal regions to fold via MD simulation which up to the knowledge, has never been reported elsewhere. Therefore, this two-step approach represents an alternative method in the structure prediction of small proteins by presenting a test-case application to the fast folder, villin headpiece subdomain (HP-36; PDB id: 1VII) with estimated folding time of 5μs [7]. The rationale behind the selection of this protein is due to its small size and its ability to fold in such a short time making it one of the most investigated systems for protein folding and protein structure prediction studies.

[1]Nurul Bahiyah Ahmad Khairudin is with the Bioprocess Engineering Department, Faculty of Chemical Engineering and Natural Resources Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Malaysia (e-mail: nurul@cheme.utm.my).

[2]Habibah A Wahab, was with School of Pharmaceutical Sciences, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia

II. Procedure

A. MD simulation on native NMR

The protein coordinate of HP-36 was obtained from the PDB [8]. The protein was immersed in a truncated octahedron water box containing 2335 molecules of TIP3P water [9] and two chloride ions to maintain the system neutrality. The system was then minimized employing 500 and 1500 cycles of steepest descent and the conjugate gradient methods, respectively. The system was then further subjected to 50 ns of MD simulation. It was initially heated from 0K to 300 K in 40 ps at constant volume and further equilibrated at constant pressure (1 bar) and constant temperature (300 K) using the Berendsen weak-coupling thermostat [10] with coupling constants of 1 ps.

The production phase was started from an equilibration phase of 960 ps at 300K and 1 bar of pressure with the system density comply with the density of liquid water. The nonbonded interactions were treated using 10 Å cutoff and PME algorithm for Lennard Jones and coulomb interactions respectively. Both of the energy minimization and MD simulation were carried out using AMBER8 [11] suite of programs utilizing the force field amber.ff03 [12].

B. Development of the core region

The 36-residues linear amino acid chain of HP-36 was subjected to sequence analysis using the web-interface

BLAST [13] to locate for the appropriate template. The sequence alignment between the template and the target was performed using CLUSTALW [14] and the 3D model was built using the program Modeller7v7 [15].

C. Development of complete protein

All the remaining residues that were not modeled were added to both end regions of the core structure using the program Modeller7v7. The developed model was then further subjected to 120 ns of MD simulation with the same condition as described for the native NMR. The complete protocol of this combined method is summarized in Figure 1.

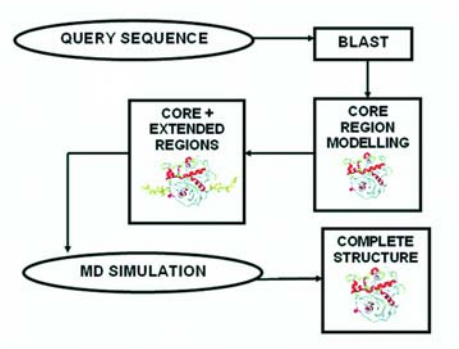

Fig. 1 Flowchart of the combined protocol. This method combines comparative modeling for the development of the core region and MD simulation for the complete 3D model generation and structure refinement.

D. Data analyses

RMSD and hydrogen bond analysis were each calculated using the ptraj module implemented in Amber8. The software NACCESS [16] was used to calculate the solvent accessible surface area (SASA). Tertiary native contacts, radius of gyration (Rgyr) and the clustering analyses were carried out using the tools from the MMTSB program [17]. The conformational free energies were estimated using the Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) protocol [18, 19] shown by Equation 1. The average free energy (ΔG) was calculated as a sum of the average gas phase molecular mechanical energies (ΔEMM) and the average of the solvation free energies (ΔGsolv). The entropy contribution was not calculated since it was previously proven that this term varied negligibly between the trajectories [20, 21] and furthermore, it was too expensive to calculate.

$$\Delta G = \Delta EMM + \Delta Gsolv \quad (1)$$

III. Result & Discussion

A. Development of the core region

From the BLAST result, the NMR structure of the human villin C-terminal headpiece subdomain (PDB id:1UNC) was randomly chosen as the modeling template. The sequence alignment between these two proteins is shown in Figure 3. The local and global percentages of sequence identity are 65% and 36%, respectively. The developed 3D model (Ala9 to Asn28) is observed to contain two native α helices. Helix 2 consist of Arg15 to Asn20 (NMRnative: Arg15- Phe18) while Helix 3 is made up of Leu23 to Asn28 (NMRnative: Leu23-Lys30). The backbone RMSD (RMSDback) with reference to the NMRnative is 1.39 Å.

