Computational Evaluation of Protein Energy Functions

Nashat Mansour and Hussein Mohsen

Department of Computer Science and Mathematics Lebanese American University, Lebanon
{nmansour,hussein.mohsen}@lau.edu.lb

Abstract. Proteins are organic compounds made up of chains of amino acids that fold into complex 3-dimensional structures based on their chemical and physical properties. A protein is characterized by its 3D structure, which defines its biological function. Proteins fold into 3D structures in a way that leads to low-energy state. Predicting these structures is guided by the requirement of minimizing the energy value associated with the protein structure. However, the energy functions proposed so far by biophysicists and biochemists are still in the exploration phase and their usefulness has been demonstrated only individually. Also, assigning equal weights to different terms in energy has not been well-supported. In this project, we carry out a computational evaluation of putative protein energy functions. Our findings show that the CHARMM energy model tends to be more appropriate for *ab initio* computational techniques that predict protein structures. Also, we propose an approach based on a simulated annealing algorithm to find a better combination of energy terms, by assigning different weights to the terms, for the purpose of improving the capability of the computational prediction methods.

Keywords: CHARMM, force field, protein structure prediction, simulated annealing.

1 Introduction

Proteins are organic compounds that are made up of combinations of amino acids and are of different types and roles in living organisms. Initially a protein is a linear chain of amino acids, ranging from a few tens up to thousands of amino acids. Proteins fold, under the influence of several chemical and physical factors, into their 3-dimensional structures which determine their biological functions and properties. Misfolding occurs when the protein folds into a 3D structure that does not represent its correct native structure, which can lead to many diseases such as Alzheimer, several types of cancer, etc… [1]. Using computational methods for predicting the native structure of a protein from its primary sequence is an important and challenging task especially that this protein structure prediction (PSP) problem is computationally intractable. The protein's sequence of amino acids defines a unique native fold which normally corresponds to a minimum energy value [2]. In theory, this free energy minimum can be computed from quantum mechanics and, thus, should help in predicting the structure from the sequence. In practice, the theoretical foundation of such functions has not been fully established and several energy functions have been proposed.

D.-S. Huang et al. (Eds.): ICIC 2014, LNBI 8590, pp. 288–299, 2014.

The energy function models proposed so far depend on a number of biophysical factors. Their usefulness has been relatively demonstrated by different researchers. But, previous work has also shown that the precision of these energy models is not well-established [3]. Also, no serious comparative evaluation of these energy functions has been reported. Furthermore, limited work has been reported on the relative importance of the terms of the energy function; many decoys per protein, for a number of proteins, were generated from molecular dynamics trajectories and conformational search using the A-TASSER program minimizing the AMBER potential [4], [5].

In this work, we carry out a computational comparison of important energy functions that have appeared in the protein structure prediction literature in association with *ab initio* algorithms. We also design a simulated annealing algorithm for deriving values that should be used as weights for the energy terms based on the native structure knowledge on existing golden proteins in the 'protein data bank'. The ultimate goal is to yield better prediction of the tertiary structure of proteins.

This paper is organized as follows. Section 2 presents our methodology. Section 3 describes the energy models used for the comparative work. Section 4 explains the algorithm used for optimizing the weights of the energy terms. Section 5 presents the experimental results. Section 6 concludes the paper.

2 Methodology

The primary structure of a protein is a linear sequence of amino acids connected together via peptide bonds. Proteins fold due to hydrophobic effect, Vander Waals interactions, electrostatic forces, Hydrogen bonding, etc.... The protein structure prediction (PSP) problem is intractable [6]. Hence, the main computational approaches are heuristics for finding good suboptimal solutions and can be classified as: Homology modeling, threading, and ab initio methods [7]. For the latter methods, the only required input is the amino acid sequence whereas for the first two methods, data of previously predicted protein structures are used.

Ab initio methods predict the 3D structure of proteins given their primary sequences without relying on protein databases. The underlying strategy is find the best possible structure based on a chosen energy function. Based on the laws of physics, the most stable structure is the one with the lowest possible energy. We have identified three energy models/force fields as the most recognized models for pure ab initio PSP methods: the CHARMM model [8], the LINUS energy function [9], and AMBER [10]. The different energy functions include different terms and make a variety of assumptions. But, the relative merits of these functions in guiding computational protein structure prediction methods have not been well-studied. In particular, a computational investigation of their applicability has not been carried out. We believe that conducting such an investigation will serve the PSP research community by providing guidelines regarding the applicability of the recognized force fields.

