Journal of Chemical and Natural Resources Engineering,Special Edition:141-154
© FKKKSA, Universiti Teknologi Malaysia

STRUCTURE PREDICTION OF LARGE PROTEIN USING THE COMBINATION OF KNOWLEDGE-BASED AND PHYSICS-BASED APPROACHES: METHOD VALIDATION ON CHOLESTEROL ESTERASE

N.B. AHMAD KHAIRUDIN[1], H. A. WAHAB[2], M.R. SAMIAN[3], N. NAJIMUDIN[3]

ABSTRACT

The objective of this study was to predict the structure of a large protein using a combined approach of knowledge-based comparative modeling and physics-based Molecular Dynamics (MD) simulation applied to the enzyme cholesterol esterase. The core region of the enzyme was modelled using information from homologous known protein structures whereby leaving the end-terminal regions (the nonhomologous regions) to fold via MD simulation. Currently, there is yet a reported study where one begins with a knowledge-based model of the core region of a protein and allowing the remaining end terminal regions to fold via MD simulation. The method was categorized into three parts; the development of the core region of the protein, the development of the complete protein structure and the MD refinement simulation. Three models were tested, CE_{CRL-87}, CE_{THG-45} and CE_{JKM-14}, with each originating from different core regions developed at three different cutoff values of sequence identity; more than 70% (%id > 70%), less than 60% but more than 30% (30% < %id < 60%) and less than 20% (%id < 20%), respectively. The remaining residues were later added using MD simulation which then followed by 20 ns of MD refinement. It was shown that the use of different starting core regions did not significantly contribute towards correct structure predictions of large proteins. Furthermore, the use of restraint of the core region would only deteriorate the model as observed in CE_{THG-45}.

Key Words : Protein structure prediction, Homology modeling, Molecular dynamics, cholesterol esterase, Fold recognition

1.0 INTRODUCTION

Despite the fact that predicting the structures of proteins using computational approach remains one of the longest standing challenges in structural biology [1-3], the field has somehow made impressive strides forward [4-7]. It has been revolutionized by the blend of knowledge-based method and physics-based *de novo* folding simulation. Over the years, these two methods have become more and more integrated. Accurate predictions of the knowledge-based method rely heavily on the sequence identities or similarities between the target and the template proteins [8]. The problem commences when the target protein does not share any significant similarities with any solved protein structures.

Molecular dynamics (MD) simulation on the other hand has been widely applied to fold proteins without having to rely on sequence conservation [9-13]. Applying proper force field and sufficient computing resources, all-atom MD simulation is

[1]Department of Bioprocess Engineering, Faculty of Chemical and Natural Resources Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Bahru, Malaysia.
[2]Malaysia Institute of Pharmaceuticals and Nutraceuticals, Universiti Sains Malaysia, 11800 Minden Penang, Malaysia.
[3]School of Biological Sciences, Universiti Sains Malaysia, 11800 Minden Penang, Malaysia
Correspondence to : Nurul Bahiyah Ahmad Khairudin (nurul@fkkksa.utm.my)

capable of folding small proteins or peptides into their functional states using only the information of the linear chain of the amino acid sequence. This method however requires a substantially high computational power to cope with the huge number of degrees of freedom in proteins [14]. As a result, MD folding simulation is currently restricted to very small proteins and peptides with the time regime limited to hundreds of nanoseconds or a few microseconds which is the upper bound of current simulation time [13, 15]. Among the many successful ones include the folding simulations of helical proteins [16-18] and β-hairpins [19-23]. While current computer speed can successfully fold small proteins and peptides into their native state, large proteins (more than 300 residues) on the lower end are trailing behind with daunting challenges. To date, there is no reported work on full atomic folding studies of large proteins. The reason is that folding simulation involving large proteins using the current all-atom folding simulation is regarded to be impractical.

This study seeks to embark upon structure prediction of a large protein using a combined approach of knowledge-based comparative modeling and physics-based MD simulation. The intention of this study is neither to solve the protein folding problem nor to investigate the folding pathways of proteins. The main aim of this work is to validate the proposed prediction method towards the accurate prediction of large proteins with more than 500 residues, which is the range of lengths of many important proteins. The general idea is to model the core region of the protein using information from homologous protein structures leaving the end-terminal regions or the nonhomologous regions to fold via MD simulation. There have been many reported studies that exploited the knowledge-based and physics-based methods such the Rosetta method which has shown to be quite effective [24]. There are also others which have contributed in better understanding of the *de novo* protein structure prediction [25, 26]. There are also abundant studies which employed MD as a refinement tool to verify the stability of the complete 3D models built using homology modeling method [27-33]. Nevertheless, there is yet a reported study where one begins with a knowledge-based model of the core region of a protein and letting the remaining end terminal regions to fold via MD simulation. It is hoped that this study could contribute to the understanding of protein structure prediction generally and to the structure prediction of large proteins specifically especially for proteins that cannot be solved using the conventional knowledge-based method alone.