Fig. 2 Sequence alignment between the target HP-36 (1VII) and the template 1UNC. The yellow box region corresponds to the core region of the model (Ala9 – Asn28) with 65% of sequence identity. Asterisks represent conserved residues.

B. Development of the complete protein

Structural Properties : Figure 3(b) shows the predicted 3D model of HP-36 with the added residues (Met1-Lys8; Leu29-Phe36) shown as yellow coils. Prior to MD simulation, the RMSDback and the Rgyr are calculated to be 6.2 Å and 12.1 Å, respectively as compared to the native structure (NMR-MDavg). When subjected to MD, these extended loops are slowly pulled towards the core region by forming contacts with surrounding residues. The all residues backbone RMSD (RMSDback-all) rapidly decreases from an initial value of 6.2 Å to ~3.34 Å within 10 ns (Figure 3(a)). The RMSDback-all somehow starts to increase back to 4.5 Å up to13.8 ns before decreases again to 3.24 Å. The conformation is stable until 81 ns where a sudden increase in the RMSDback-all is observed (5.34 Å) which then rapidly decreases back to 3.5 Å for the remaining simulation. The evolution of RMSD for the core region (RMSDback-core) is observed to be fairly stable with value varies between 0.66 Å to 1.56 Å. This finding indicates that the core region has probably achieved the near-native state with deviation comparable to that of high resolution X-ray diffraction. The starting structure of the model is found to be distended with Rgyr ~12 Å compared to the NMR-MDavg (Rgyr ~9.5 Å). Throughout the simulation, the structure is constantly contracting and expanding with value fluctuates between 9.5 Å and 11.3 Å due to the high flexibility of the terminal regions (Figure 3(b)). As expected, there seems to be a sudden rise in the size of the structure to 12.4 Å at 81 ns which strongly correlates with the abrupt increase in the value of RMSD as noted previously. The development of the native tertiary contacts along the folding coordinates is demonstrated in Figure 3(c). Overall, the initial contacts is 54% and rapidly increases to 81% within the first 28 ns follows by wild fluctuations from 36% to 73% for the remaining of the simulation. Although it is not our intention to investigate the folding pathway, the results do suggest the involvement of the collapse phase which is also in good agreement with other studies [4, 22]. The "collapse phase" or the "burst phase" takes place when the protein quickly reaches a high level of native contacts. This phase can also be observed in the reduction of the RMSDback-all and the Rgyr during the first

30 ns. The high level of native contacts corresponds to the formation of new tertiary contacts formed by the interactions between the terminal regions and the knowledge-based core region. For instance, Val10 from the core region and Lys33 from the C-terminal region form a contact in the NMR-MDavg with a distance of 3.46 Å and this contact is also observed to form in the model at 28.1 ns with a distance of 3.44 Å. The hydrogen bond that forms between the donor Glu32@OE2 and the acceptor Val10@N mainly contributes in preserving this native tertiary contact. This hydrogen bond is very stable since it shows the highest residence time i.e., 69% among all the other hydrogen bonds observed in the system. It even reaches its highest stability in the last 10 ns with 95% occupancy. However, this bond is not present in the NMR-MDavg simulation.

The total SASA of HP-36 quickly decreases from ~3800 Å2 to ~3200 Å2 in the first 20 ns and remains fluctuated around this value (Figure 5(d)). At the beginning of the simulation, the N-terminal region rapidly moves towards the core region to protect the nonpolar part from being exposed to the solvent and this explains the reduction in SASA. Residues in the hydrophobic core region have lower SASA values compared to those on the exterior part. As compared to the nonpolar part, the polar SASA does not decrease much due to the high affinity of the hydrophilic residues to water molecules.