Our methodology is based on the following steps and activities. We employ our recently developed computational method for PSP, namely the adapted scatter search metaheuristic [3], [11], as the basic platform for analyzing the behavior of the different energy functions. The selected energy functions are simulated and incorporated into the scatter search algorithm to create different versions of the scatter search based program. Then, real-world proteins are selected from a protein databank. The impact of the energy functions will be evaluated by computing the widely used root mean square deviation (RMSD) of the target structure with respect to the reference/golden protein structure. Then, we computationally derive sub-optimal weights for the energy terms in order to further improve the prediction. Then, for the ‘winner’ energy function, we find alternative weights for its terms to replace the commonly-used equal weights. This is done by adapting a simulated annealing algorithm that aims to simultaneously minimize the energy values of several (golden) proteins whose structure is already known.

3 Energy Functions/Models

The stability of the three-dimensional structure for protein is determined by the intramolecular interactions and interactions with the external environment. The search for stable conformations of proteins is based on the minimum total energy of interaction. The three recognized energy models, which are selected for our experimental study, are described in the following subsections.

3.1 CHARMM Energy Model

The Chemistry at HARvard Molecular Mechanics (CHARMM) function [12] is based on the dihedral planes representation of proteins that can be defined by the degrees of freedom given by the torsion angles. There are four torsion angles present in each amino acid in a protein: phi φ, psi ψ, omega ω, and chi χ. Phi is the angle between the planes C-N-Cα and N-Cα-C, where N-Cα is the axis of rotation. This angle decides the distance of C-C of two amino acids. The chi angle is between the planes formed by the atoms of the side chains, and side chains can have as many as five chi angles. Fig. 1 shows a segment of a protein backbone.

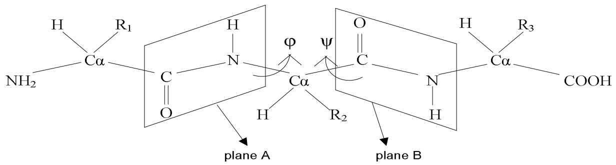

Fig. 1. Segment of a protein backbone with planes of bonds N-Cα and Cα-C, plane A and plane B, respectively

The CHARMM energy function is given by

$$
\begin{aligned}
E(\bar{c}) = {} & \sum_{\text{bonds}} K_b (b - b_o)^2 + \sum_{\textit{angles}} K_\theta (\theta - \theta_o)^2 + \\
& \sum_{\textit{improper dihedrals}} K_{\text{imp}} (\varphi - \varphi_o)^2 + \sum_{\textit{dihedrals}} K_\chi (1 + \cos(n\chi - \delta)) \\
& + \sum_{\textit{hydrogen}} \left(\frac{A_{ij}}{r_{ij}^{10}} - \frac{B_{ij}}{r_{ij}^{12}} \right) + \sum_{\textit{electrosta tic}} \frac{q_i q_j}{4\pi\varepsilon_o \varepsilon_r r_{ij}} \\
& + \sum_{\textit{van der waals}} 4\varepsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right)
\end{aligned}
$$

3.2 AMBER Energy Model

The AMBER99 model is composed of several all-atom force fields. These fields include parameters for bonded potential energy terms (stretching, bending, and torsion) and nonbonded terms (charge and van der Waals). The original version was AMBER94, which was developed to improve on peptide backbone torsion parameters; Kollman and co-workers used RESP charges derived from high-level ab initio calculations to parameterize energies [13]. The subsequent force field, denoted as AMBER99, is intended for use both with and without polarization effects [14]. AMBER99 includes the following terms:

$$
V_{bounded} = \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\textit{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\textit{dihedrals}} \frac{Vn}{2} [1 + \cos(n\varphi - \gamma)]
$$

$$
V_{nonbounded} = \sum_{i<j} \left\{ \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right] + \frac{q_i q_j}{4\pi\varepsilon R_{ij}} \right\}
$$

$$
V_{pol} = -\frac{1}{2} \sum_i \mu_i E_i^o
$$

The total energy includes the sum of all the potential fields and the polarization potential energy added in AMBER99. In the first equation , three terms represent contribution to the total energy from “bond stretching, bond angle bending, and torsion angle”, in the second one $V_{\text{nonbonded}}$ is the sum of non-bonded energy as “van der Waals and electrostatic energies”. In the third equation, V_{pol} the polarization value is calculated for each pair of point charge and induced point dipole moment as μ_i and E_i is the electrostatic field at the ith atom generated by all other point charges q_j.