2.0 MATERIALS AND METHODS

Cholesterol esterase

The enzyme cholesterol esterase (CE) was chosen as the system to be studied. Figure 1 shows the 3D structure of CE obtained from the asporogenic yeast *Candida cylindracea* solved at the resolution of 2.0 Å using X-ray Crystallography (PDB id: 1CLE) [34]. Containing 534 amino acid residues, this enzyme reversibly hydrolyzes cholesterol and other cholesterol-containing compounds into sterol and fatty acids. Its tertiary structure consists of the α/β domain with a combination of 13 β-strands and 16 α-helices with the core region comprises of seven β-strands forming parallel β sheet.

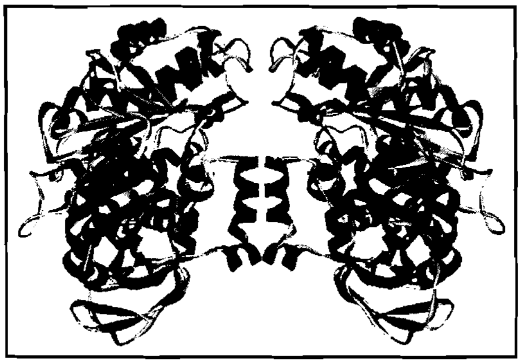

Figure 1 Ribbon representation of the 3D tertiary structure of cholesterol esterase. Red = helices, blue = β-strands, grey=random coils, green= β-turns.

Development of the Core Region of CE

The core regions of CE were developed at three different cutoff values of sequence identity (id); more than 70% (%id > 70%), less than 60% but more than 30% (30% < %id < 60%) and less than 20% (%id < 20%). The rationale behind the variety of the cutoffs was to investigate how the different identity values of the core region would contribute to the extent of the success of this proposed method. Thus, this study involved three independent study cases with different starting structures.

The linear chain of CE was subjected to sequence analysis using BLAST [35] to locate for appropriate templates. The search for the appropriate template was also performed using the fold recognition methods of mGenThreader [36], 3DPSSM [37] and FUGUE [38]. Results showed that CE shared a wide range of similarities among solved structures. The possible templates were ranked according to the highest percentage of sequence identity.

Table 1 Summary of the BLAST results with scores and % global identity.

Possible Templates (Lipase)	Score	E-value	Global Identities (%)
1LPS, 1LPP, 1LPO, 1LPN,1LPM, 1TRH, **1CRL**	974	0.0	**88**
1GZ7	908	0.0	82
1THG	411	1e-115	**41**

From Table 1, it was observed that CE shared high sequence identity (88%) with lipases 1LPS, 1LPP, 1LPO, 1LPN, 1LPM, 1TRH and 1CRL. Out of these lipases, 1CRL [39] was chosen as the template to build the comparative model for the first case study (% id > 70%). The lipase 1THG [40] shared more than 30% but less than 50% identity with CE. Therefore, this structure was chosen as the template to build the core region of CE for the second case study. Table 2 lists the possible templates obtained from the fold recognition methods. Protein 1JKM [41] was found to be the most suitable template for the third case study, with % id cutoff < 20%. Not only was it in the range of the cutoff threshold, it was also the only protein that appeared in all the results obtained from the three methods. Among other criterion considered in the template selection was the quality of the template structure. For X-ray solved structure, the resolution must be as low as possible and that there should be no missing residues. Table 3 summarizes the selected templates for the modeling of the core regions of CE. The sequence alignments between the templates and the proteins were performed using CLUSTALX [42] and the models were developed using Modeller7v7 [43].

Table 2 Proposed templates obtained from the three threading methods along with the % id and the calculated scores.

Methods	% identity	Score
3D-PSSM		***E-value***
• 1qid (acetylcholinesterase)	27	4.8e-06
• 1ehq (butyryl cholinesterase)	27	0.0027
• 2bce (cholesterol esterase)	25	0.0027
• 1qe3 (PNB esterase)	27	0.0027
• **1jkm (brefeldin A esterase)**	**13**	0.0114
• 1gkl (Feruloyl esterase)	10	0.0485
FUGUE		***Z-score***
• 1jji (Carboxylesterase)	14.52	16.42
• 1vkh (serine hydrolase)	9.52	13.09
• **1jkm (brefeldin A esterase)**	**13.60**	11.84
• 1h2w (oligopeptidase)	13.57	11.16
• 1jfr (lipase)	10.63	4.15
MGenThreader		**Z-score**
• 1c7i (PNB esterase)	27.1	3e-05
• 2bce (cholesterol esterase)	23.5	3e-05
• 1ukc (esterase)	25.9	4e-05
• 1evq (carboxylesterase)	17.2	5e-05
• **1jkm (brefeldin A esterase)**	**13.7**	6e-05
• 1mpx (hydrolase)	8.20	9e-05

Table 3 The chosen templates for the 3D core region structure prediction of CE at various cutoff of sequence identity

Study Cases	**Templates**	**Sequence identity (% id)**
% id cutoff > 70%	1CRL	87.0
30 < % id cutoff < 60%	1THG	45.0
% id cutoff < 20%	1JKM	13.6

Development of the Complete 3D Mode

For the development of the complete structure of the protein, the remaining amino acids that were not modeled were added to the core region, 10 residues at a time, consecutively. This was followed by energy minimization and short MD simulation in order to quickly fold the extended segments towards the core region. The whole process was repeated by adding another group of 10 amino acids until the complete structure was formed. The MD simulations were performed using AMBER8 [44] suite of programs. The force field amber.*ff03* [45] was used in all the simulations. Figure 2 summarized the general procedure applied for the three study cases.