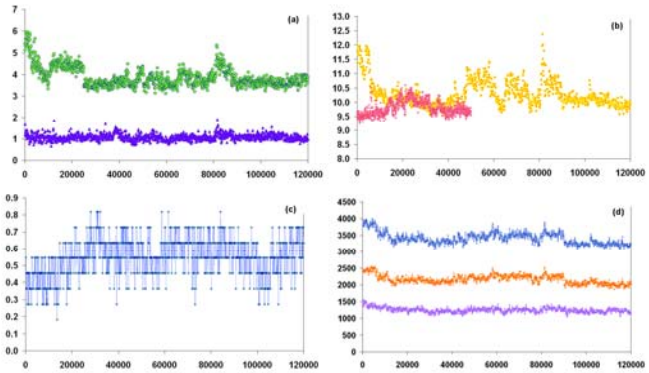

Fig. 3 Structural analyses after 120 ns of MD simulation. Time evolution of (a) backbone RMSD; Green squares represent the backbone RMSD for all residues while the purple triangles represent the RMSD for the core region (residue Ala9- Asn28). (b) radius of gyration for model (yellow) and the NMR-MDavg (pink). (c) fraction of native contacts for the model (d) total solvent accessible surface area (Å2) (blue) with the contributions from the polar (pink) and the nonpolar (orange) residues. Time in picoseconds (ps).

The three α-helices in the NMR-MDavg are made up of residues Asp4-Phe11, Arg15-Asn20 and Leu23-Lys33, respectively. However, the helices that form in our model are shorter than that of the NMR-MDavg, Helix1 (Leu2-Glu5); Helix2 (Arg15-Phe18); Helix3 (Leu23-Gln26). Another strange finding is the formation of an intermittent β-sheet by two β-strands. Up to our knowledge, this event has never been reported in the previous folding studies of HP-36. Lys8 and Ala9 form the first β-strand while the second strand is formed by residues Met13 and Thr14 connected by a hydrophobic turn comprising Val10, Phe11 and Gly12. These three nonpolar residues are rapidly drawn towards the core region in the beginning of the simulation in order to avoid contacts with water molecules on the protein surface (Figure 4).

Since Phe7 is surrounded by polar residues, it requires a hydrophobic shield from the surrounding water; and this should be provided by Phe11, Ala9 and Val10. Instead, both Val10 and Phe11 collapse to the core region causing Lys8 and Ala9 to adopt ϕ and ψ angles of that of β-sheet. It is possible that given longer simulation time, we may observe Phe7 to form a hydrophobic interaction with these residues. This perhaps will automatically disrupt the transient β-sheet to form a more stable α-helix. It is a common fact that both lysine and alanine show high propensities towards α-helices compared to β-sheets [23]. Thus, we believe that Phe7 might initiate the stable form of helix 1 as it is claimed that the coil to helix transition is driven primarily by non-polar interactions [24]. In another folding study on the same protein [25], claimed that the formation of the non-native contact between Phe36 with other phenylalanines (Phe7, Phe11 and Phe18) hindered the folding rate and they suggested that this hydrophobic contact needs to be broken in order for the protein to fold [3]. However, this phenomenon does not occur in our result as Phe36 is not making hydrophobic interaction with the core residues of the phenylalanines. If their notion holds, our protocol will have an advantage of escaping the formation of the non-native contact whose presence will delay the folding process.

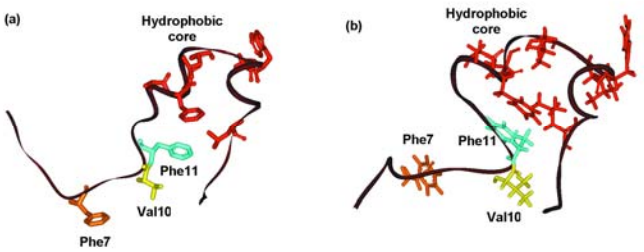

Fig. 4 Atomic interactions in the developed knowledge-based core region of HP-36. (a) The starting model prior to MD simulation; (b) Model after 100 ps of MD simulation. The collapse of Phe11 and Val10 towards the hydrophobic cluster core region is shown.

Clustering and energetic analysis:

The trajectories are subjected to clustering analysis in order to discriminate the native-like model from the non-natives or to put it differently, to locate for the 3D model that best represents the native structure of HP-36. Six distinctive clusters are obtained and the most populated is Cluster 2 with 513 populations covering 42.7% of the trajectories (Table I). A centroid structure is also generated for each clusters and the corresponding RMSDback of these six centroids with reference to the NMR-MDavg are described in Table I.

In general, almost all of the six clusters are scattered along the trajectories especially Cluster 2, 3 and 5 as illustrated in

Figure 5(a). Cluster 1, 4 and 6 however are more localized but less important. Further inspection of Cluster 1 reveals that all the members are extracted from the beginning of the folding coordinates and can be regarded as insignificant. At this stage, all of the models are very much non-native with the centroid having the highest RMSD towards the NMR-MDavg (4.45Å). Cluster 4 which occurs among trajectories mostly at 49-58 ns can also be regarded as insignificant, as it comprises of very loosely packed structures (Figure 5(b)) with a low native tertiary contacts (Figure 5(c)). Cluster 6 on the other hand represents a collection of trajectories extracted from the beginning of the simulation (10-25 ns), thus are also very much different from the NMR-MDavg.