3.3 LINUS Energy Model

LINUS (Local Independently Nucleated Units of Structure) is an ab initio method for simulating the folding of a protein on the basis of simple physical principles. LINUS involves a Metropolis Monte Carlo procedure, represents a protein by including all its heavy atoms (i.e., non-hydrogen atoms). LINUS developed by Srinivasan and Rose in 1995 is based on simple scoring function includes three components: hydrogen bonding, scaled contacts, and backbone torsion. The hydrogen bonding and the hydrophobic contact scores are calculated between pairs of atoms from residues. The contact energy is given by

$$\text{Contact energy} = \text{maximum value x } \left[1.0 - \frac{d_{ij}^2 - \sigma^2}{(\sigma + 1.4)^2 - \sigma^2} \right]$$

The scaled contact has both "Repulsive and attractive" terms. The repulsive term is implemented by rejecting conformations when the interatomic distance between any two atoms is less than the sum of their van der Waals radii. All pairwise interactions are evaluated except those involving carbonyl carbons and the attractive term is applied between the side chain pseudo-atoms. The total energy of the LINUS function scoring is given by the negative sum of the three preceding terms: hydrogen bonding, scaled contact, and backbone torsion [9].

4 Simulated Annealing Algorithm for Energy Terms' Weights

Simulated Annealing (SA) is a metaheuristic that deals with optimization problems with large search space to find near-optimal solutions. In this paper, simulated annealing is employed to assign appropriate weights to different energy terms. SA aims to yield a good solution with a minimized objective function value.

4.1 Algorithm Steps

The outline of the SA algorithm is shown in Fig. 2

```
Initialize Solution X and initial Temperature T
while (T > Steady State Temperature
                and iteration < MAX_ITERATIONs)
     for (NUMBER_OF_PERTURBATIONS)
    Y = Perturb(X)
    if (F(Y) < F(X) or e (-delta E/T) > random (0,1))
          X = Y
    // end if
    // end for
    Update (T)
// end while
```

Fig. 2. Outline of the SA algorithm

4.2 Solution Representation

Each energy term is given a random weight between 0 and 1, where the sum of weights must be 1 in any solution. The weight given to ES is called α, that to VW is called β, and to Torsion energy is δ. Thus, the obtained total energy for all solution is calculated according to the following formula:

$$E_{protein} = \alpha . Golden\text{-}ES_{protein} + \beta . Golden\text{-}VW_{protein} + \delta . Golden\text{-}Torsion_{protein}$$

4.3 Initial Solution

The initial solution is randomly generated. α, β, and δ are randomly generated with values between 0 and 1, such that α + β + δ = 1 and no single value falls below 1.

4.4 Initial Temperature, Maximum Iterations, and Number of Perturbation

Initial temperature is chosen as 400,000 and that of steady state is 15. The maximum number of iterations is chosen to be 200 with 3 perturbations in the 'for' loop of each iteration.

4.5 Perturbation Method

The implemented perturbation method alters the weights distribution among the energy terms. It either takes a certain percentage (0.1) from a target weight factor and distributes it to the other 2 weight factors or takes this percentage from two factors and adds it to the target third one.

The method has 3 decisions to take randomly. The first one is which of the weight factors (α, β, and δ) to target (whether to take from or add to). The second decision is whether to take from the chosen factor or give it more energy. The third decision is how to distribute the amount of energy taken/given.

For example, the algorithm may randomly first choose α as the target, then randomly choose to take from it 0.1 of its value, and then randomly chooses to give 0.3 of the amount to be taken from α to β and 0.7 to δ.

4.6 Objective Function

The objective function to be minimized is the following:

$$\sum \ (E_i - Avg_{Ei})^2, \qquad for\ i=1\ to\ n$$

where n is the number of the selected proteins, E_i (defined in subsection 4.2) is the total energy obtained according to the weights distribution by α, β, and δ, and Avg_{Ei} is the average of the calculated energies, E_i for i=1..n, in the set of proteins according to weight distributions. Effectively, we are minimizing the mean square deviation in the values of E_i over the n proteins as the values of the weights vary.

The SA metaheuristic algorithm enforces a lower bound of 0.1 for each of α, β, and δ in order not to prevent making any of the energy terms negligible. Thus, in any perturbed combination of values, the SA algorithm rejects any perturbation that yields weight values below 0.1.