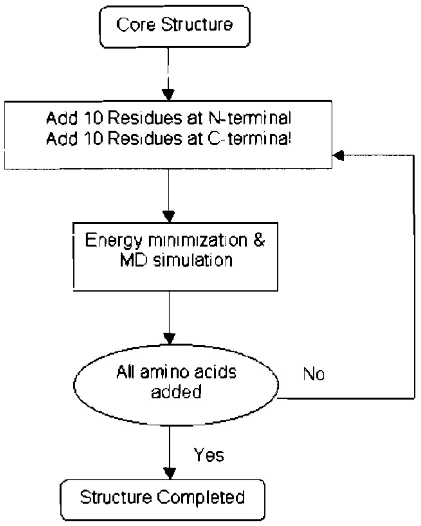
Core Structure
Add 10 Residues at N-terminal
Add 10 Residues at C-terminal
Energy minimization & MD simulation
All amino acids added
No
Yes
Structure Completed

Figure 2 Protocol for the development of the 3D complete protein

The energy minimization was carried out using 500 cycles of steepest descent and another 500 cycles of conjugate gradient. The MD simulation was carried out for 1 ns. In order to save computing time, the system was simulated in vacuum environment with a crude estimation of the solvent using distance dependent dielectric constant. Non-bonded interactions were truncated by using a 16 Å cutoff. The system was coupled to a temperature bath using Berendsen [46] thermostat to maintain the temperature at 300K with coupling constants of 1.0 ps. Bond constraints were imposed on all bonds involving hydrogen atoms via SHAKE [47]. The trajectories were produced by numerical integration of the Newton's equation of motion using the Verlet-leap frog [48] algorithm with a time step of 2 fs. The protein models (CE_{JKM-14} and CE_{CRL-87}) were simulated without any restraint except for the second case (CE_{THG-45}) in which restraint was imposed on the core region.

Long MD Simulation

The final part of the method was the long MD refinement in the presence of explicit water molecules. The model was first subjected to energy minimization using 500 cycles of steepest descent and another 500 cycles of conjugate gradient. The simulation was performed in a periodic boundary condition with the molecule immersed in a truncated octahedron water box filled with TIP3P water model. Table 3.4 showed the initial properties of the system under investigation. The nonbonded interactions were treated using 16Å cutoff and PME for Lennard Jones and coulombic interactions, respectively. The system was simulated at constant pressure (1 bar) and constant temperature (300 K) using Berendsen weak-coupling thermostat with coupling constants of 1 ps. Bond constraints were imposed on all bonds involving hydrogen atoms via SHAKE [47]. The production phase was started from an equilibration phase of 1 ns at 300K and 1 bar of pressure. The simulation was conducted for 20 ns. As a control, the crystal structure of native CE was also subjected to MD simulation for 10 ns using the same conditions described above.

3.0 RESULTS

Figure 3 illustrates the development of the $RMSD_{back}$ as a function of the simulation time for the three models, CE_{JKM-14}, CE_{THG-45} and CE_{CRL-87}. As observed from the trendlines, all of the models had very large RMSD with more than 20Å. This suggested that the models were largely different from the native conformation. This finding was supported by the small fraction of the tertiary native contacts as presented in Table 4. The very small percentages (less than 18%) portrayed an almost zero occurrence of tertiary native contacts, whereas there were around 84.5% contacts in the CE_{MD-avg}.

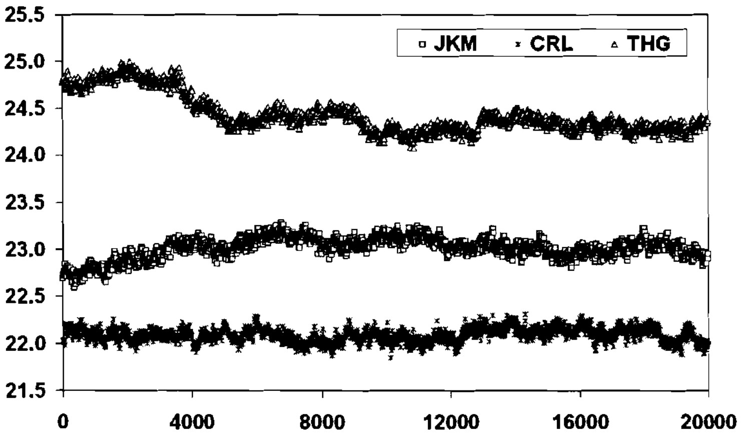

Figure 3 Time evolution (ps) of $RMSD_{back}$ (Å) for CE_{CRL-87}, CE_{THG-45} and CE_{JKM-14}

Table 4 Average values of the structural properties of the predicted 3D structures during the last 10ns of MD refinement

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
$RMSD_{all}$	-	22.09	24.30	23.02
$RMSD_{core}$	-	6.56	5.25	17.79
Gyration	22.51	22.78	23.94	24.15
Total SASA	19,516.94	22,258.72	24,056.09	22,534.03
Nonpolar SASA	11,807.53	14,746.22	15,525.23	14,657.38
Polar SASA	7,709.41	7,512.50	8,530.85	7,876.65
Fraction	0.85	0.14	0.15	0.11