TABLE I
COMPARISON OF STRUCTURAL PROPERTIES BETWEEN EACH CONFORMATIONAL CLUSTERS

Cluster	N	R_{gyr}	BS (ns)	§B-C	§C-AVG		§B-AVG	
					All	Core	All	Core
1	52	11.51	40.0	1.85	4.45	1.02	4.84	1.28
2	***513***	***10.00***	***100.2***	***1.72***	***3.40***	***0.94***	***3.53***	***1.01***
3	126	10.84	84.9	1.40	3.97	1.05	4.01	1.18
4	83	10.77	50.8	1.37	3.46	0.97	3.53	1.13
5	236	10.26	72.6	1.35	3.40	0.88	3.61	1.05
6	190	10.21	14.6	1.47	3.95	0.91	4.35	1.01

N= number of structures in the cluster; Rgyr = average measure of compactness (Å); BS = Native-like conformation extracted at simulation time in nanosecond (ns); §B-C = RMSDback between the best model and the centroid structure; §C-AVG = RMSDback between the centroid structure and the NMR-MDavg; §B-AVG = RMSDback between the best structure and the NMR-MDavg; All = All residues from Met1 to Phe3; Core = Residues only in the core region covering Ala9 to Asn28.

Cluster 2 has been identified as the best cluster containing the most near-native like conformations. The definition for the best cluster in this study is the cluster that shows the lowest RMSD between the centroid structure and the NMR-MDavg. However, extra things have been put into considerations such as the number of population within the cluster, the cluster stability as well as the compactness of the conformations within. Although both Cluster 2 and 5 have the same lowest RMSDback-all, Cluster 2 however is found to be more stable with longer residence time and shows higher packing density (10Å).

Furthermore, it is observed that all the states sampled in the last 30 ns of the simulation are grouped in this cluster. A conformation extracted at 10,020 ps (from Cluster 2) has been identified as the 3D model that best represents HP-36 since it gives the lowest RMSDback-all to the centroid. A superimpose of the best model and the NMR-MDavg is shown in Figure 5(b). Figure 5(c) on the other hand, illustrates the differences between the SASA values of the best model and that of the NMR-MDavg. The SASA are classified into nonpolar and polar contributions. The C-terminal part shows strong deviation from that of the NMR-MDavg thus signifying that all the residues are largely exposed to the solvent (positive values) or largely excluded from the solvent (negative values). Phe36 is found to have 186.46 Å2 of its surface exposed to the solvent whereas in the NMR-MDavg is just 88.35 Å2.

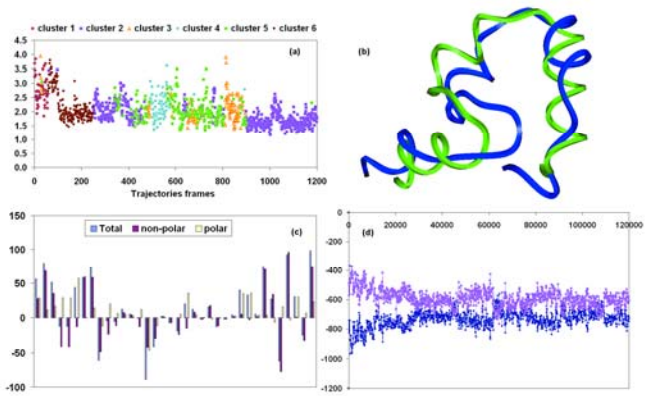

Figure 5. Representations of the cluster, SASA and energetic analyses. (a) A distribution of six clusters with each containing different sets of MD trajectories. Cluster 2 is identified as the set consisting of most compact and stable conformations compared to other sets. (b) Superimpose between the NMR-MDavg (green ribbon) and the best model from cluster 2 (blue ribbon) with backbone RMSD 3.53 Å. (c) Deviation from the NMR-MDavg SASA presented by the total SASA, nonpolar and polar contributions for each residue of the best model. The core region is examined to have little variation. Residues Met13, Ala17, Pro22, Lys25 and Gln26 show native-like SASA values with deviation less than 5 Å2 for both nonpolar and polar contributions. The negative values correspond to the residue in the model having less accessible surface area compared to the native structure and vice versa. (d) Time evolution of the solvation free energy (pink) and the internal electrostatic energy (blue). Time in picoseconds (ps)