5 Experimental Results

In this section, we demonstrate the performance of our proposed methods using the 3 different energy models for generating native-like structures for the backbones of the three target proteins.

5.1 Experimental Procedure for Energy Functions

In our experiments, we use three subject proteins with known structures in PDB. Our reference PDB is the Brookhaven database [15]. The 3 proteins are: 1CRN (CRAMBIN) is Plant seed protein (46 AAs); 1ROP (ROP Protein) is Transcription Regulation (56 AAs); 1UTG (UTEROGLOBIN) is Steroid Binding (70 AAs).

We run the scatter search program for predicting protein structures based on each of the three energy models. The results are evaluated by computing the target protein's structural difference from the reference/golden protein. This is accomplished by calculating the root mean square deviation (RMSD), in Angstrom, of the Cα atoms of the two proteins [16].

5.2 Experimental Results for Energy Functions

Table 1 gives the RMSD values by Scatter Search using the 3 types of potential energy function. These results show that CHARMM generates the lowest RMSD values for the 3 proteins. Also, the limited experiments reveal that the polarization term added in AMBER99 may not cause an improvement in the predicted protein structure. CHARMM also runs faster than AMBER and has comparable execution time to LINUS.

Although the results in Table 1 demostrate a rather clear tendency in favor of CHARMM, this does not yield a final conclusion since the number of proteins that are used in the experiment is small. Future work should employ many proteins with various sizes and functions in order to establish whether our result may vary with protein size and type.

5.3 Experimental Procedure for Energy Weights

In order to experiment with weights for energy terms, the real energy values have to be obtained. These are the golden energy values obtained from the data retrieved from the Protein Data Bank (PDB) file of each protein. The data are the 3D coordinates of atoms, torsion angles and amino acids. The total energy value of a protein is the sum of the terms of the CHARMM energy function.

Two sets of non-homologous proteins have been chosen, where each is made up of 5 different proteins. The first set contains proteins each with less than 100 amino acids (AAs), whereas the second contains proteins with more than 100 AAs in each. We run the SA algorithm on the data obtained from each set of proteins to generate values for the weights of the energy terms. The variation in the weights is also plotted over the SA iterations until convergence. Finally, to give an indication of the validity of the weight results, we rerun the scatter search program of Mansour et al. [11] to predict the structure of a protein, by using the new weights, and compare the resulting RMSD with that of the structure produced by using equal weights for the energy terms.

Table 1. Experimental results for SS on 3 energy functions

Energy function	1CRN (46 AAs, 326 atoms)		1ROP(56 AAs, 420 atoms)		1UTG(70 AAs, 547 atoms)	
	Time (min)	RMSD	Time (min)	RMSD	Time (min)	RMSD
CHARMM	964	9.39	1059	11.52	1182	13.56
AMBER96	1140	11.80	1320	15.84	1800	16.21
AMBER99 with Polarization	3600	14.39	4200	16.04	5005	18.07
LINUS	960	13.05	1005	15.22	1200	18.36

5.4 Results for Energy Weights

Results of SA for Proteins of Less than 100 Amino Acids. For proteins with less than 100 amino acids (1RPO, 1UBQ, 1CRN, 1ROP, and 1UTG), SA converged after 20 iterations to the objective function value of 5.159x10^{15}. The initial randomly generated solution was of objective function value of 2.178x10^{17}. Fig. 3 shows the variations of α, β, and δ as a function of iterations of the SA. The values of the three weights at convergence are: α = 0.352; β = 0.105; δ = 0.543.

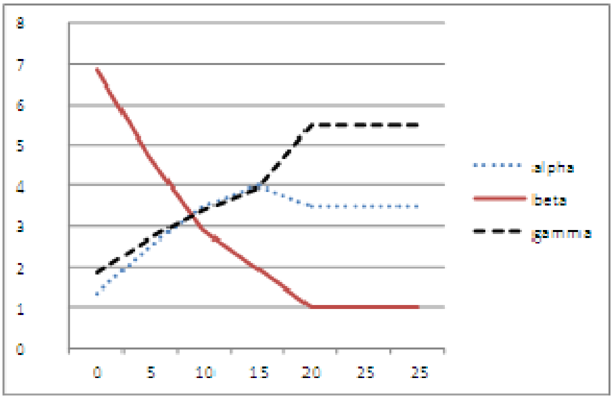

Fig. 3. Variations of α, β, and δ (x10^{-1}) as a function of iterations

Comparing the energy values obtained by using these weights with those obtained from the equal weights for the 1ROP protein show that the former ones (using the new weights) are lower. Equal weights (0.333) give the energy value of 1.125x10^{6}, whereas the obtained solution (with α, β, and δ values) gives the value of 426,072. Fig. 4 shows the values of ES, VW and Torsion energy terms as a function of iterations and the corresponding values of α, β, and δ.