The $RMSD_{back}$ for CE_{THG-45} and CE_{JKM-14} increased in the initial part of the MD refinement. This rise was due to the release of strain in the structure which in turn led the models to adopt other conformational geometries. After 2 ns, the $RMSD_{back}$ of CE_{THG-45} started to slowly decrease until it reached 5 ns and began to stabilize for the remaining sampling period with values varying between 24.2 to 24.5 Å from that of CE_{MD-avg}. In contrast, the $RMSD_{back}$ for CE_{JKM-14} continued to increase modestly from 22.6 Å in the beginning up to 23 Å until it reached 8 ns where the value started to remain plateau. As for CE_{CRL-87}, it was observed that the structure did not undergo any significant structural change. This was demonstrated by the stable trend of the $RMSD_{back}$ throughout the simulation. As expected, the $RMSD_{back}$ for CE_{CRL-87} was found to be lower by 1Å and 2.5Å from that of CE_{JKM-14} and CE_{THG-45}, respectively.

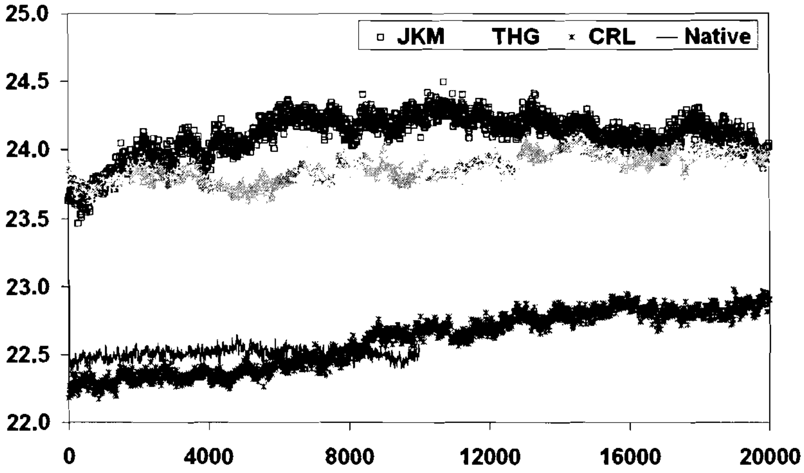

Figure 4 Time evolution (ps) of the R_{gyr} (Å) for CE_{CRL-87}, CE_{THG-45}, CE_{JKM-14} and the native

As demonstrated in Figure 4, both CE_{THG-45} and CE_{JKM-14} displayed larger R_{gyr} compared to that of CE_{CRL-87} and the native. There was a slight increment in the values during the first 4 ns of the simulation and it continued to rise gradually until it reached 16 ns where the value started to stabilize around 23.8 to 24.2 Å. This showed that both CE_{THG-45} and CE_{JKM-14} were highly expanded compared to the native with averaged value of 22.5 Å. This indicated that the added residues in both models did not pack well with residues in the core region. On the other hand, R_{gyr} for CE_{CRL-87} almost resembled that of the native with compactness of around 22.3 Å. However, as the simulation progressed, the value slowly raised up to 23 Å at the end of the simulation.

The total SASA for the models and the native were calculated and presented in Figure 5. The SASA of the native remained stable throughout the simulation with an averaged exposed area of 19516 Å^2. There seemed to be a drastic steady increase in SASA for CE_{JKM-14} from 19980 Å^2 to 22559 Å^2 in the first 4 ns. As for CE_{THG-45}, the increment was quite modest but showed higher SASA value of 24,056 Å^2. In contrast to CE_{THG-45} and CE_{JKM-14}, the SASA for CE_{CRL-87} started to rise only after 4 ns and continued to increase up to the end of the refinement to achieve the same SASA value of that CE_{JKM-14}. The elevated SASA values were mostly contributed by the large exposed surface area involving the nonpolar amino acids as depicted in Figure 6. The nonpolar SASA increased over time except for CE_{THG-45} in which the SASA remained fluctuated in the range of 15,000 to 16,000 Å^2. At the end of the simulation, the three models shared the same nonpolar SASA with values varying between 14,000 to 15,000 Å^2.

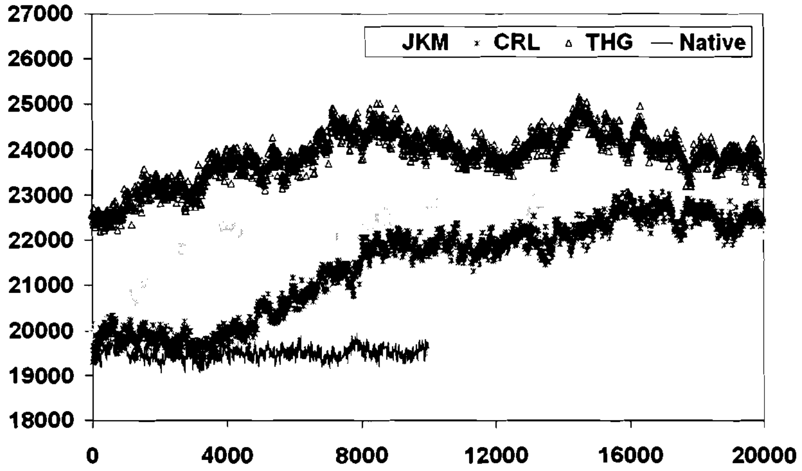

Figure 5 Time evolution (ps) of Total SASA (Å^2) for CE$_{CRL-87}$, CE$_{THG-45}$, CE$_{JKM-14}$ and the native