The individual terms of the energies are summarized in Table III. The energetic calculation for the NMR-MDavg covers the last 20 ns of the simulation (30-50 ns). Meanwhile, the calculations for the models are not only done on the six clusters, but also on another two trajectory sets with each correspond to two different simulation periods (1-120 ns and 60-120 ns). The reason is to investigate the energy difference between the set containing the initial unstable trajectories (1-120 ns) while the other concerning only the stable part of the conformations (60-120 ns).

As expected, conformations taken from 60 to 120 ns give lower energy compared to the other set (1-120 ns) with a significant energy gap of 4.26 kcal/mol. Thus, we believe that the conformations obtained in the first 60 ns of the simulation are insignificant and can be ignored especially in the process of ranking the native-like models. It is very interesting to observe that the free energy corresponding to Cluster 2 appears to be the lowest of all the six clusters which suggests that this cluster contains most of the lowest energy conformations. This is expected since all the conformations in Cluster 2 appear to have native-like properties as discussed previously and this finding further strengthens our belief that the best model resides in this cluster. Furthermore, it is also shown that the best model exhibits lower conformational free energy compared to that of the NMR-MDavg with energy difference of 5.85 kcal/mol.

The three main terms that differ largely between the

NMR-MDavg and the models are van der Waals, electrostatic and solvation energies. The variations are insignificant for the other energy terms such as bond, angle and dihedral energies. The van der Waals energy reflects the packing of the protein side chains. The higher energy, the less favorable the van der Waals interactions between the atoms. Overall, the model obtained is not able to replicate the van der Waals energy of that of the NMR-MDavg due to the less precise packing of the amino acid side chains in the model. For example, being highly expanded with a large amount of residues exposed to the solvent, it is not surprising that the starting structure (raw model in Table II shows unfavorable van der Waals interactions (1831.36 kcal/mol). However, as the simulation progress, the magnitude of the van der Waals energy is reduced as the structure becomes more compact. It is inspected that Cluster 1 exhibits the highest van der Waals energy follows by Cluster 6. This is expected since almost all of the conformations in these two clusters are loose and less compact.

TABLE II
SUMMARY OF VARIOUS ENERGY TERMS CALCULATED FOR MD TRAJECTORIES OF H-36

	$\langle E_{strain}\rangle$	$\langle E_{vdW}\rangle$	$\langle E_{eel}\rangle$	$\langle E_{gas}\rangle$	$\langle \Delta G_{NP}\rangle$	$\langle \Delta G_{pol}\rangle$	$\langle \Delta G_{sol}\rangle$	$\langle G\rangle$
Native	760.12	-118.20	-650.59	-8.67	17.56	-683.75	-666.15	-674.83
(30-50 ns)	(16.96)	(9.13)	(40.01)	(42.69)	(0.93)	(37.20)	(37.09)	(20.11)
Model	754.68	-90.31	-592.69	71.69	19.41	-752.40	-733.00	-661.31
(1-120 ns)	(16.93)	(11.84)	(56.33)	(62.55)	(0.94)	(54.83)	(54.31)	(19.87)
Model	754.97	-93.51	-611.05	50.42	19.17	-735.16	-715.99	-665.57
(60-120 ns)	(16.90)	(10.21)	(48.84)	(52.90)	(0.85)	(46.68)	(46.39)	(18.89)
Cluster 1	755.67	-67.88	-503.08	184.71	20.04	-846.66	-826.63	-641.92
	(15.81)	(6.20)	(64.33)	(64.50)	(0.43)	(64.71)	(64.49)	(17.06)
Cluster 2	755.83	-96.77	-615.76	43.30	17.08	-729.02	-711.94	-668.63
	(17.25)	(9.27)	(43.37)	(47.47)	(0.50)	(42.75)	(42.56)	(18.40)
Cluster 3	755.85	-86.10	-577.76	92.00	18.59	-770.18	-751.59	-659.59
	(16.10)	(10.49)	(48.71)	(51.65)	(0.49)	(43.12)	(42.95)	(17.99)
Cluster 4	751.07	-86.59	-584.46	80.02	18.06	-756.25	-738.19	-658.17
	(19.45)	(9.99)	(42.27)	(41.85)	(0.46)	(40.65)	(40.56)	(17.77)
Cluster 5	753.62	-89.97	-606.83	56.82	18.07	-738.23	-720.16	-663.34
	(16.48)	(9.61)	(52.64)	(57.04)	(0.65)	(47.35)	(47.08)	(20.08)
Cluster 6	753.44	-83.83	-550.86	118.74	17.89	-793.62	-775.73	-656.99
	(16.05)	(10.20)	(47.82)	(50.16)	(0.65)	(46.39)	(46.09)	(19.63)
Best model	754.70	-107.14	-577.69	69.87	18.38	-768.93	-750.55	-680.68
Raw model	438.21	1831.36	-289.88	1979.7	22.15	-943.88	-921.73	1057.97