After running the scatter search code, the obtained 1ROP protein structure has an RMSD with respect to the golden protein of 7 Angstrom in comparison with about 12 Angstrom, obtained with equal weights [11].

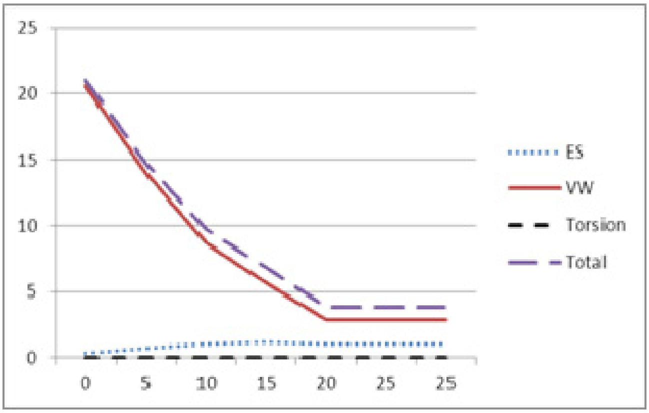

Fig. 4. Variations of ES, VW and Torsion Energies for 1ROP (in 10^{5} charge units) as a function of iterations

Results of SA for Proteins of More Than 100 Amino Acids. For proteins with more than 100 amino acids (1AAJ, 1BP2, 1RPG, 1RRO, and 1YCC), SA converged after 43 iterations to the objective function value of 1.457x10^{12}. The initial randomly generated solution was of objective function value of 3.278x10^{13}. Fig. 5 shows the variations of α, β, and δ as a function of iterations of the SA. The values of the three weights at convergence are: α = 0.104; β = 0.108; δ = 0.788.

Comparing the energy values obtained by using these weights with those obtained from the equal weights for the 1AAJ protein shows that the new weights yield lower values. Specifically, equal weights for all energy terms give the value of: 2.305x10^{6}. The obtained solution (with different α, β, and δ) gives the value of: 742,844. Fig. 6 shows the values of ES, VW and Torsion energy terms as a function of iterations and the corresponding values of α, β, and δ.

After running the Scatter Search code, the obtained 1AAJ protein structure has an RMSD with the golden of 5.84 Angstroms, which is a promising result.

By inspecting the values of each of the 3 energy terms, it is clear that in the case of equal weights, the VW term used to dominate the energy function and, consequently, the resulting protein structure. The change in the weight values provides fairer shares to the other two terms and, thus, allows them to influence the evolution of the predicted structure. We also note that the relative weights are different for different protein sizes.

These results show that assigning equal weights to the energy terms does not yield the best possible protein structure prediction. But, more experiments are required to establish what weights should be assigned for what size-category of proteins. Also, it will be useful to apply our approach to other recognized energy functions.