Apart from the nonpolar SASA, the evolution of the polar component of the SASA was also investigated (Figure 7). There was a rapid increase of the polar SASA for CE$_{JKM-14}$ while a more stable increase for both CE$_{THG-45}$ and CE$_{CRL-87}$. In the beginning of the simulation, all models had lower polar SASA compared to that of the native indicating that most of the polar residues in the structure models were not located on the protein surface. However, in the last 4 ns, the polar SASA for both CE$_{CRL-87}$ and CE$_{JKM-14}$ stabilized and fluctuated around 7700 Å^2 resembling the averaged polar SASA of the native. As for CE$_{THG-45}$, the polar SASA remained stable at much higher value of 8500 Å^2. Even though the high deviation of the total SASA from that of the native was caused by the large amount of the nonpolar SASA, the continuous rise in the total SASA was however largely contributed by the substantial rise of SASA involving polar residues. About 66.4%, 71.3% and 45.3% of the increment in the total SASA was due to the exposure of the polar surface area for CE$_{THG-45}$, CE$_{JKM-14}$ and CE$_{CRL-87}$ respectively.

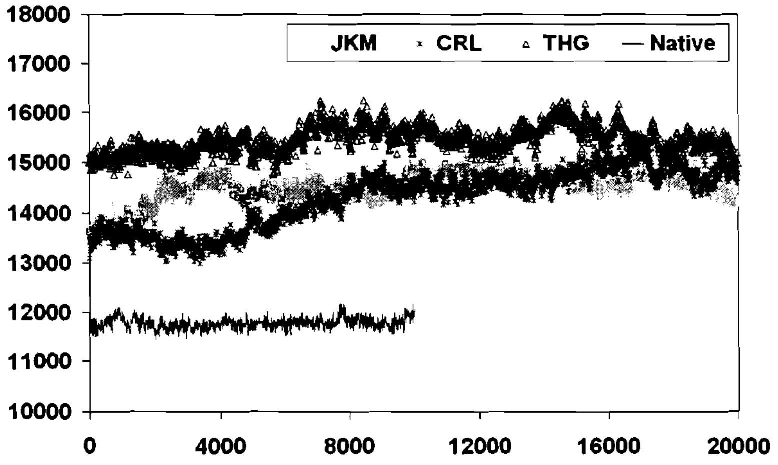

Figure 6 Time evolution (ps) of Nonpolar SASA (Å^2) for CE$_{CRL-87}$, CE$_{THG-45}$, CE$_{JKM-14}$ and the native

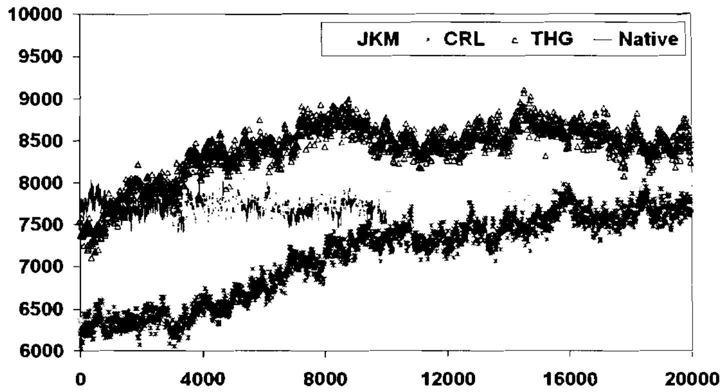

Figure 7 Time evolution (ps) of Polar SASA (Å^2) for CE$_{CRL-87}$, CE$_{THG-45}$, CE$_{JKM-14}$ and the native

Another interesting feature worth mentioning was the changes observed in the secondary structures throughout the simulation. In general, there were very low occurrences of secondary structures in the models. The models were mainly composed of random coils and loops which occupied more than 450 residues in average instead of less than 300 residues as what was observed in the native. Table 5 summarized the percentages as well as the total number of residues involved in the formation of the native α-helices and β-strands for the native crystal and for the CE models taken at 20 ns of the MD simulation. A feature worth highlighting was the increase in the percentage of the native helices for CE$_{CRL-87}$ from 6.74% to 12.17% following the 20ns refinement. However, these helices usually were either shorter or longer by a few residues as compared to the helices in the native. This increment was not observed to occur in other models. Apart from the α-helices, the formation of the native strands was also observed to increase by 0.38% and 1.5% in the models CE$_{CRL-87}$ and CE$_{THG-45}$, respectively. In the native structure, 31.65% and 12.36% of the 534 amino acids formed the α-helices and the β-strands, respectively while the remaining consisted of loops. In contrast, there were only 65 and 57 residues forming native α-helices in CE$_{CRL-87}$ and CE$_{THG-45}$, respectively, more than twofold lower from that of the native. On the lower end, the model CE$_{JKM-14}$ was almost entirely made up of random loops.

Table 5 Total number of residues involved in the formation of the secondary structures for the native crystal, CE$_{MD-avg}$ and the CE models.