Native = NMR-MDavg; Model (1-120 ns) = Conformations taken from 1-120 ns of the simulation; Model (60-120 ns) = Conformations taken from 60-120 ns of the simulation; Best model = the best representative of the HP-36 model; Raw model = the starting model with the core region and the extended terminal segments; $\langle G\rangle = \langle E_{gas}\rangle + \langle \Delta G_{sol}\rangle$; $\langle \Delta G_{sol}\rangle$= solvation energy; $\langle E_{gas}\rangle$= gas phase energy; $\langle E_{strain}\rangle$= strain energy; $\langle E_{vdW}\rangle$= van der Waals energy; $\langle E_{eel}\rangle$= electrostatic energy; $\langle \Delta G_{NP}\rangle$ = solvation non-polar energy; $\langle \Delta G_{pol}\rangle$= solvation polar energy; All energies are in Kcal/mol.

The protein-protein electrostatic energy (E_{eel}) also shows similar trend as the van der Waals term; none of the conformations achieve the NMR-MD_{avg} energy (-650.59 kcal/mol). The $Model_{(1\text{-}120\ ns)}$ and $Model_{(60\text{-}120\ ns)}$ have internal electrostatic energy higher by 57.9 kcal/mol and 39.54 kcal/mol than that of the NMR-MD_{avg}, respectively. The solvent polarization energy (ΔG_{pol}) calculated by the PB equation reveals that the energy of the NMR-MD_{avg} is much higher (-683.75 kcal/mol) compared to the energies of the other conformations. The result demonstrates that the E_{eel} term favors the native compact state while the ΔG_{pol} term prefers the expanded non-native form. In return, this also signifies that the solvation energy, ΔG_{sol} disfavors the native state. Apart from this, it is also seen that loose conformations exhibit lower ΔG_{sol} and higher E_{eel} compared to compact conformations. For example, the energy gaps for both the E_{eel} and ΔG_{sol} between the raw model and the best model (287.81 kcal/mol and 171.18 kcal/mol, respectively) are found to be much larger than the energy gaps between the best model and the conformations from Cluster 1 (74.61 kcal/mol and 76.08 kcal/mol, respectively). Most of the charged atoms in the NMR-MD_{avg} are buried and this incurs large penalties on the ΔG_{sol} since a large contribution of the solvation term comes from the electrostatic interaction between protein and water. The protein however cleverly eliminates the penalties for burying the charged atoms by forming a smooth equilibrium between the ΔG_{sol} and the E_{eel} terms. These two terms are found to be inversely correlated with an excellent correlation coefficient of 0.98. The burial of these charged atoms thus result in more favorable energy for E_{eel} due to a better atomic charge distribution. This finding is also in agreement with previous studies [21, 26]. As the ΔG_{sol} decreases (more negative), the E_{eel} will compensate by increasing the magnitude (less negative) and vice versa (Figure 5(d)). This cooperative balance between the E_{eel} and the ΔG_{sol} answers to the question of why the model has lower ΔG_{sol} compared to the NMR-MD_{avg}.

IV. Conclusion

This study reports on the combined approach of the knowledge-based method and MD simulation as an alternative protocol for predicting the 3D structures of proteins. It is shown that this procedure is effective in predicting the 3D structure of small globular protein, HP-36 with a considerable accuracy in much shorter time compared to the conventional MD simulation alone. From this work, we concluded that the presence of the well-developed knowledge-based core region can serve as a 'nucleation center' for subsequent rapid folding. We also propose that our combined method is capable in preventing the formation of non-native contact that will hinder the folding rate. However, further works are critical as to enhance this protocol and perhaps benchmark according to the accuracy of the knowledge-based core region and the size of the proteins.