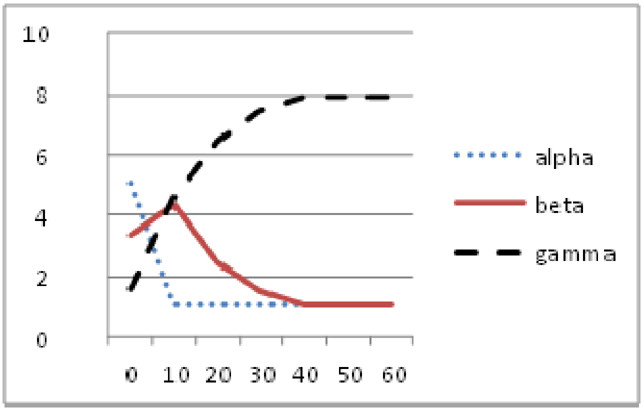

Fig. 5. Variations of α, β, and δ (x10^{-1}) as function of iterations

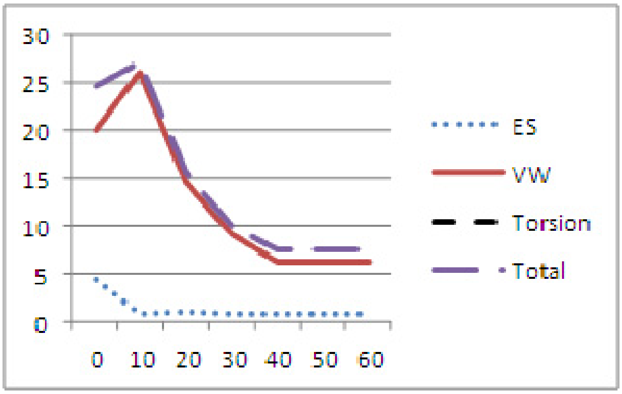

Fig. 6. Variations of ES, VW and Torsion Energies in 1AAJ (in 10^5 charge units) as a function of iterations

6 Conclusions

We have carried out a computational assessment of the ability of three recognized energy models for predicting the tertiary structure of proteins using a pure ab initio algorithm. We have also investigated the merit of assigning unequal weights to the terms included in energy functions employed for ab initio protein structure prediction.

Our experimental results show that the CHARMM energy model tends to yield better protein structures than the other two energy functions, next is AMBER without polarization, followed by LINUS and the polarization version of AMBER. We have also found that it is more favorable for computational protein prediction methods to assign unequal weights to the terms in the energy function. As a result, we recommend the following weight values for the CHARMM energy function: 0.1-0.3 for the electrostatic term, 0.1-0.2 for the Vander Waals term, and 0.6-0.8 for the Torsion term, when applied to small to medium size proteins.

This work has established the merit of the proposed approach. Further work can focus on extending the experimental work to a larger number of protein sets that include proteins with different sizes and functions and on extending to other energy models.

Acknowledgments. This work was partially supported by the Lebanese American University and Lebanon's National Council for Scientific Research. We thank Mirvat Sibai for guidance on biology issues and acknowledge the coding support of Rachid Sayed.

References

1. Prusiner, S.B.: Prions. Proceedings of the National Academy of Sciences of the United States of America 95, 13363–13383 (1998)
2. Anfinsen, C.B.: Principles that Govern the Folding of Proteins. Science, 181–187 (1973)
3. Mansour, N., Kehyayan, C., Khachfe, H.: Scatter Search Algorithm for Protein Structure Prediction. Int. J. Bioinformatics Res. Appl. 5, 501–515 (2009)
4. Wroblewska, L., Skolnick, J.: Can a Physics-Based, All-Atom Potential Find a Protein's Native Structure Among Misfolded Structures. J. Comput. Chem. 28, 2059–2066 (2007)
5. Wroblewska, L., Jagielska, A., Skolnick, J.: Development of a Physics-Based Force Field for the Scoring and Refinement of Protein Models. Biophysical J. 94, 3227–3240 (2008)
6. Unger, R., Moult, J.: Genetic Algorithms for Protein Folding Simulations. J. Mol. Biol. 231, 75–81 (1993)
7. Sikder, A.R., Zomaya, A.Y.: An Overview of Protein-Folding Techniques: Issues and Perspectives. Int. J. Bioinformatics Res. Appl. 1, 121–143 (2005)
8. Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., Mackerell Jr., A.D.: CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. J. Comput. Chem. 31, 671–690 (2009)
9. Srinivasan, R., Fleming, P.J., Rose, G.D.: Ab Initio Protein Folding Using LINUS. Methods in Enzymology 383, 48–66 (2004)
10. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., Kollman, P.A.: A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J. Am. Chem. Soc. 117, 5179–5197 (1995)
11. Mansour, N., Ghalayini, I., Rizk, S., El Sibai, M.: Evolutionary Algorithm for Predicting All-Atom Protein Structure. In: Int. Conf. on Bioinformatics and Computational Biology, New Orleans (2011)

12. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., Karplus, M.: CHARMM: a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. Journal of Computational Chemistry 4, 187–217 (1983)
13. Kollman, P.A.: Advances and Continuing Challenges in Achieving Realistic and Predictive Simulations of the Properties of Organic and Biological Molecules. American Chemical Society (1996)
14. Wang, J., Cieplak, P., Kollman, P.A.: How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? J. Comp. Chem. 21, 1049–1074 (2000)
15. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.: The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535–542 (1977)
16. Carugo, O., Pongor, S.: A Normalized Root-Mean-Square Distance for Comparing Protein Three-Dimensional Structures. Protein Sci. 10, 1470–1473 (2001)