Sec.	Native	CE$_{CRL-87}$		CE$_{THG-45}$		CE$_{JKM-14}$	
elements		Raw	MD	Raw	MD	Raw	MD
α-helices	169 (31.65)	36 (6.74)	65 (12.17)	61 (11.42)	57 (10.67)	23 (4.31)	13 (2.43)
β-strands	66 (12.36)	6 (1.12)	8 (1.50)	13 (2.43)	21 (3.93)	25 (4.68)	12 (2.25)
loops	299 (55.99)	492 (92.13)	461 (86.33)	460 (86.14)	456 (85.39)	486 (91.01)	509 (95.32)

Sec. elements = secondary structural elements
Avg MD = average structure of the native after 10 ns MD simulation
Raw = model before refinement
MD = model after 20 ns refinement
All numbers in the parenthesis correspond to percentages (%).

4.0 DISCUSSION

The central objective of this study was to validate the proposed combined knowledge-based and physics-based structure prediction method when applied to large protein. Three models were tested with each originating from different starting core regions developed using the knowledge-based method. The remaining residues were later added and folded via MD simulation followed by 20 ns of MD refinement. Initially, it was hoped that the core region could provide a scaffold for the remaining residues to fold correctly. However, throughout the 20 ns MD refinement, the models did not closely resemble the native conformation. The high $\mathrm{RMSD}_{\mathrm{back}}$, the large R_{gyr} and the significant increase in the total SASA indicated that the models were experiencing large conformational changes involving expansion of the protein size and errors in packing within the protein interior. Result also showed that there was no bias between the three models which suggested that the core region, regardless of its accuracy, was not significantly exploited to correctly predict the structure of the protein. This was due to the huge number of atoms involved and 20 ns was insufficient to correctly fold the added residues.

It is believed that hydrophobic interactions are crucial in forming and stabilizing the protein structures [49]. The aggregation of the nonpolar side chains in the interior part of the protein to avoid contact with surrounding water molecules is indeed the basis of the hydrophobic effects. In this study, this effect was investigated by SASA and R_{gyr}. Since the models were less compact, with almost zero of native tertiary contacts, the chances of the occurrences of the secondary structures such as α-helices and β-strands were slim. This suggested that the network of protein-protein hydrogen bonds was very poor which forced the buried polar residues to traverse to the protein surface to seek interactions with water molecules. This behavior led to the opening up of the structures as if they were unfolding as shown by the high SASA values. Since it was agreed that SASA for denatured state is larger than that of the native state [50-52], it could be pointed that the models obtained in the current work were denatured in which all the native attributes were nowhere closed.

In general, if a model was poorly developed, the structure would likely to unfold into a different conformation following MD refinement. Thus, structural restraint on the core region was imposed for $\mathrm{CE}_{\mathrm{THG\text{-}45}}$, in order to investigate whether the restrained core structure could provide a better channel for the protein to fold into its native form. Surprisingly, not only was the RMSD for $\mathrm{CE}_{\mathrm{THG\text{-}45}}$ higher compared to that of $\mathrm{CE}_{\mathrm{JKM\text{-}14}}$ which had the core region developed less accurately, the result also indicated that the polar SASA for $\mathrm{CE}_{\mathrm{THG\text{-}45}}$ was greatly diverged from the native polar SASA. This finding suggested that the use of restraint on the core region caused the protein to be trapped in a high energy conformation.

It is widely agreed that all-atom folding study of large proteins is highly impractical and could not be accomplished even in the foreseeable future. It is unrealistic to expect large protein to fold into its native structure within the current limit of simulation time. In view of this difficulty, this study aimed to contribute a new perspective to interpret the problem only to find that the massive amount of degrees of freedom present in large proteins still could not be handled even in the presence of an accurate core region. Thus, the method proposed in this study cannot be applied to predict the 3D structure of a large protein. It might work should the MD simulation were extended to the order of milliseconds to minutes. However, this is impossible for even the most powerful parallel machines today could only simulate all-atom folding in the limits of microseconds using only small proteins typically less than 100 amino acids [53, 54].The current all-atom folding simulation studies were restricted by the efficiency of the MD simulation itself to effectively sample the conformational space of the system. Efficient conformational sampling over the entire phase space is vital to obtain an accurate thermodynamic ensemble at certain temperature which obeys the ergodic hypothesis. However, for large structures such as proteins, the sampling process is not trivial due to the rugged energy landscape that contains numerous local energy minimas separated by high energy barriers.

5.0 CONCLUSIONS

The attempt to predict the 3D structure of a large protein by exploiting both the knowledge-based and the physics-based methods did not succeed. Although both methods were commonly applied in the field of protein structure prediction, the challenge however was far from trivial when it involved large proteins. Three models CE_{CRL-87}, CE_{THG-45} and CE_{JKM-14} were developed using different starting core regions. Each region represented a different level of accuracy spanning from the most accurate (CE_{CRL-87}) to the least accurate (CE_{JKM-14}) models. MD simulation was further applied to add the remaining residues onto the core regions followed by a 20 ns MD refinement simulation. However, the predicted structures were found to be denatured and had lost the native attributes. Nevertheless, this study could be regarded as important in the sense that it provided the understanding of using different starting models of different accuracies towards predicting the 3D structure of a large protein using the proposed structure prediction technique. It was shown that the use of different starting core regions did not significantly contribute towards correct predictions even when the region showed very high sequence identity (CE_{CRL-87}). Furthermore, it was also shown that the use of restraint in this combined method deteriorated the structure, as observed in the model CE_{THG-45}. The task to predict the functional form of a large protein via conventional MD simulation is not a reliable method even with the inclusion of knowledge-based information. This is due to the massive conformational space that need to be sampled and current computational power still could not cope with such exhaustive conformational space.