Acknowledgment

This work is supported by the Top Down Grant No 09-02-04-001 BTK/TD/004 awarded by the National Biotechnology Directorate, Ministry of Science, Technology and Innovation, Malaysia. The authors wish to acknowledge MIMOS (M) Berhad for providing the computing time.

References

[1] Sanchez R, Sali A: Advances in comparative protein-structure modeling. Curr Op Struc Biol 1997, 7:206-214.
[2] Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A: Critical assessment of methods of protein structure prediction (CASP) - Round 6. Proteins 2005, 61:3-7.
[3] Zagrovic B, Snow CD, Shirts MR, Pande VS: Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. J Mol Biol 2002, 323:927-937.

[4] Duan Y, Kollman PA: Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998, 282:740-744.
[5] Jang S, Kim E, Shin S, Pak Y: Ab initio folding of helix bundle proteins using molecular dynamics simulations. J Am Chem Soc 2003, 125:14841-14846.
[6] Krautler V, Aemissegger A, Hunenberger PH, Hilvert D, Hansson T, van Gunsteren WF: Use of molecular dynamics in the design and structure determination of a photoinducible b-hairpin. J Am Chem Soc 2004, 127:4935-4942.
[7] Kubelka J, Eaton WA, Hofrichter J: Experimental tests of villin subdomain folding simulations. J Mol Biol 2003, 329:625-630.
[8] Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE: Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. Acta Cryst D Biol Cryst 1998, 54:1078-1084.
[9] Jorgensen WL, Chandrasekar A, Madura JD, Impey RW, Klein ML: Comparison of simple potential functions for simulating liquid water. J Chem Phys 1983, 79:926-935.
[10] Berendsen HJC, Postma JPM, van Gunsteren WF, Dinola A, Haak JR: Molecular dynamics with coupling to an external bath. J Comp Phys 1984, 81:3684-3690.
[11] Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham III TE, DeBolt S, Ferguson N, Seibel G, Kollman P: AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics, and free energy calculations to simulate the structural and energetic properties of molecules. Comp Phys Comm 1995, 91:1-41.
[12] Duan Y, Wu C, Chowdhury S, Lee M.C, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, et al: A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comp Chem 2003, 24:1999-2012.
[13] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215:403- 410.
[14] Thompson JD, Higgins DG, Gibson TJ: CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matric choice. Nuc Acids Res 1994, 22:4673-4680.
[15] Sali A, Blundell TL: Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993, 234:779-815.
[16] Hubbard SJ, Thornton JM: "NACCESS", Computer program. Department of Biochemistry and Molecular Biology, University College London; 1993.
[17] Feig M, Karanicolas J, Brooks III CL: MMTSB Tool set: Enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Biol 2004, 55:379-400.
[18] Kollman P, Massova I, Reyes CM, Kuhn B, Huo S, Chong LT, Lee MR, Lee T, Duan Y, Wang W, et al: Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. Acc Chem Res 2000, 33:889-897.
[19] Srinivasan J, Cheatham III TE, Cieplak P, Kollman P, Case DA: Continuum solvent studies of the stability of DNA, RNA and phosphoramidate-DNA helices. J Am Chem Soc 1998, 120:9401-9409.
[20] Vorobjev YN, Hermans J: ES/IS: Estimation of conformational free energy by combining dynamics simulations with explicit solvent with an implicit solvent continuum model. Biophys Chem 1999, 78:195-205.
[21] Lee MR, Duan Y, Kollman PA: Use of MM-PB/SA in estimating the free energies of proteins: Application to native, intermediates, and unfolded villin headpiece. Proteins 2000, 39:309-316.
[22] Alonso DOV, Daggett V: Molecular dynamics simulations of hydrophobic collapse of ubiquitin. Protein Sci 1998, 7:860-874.
[23] Creighton TE: PROTEINS: Structures and molecular properties. 2nd edn. New York: W.H. Freeman and Company; 1993.
[24] Yang A, Honig B: Free energy determinants of secondary structure formation: I. a-helices. J Mol Biol 1995, 252:351-365.
[25] Shirts MR, Pande VS: Screensavers of the world, unite! Science 2001, 290:1903-1904.
[26] Vorobjev YN, Hermans J: Free energies of protein decoys provide insight into determinants of protein stability. Protein Sci 2001, 10:2498-2506.