REFERENCES

[1] Aloy P., M. Pichaud and R.B. Russell. 2005. Protein complexes: Structure prediction challenges for the 21st century, *Curr. Op. Struc. Biol.* 15: 15-22.

[2] Skolnick J. and A. Kolinski. 1989. Computer simulations of globular protein folding and tertiary structure, *Annu. Rev. Phys. Chem.* 40: 207-235.

[3] Westhead D.R. and J.M. Thornton 1998. Protein structure prediction, *Curr. Op. Biotechnol.* 9: 383-389.

[4] Honig B. 1999. Protein folding: From the levinthal paradox to structure prediction, *J. Mol. Biol.* 293: 283-293.

[5] Baker D. 2006. Prediction and design of macromolecular structures and interactions, *Phil. Trans. R. Soc. B.* 361: 459-463.

[6] Schueler-Furman O., C. Wang, P. Bradley, K. Misura and D. Baker. 2005. Progress in Modeling of protein structures and interactions, *Science.* 310: 638-642.

[7] Daggett V. 2006. Protein folding - simulation, *Chem. Rev.* 106: 1898-1916.

[8] Rost B. in (von Rague-Schleyer, P., Allinger, N.L., Cclark, T., Gasteiger, J., Kollman, P.A. and Schaefer, H.F., eds.) Encyclopedia of Computational Chemistry, John Wiley, Chisester 1998, pp. 2242-2255.

[9] Karplus M. and A. Sali. 1995. Theoretical studies of protein folding and unfolding, *Curr. Op. Struc. Biol.* 5: 58-73.

[10] Daura X. 2005. Molecular dynamics simulation of peptide folding, *Theor. Chem. Acc.*

[11] Daggett V. 2002. Molecular dynamics simulations of the protein unfolding/folding reaction, *Acc. Chem. Res.* 35: 422-429.

[12] Lee M.R., Y. Duan and P.A. Kollman. 2001. State of the art in studying protein folding and protein structure prediction using molecular dynamics methods, *J. Mol. Graph. Model.* 19: 146-9.

[13] Seibert M.M., A. Patriksson, B. Hess and D. van der Spoel. 2005. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations, *J. Mol. Biol.* 354: 173-183.

[14] Fersht A.R. and V. Daggett. 2002. Protein folding and unfolding at atomic resolution, *Cell.* 108: 573-582.

[15] Gnanakaran S., H. Nymeyer, J. Portman, K.Y. Sanbonmatsu and A.E. Garcia. 2003. Peptide Folding Simulations, *Curr. Op. Struc. Biol.* 13: 168-174.

[16] Zagrovic B., C.D. Snow, M.R. Shirts and V.S. Pande. 2002. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing, *J. Mol. Biol.* 323: 927-937.

[17] Duan Y. and P.A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution, *Science.* 282: 740-744.

[18] Jang S., E. Kim, S. Shin and Y. Pak. 2003. *Ab initio* folding of helix bundle proteins using molecular dynamics simulations, *J. Am. Chem. Soc.* 125: 14841-14846.

[19] Lee J., S. Jang, Y. Pak and S. Shin. 2003. Folding dynamics of β-hairpins: Molecular dynamics simulations, *Bull. Korean Chem. Soc.* 24: 785-791.

[20] Zagrovic B., E.J. Sorin and V.S. Pande 2001. β-hairpin folding simulations in atomistic detail using an implicit solvent model, *J. Mol. Biol.* 313: 151-169.

[21] Jang S., S. Shin and Y. Pak. 2002. Molecular dynamics study of peptides in implicit water: *Ab initio* folding of β–Hairpin, β–sheet, and ββα-motif, *J. Am. Chem. Soc.* 124:,4976-4977.

[22] Krautler V., A. Aemissegger, P.H. Hunenberger, D. Hilvert, T. Hansson and W.F. van Gunsteren. 2004. Use of molecular dynamics in the design and structure determination of a photoinducible β-hairpin, *J. Am. Chem. Soc.* 127: 4935-4942.

[23] Paci E., A. Cavalli, M. Vendruscolo and A. Caflisch. 2003. Analysis of the distributed computing approach applied to the folding of a small β peptide, *Proc. Natl. Acad. Sci. USA.* 100: 8217-8222.

[24] Rohl C.A., C.E.M. Strauss, K.M.S. Misura and D. Baker. 2004. Protein structure prediction using Rosetta, *Methods Enzymol.* 383: 66-93.

[25] Karplus K., R. karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans and R. Hughey. 2003. Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction., *Proteins: Structure,Function and Bioinformatics.* 53: 491-496.

[26] Skolnick J., A. Kolinski, D. Kihara, M. Betancourt, P. Rotkiewicz and M. Boniecki. 2001. *Ab initio* protein structure prediction via a combination of threading, lattice folding, clustering and structure refinement, *Proteins.* 5: 149-156.

[27] Lee M.R., J. Tsai, D. Baker and P.A. Kollman. 2001. Molecular dynamics in the endgame of protein structure prediction, *J. Mol. Biol.* 313: 417-30.

[28] Simmerling C., M.R. Lee, A.R. Ortiz, A. Kolinski, J. Skolnick and P. Kollman. 2000. Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: Application to the small protein CMTI-1, *J. Am. Chem. Soc.* 122: 8392-8402.

[29] Fan H. and A.E. Mark. 2004. Refinement of homology-based protein structures by molecular dynamics simulation techniques, *Prot. Sci.* 13: 211-220.

[30] Cazalis R., T. Aussenac, L. Rhazi, A. Marin and J. Gibrat. 2003. Homology modeling and molecular dynamics simulations of the N-terminal domain of wheat high molecular weight glutenin subunit 10, *Prot. Sci.* 12: 34-43.

[31] Ma B., J. Xiong, J. Lubkowski and R. Nussinov. 2000. Homology modeling and molecular dynamics simulations of lymphotactin, *Prot. Sci.* 9: 2192-2199.

[32] Li L., T. Darden, C. Foley, R. Hiskey and L. Pedersen. 1995. Homology modeling and molecular dynamics simulation of human prothrombin fragment 1, *Prot. Sci.* 4: 2341-2348.

[33] Pisabarro M.T., A.R. Ortiz, L. Serrano and R.C. Wade. 1994. Homology modeling of the Abl-SH3 domain, *Proteins.* 20: 203-215.

[34] Ghosh D., Z. Wawrzak, V.Z. Pletnev, N. Li, R. Kaiser, W. Pangborn, H. Jornvall, M. Erman and W.L. Duax. 1995. Structure of uncomplexed and linoleate-bound *Candida cylindracea* cholesterol esterase, *Structure.* 15: 279-288.

[35] Altschul S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman. 1990. Basic local alignment search tool, *J. Mol. Biol.* 215: 403- 410.

[36] Jones D.T. 1999. An efficient and reliable protein fold recognition method for genomic sequences, *J. Mol. Biol.* 287: 797-815.

[37] Kelley L.A., R.M. MacCallum and M.J.E. Sternberg. 2000. Enhanced genome annotation using structural profiles in the program 3DPSSM, *J. Mol. Biol.* 299: 499-520.

[38] Shi J., T.L. Blundell and K. Mizuguchi. 2001. FUGUE: Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties, *J. Mol. Biol.* 310: 243-257.

[39] Grochulski P., Y. Li, J.D. Schrag, F. Bouthillier, P. Smith, D. Harrison, B. Rubin and M. Cygler. 1993. Insights into interfacial activation from an open structure of *Candida rugosa* lipase., *J. Biol. Chem.* 268: 12843-12847.

[40] Schrag J.D. and M. Cygler. 1993. 1.8 A refined structure of the lipase from *Geotrichum candidum*, *J. Mol. Biol.* 230: 575-591.

[41] Wei Y., A.J. Contreras, P. Sheffield, T. Osterlund, U. Derewenda, R. Kneusel, U. Matern, C. Holm and Z.S. Derewenda. 1999. Crystal structure of Brefeldin A esterase, a bacterial homolog of the mammalian hormone-sensitive lipase, *Nat. Struc. Biol.* 6: 340 - 345.

[42] Thompson J.D., T.J. Gibson , F. Plewniak, F. Jeanmougin and D.G. Higgins. 1997. The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nuc. Acids Res.* 24: 4876-4882.

[43] Sali A. and T.L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.* 234: 779-815.

[44] Pearlman D.A., D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham III, S. DeBolt, N. Ferguson, G. Seibel and P. Kollman. 1995. 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.* 91: 1-41.

[45] Duan Y., C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang and P. Kollman. 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations, *J. Comp. Chem.* 24: 1999-2012.

[46] Berendsen H.J.C., J.P.M. Postma, W.F. van Gunsteren, A. Dinola and J.R. Haak. 1984. Molecular dynamics with coupling to an external bath, *J. Comp. Phys.* 81: 3684-3690.

[47] Ryckaert J.P., G. Ciccotti and H.J.C. Berendsen. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes, *J. Comput. Phys.* 23: 327-341.

[48] Verlet L. 1967. "Computer experiments" on classical fluids I. Thermodynamics properties of Lennard-Jones molecules, *Phys. Rev.* 159: 98-103.

[49] Kauzmann W. 1959. Some factors in the interpretation of protein denaturation, *Adv. Protein Chem.* 14: 1-63.

[50] Teller D.C. 1976. Accessible area, packing volumes and interaction surfaces of globular proteins., *Nature.* 260: 729-731.

[51] Janin J. 1976. Surface area of globular proteins., *J. Mol. Biol.* 105: 13-14.

[52] Chotia C. 1976. The nature of the accessible and buried surfaces in proteins., *J. Mol. Biol.* 105: 1-12.

[53] Doniach S. and P. Eastman. 1999. Protein dynamics simulations from nanoseconds to microseconds, *Curr. Op. Struc. Biol.* 9: 157-163.

[54] Daggett V. 2000. Long timescale simulations, *Curr. Op. Struc. Biol.* 10: 160-164.