Comparative Study of Computational Strategies for Protein Structure Prediction

Fanny G. Maldonado-Nava, Juan Frausto-Solís, Juan Paulo Sánchez-Hernández, Juan Javier González Barbosa and Ernesto Liñán-García

Abstract Protein Folding Problem (PFP) is one of the most challenging problems of combinatorial optimization with applications in bioinformatics and molecular biology. The aim of PFP is to find the three-dimensional structure of a protein, this structure is known as Native Structure (NS), which is characterized by the minimal energy of Gibbs and it is commonly the best functional structure. To find an NS knowing only the amino acids sequence (primary structure) of a protein is known as ab initio problem. A protein can take a huge number of different conformational structures from its primary structure to the NS. For solving PFP, several computational strategies are applied in order to search structures of protein on a huge space of possible solutions. In this work, the most popular methods and strategies are compared, and advantages and disadvantages of them are discussed.

Keywords Protein folding problem · Computational strategies
Ab initio · Threading · Homology

F. G. Maldonado-Nava (✉) · J. Frausto-Solís · J. J. González Barbosa
TecNM/Instituto Tecnológico de Ciudad Madero, Ciudad Madero, Mexico
e-mail: fanny_mn@hotmail.com

J. Frausto-Solís
e-mail: juan.frausto@itcm.edu.mx

J. J. González Barbosa
e-mail: jjgonzalezbarbosa@itcm.edu.mx

J. P. Sánchez-Hernández
Universidad Politécnica del Estado de Morelos, Jiutepec, Mexico
e-mail: juan.paulosh@upemor.edu.mx

E. Liñán-García
Universidad Autónoma de Coahuila, Saltillo, Mexico
e-mail: ernesto_linan_garcia@uadec.edu.mx

O. Castillo et al. (eds.), *Fuzzy Logic Augmentation of Neural and Optimization Algorithms: Theoretical Aspects and Real Applications*, Studies in Computational Intelligence 749, https://doi.org/10.1007/978-3-319-71008-2_32

1 Introduction

Proteins are molecules, which play a central role in our body. Proteins are needed to catalyze reactions, transport molecules, and other important functions. Proteins consist of smaller units named amino acids, attached to one another in long chains by peptide bonds. A functional protein has a specific three-dimensional structure, usually named Native Structure (NS), which takes when it is correctly folded. The NS is biologically active, in which the protein correctly performs its functions. The natural process of protein folding is not completely understood; this is because nature takes an unknown path to achieve the native structure in a very fast way [1].

The process of protein folding in living organisms (Natural process of protein folding, folding of proteins in vivo; or in short, folding) occurs within cells, which as is well known are prokaryotes in all bacteria and eukaryotes for animals, plants, and fungi. Understanding the process of protein folding is important because many human diseases are related to improper folding in vivo; some of these diseases are [2–4]: Alzheimer's, Parkinson's, Prion, Tauopathy, Huntington's disease, Creutzfeldt-Jakob disease, Cystic Fibrosis, Gaucher disease, and Sickle Cell Anemia. In fact, recent specialized publications have noticed that incorrect folding of proteins (or misfolded) is involved with most of the diseases not caused by infectious agents and is involved in the progression of hundreds of diseases [4, 5].

Protein Folding Problem has been studied for the last 50 years and is one of the biggest unsolved problems in science [3, 6]. PFP is an NP-Hard problem [7, 8], which consists in determining the native structure of a protein, this structure is the one in which the Gibbs free energy is the lowest [9]. Due to the amount of conformations that a protein can take, computational methods are becoming important. Some methods for the study of the tertiary structure of the proteins have been developed are X-ray Crystallography and Nuclear Magnetic Resonance (NMR). These methods are regularly very expensive and their processes can consume very long time [10, 11]. Thus, the NS prediction is necessary and it has become one of the most important challenges of modern computational biology [7]. Different computational approaches for finding the three-dimensional structure have been proposed over the last decades. These approaches can be classified into three categories: (a) ab initio, (b) homology, and (c) threading. The main challenge is to understand how the information included in the amino acids sequence can be translated into a three-dimensional structure (functional structure), in order to develop computational algorithms that can predict a protein structure correctly.

Over the last decades, many algorithms have been proposed and tested as a solution to PFP. Most common algorithms are Simulated Annealing (SA), Genetic Algorithms (GA), Ant Colony Optimization (ACO), Tabu Search (TS), and among other. The most successful algorithms for solving PFP are SAL algorithms (Simulated Annealing Like algorithms) [12]; these successful methods are usually hybridized with other heuristics. Despite the efforts made so far, just a little number of protein sequences have been solved, which has motivated the scientific community on working on more powerful algorithms [10]. Recently, new and more efficient SAL algorithms have been proposed; as Golden Ratio Simulated Annealing (GRSA), which is part of these successful SAL algorithms [13]. GRSA is important because has obtain very good results in the case of peptides, particularly the Met-enkephalin, which is commonly studied in PFP area.

This paper is organized as follows: in Sect. 2, three strategies for Protein Folding Problem are presented. Section 3, describes three important methods presented in CASP. In Sect. 4, PFP for ab initio approach is described and an energy function is presented. Finally, conclusions for this work are discussed.

2 Computational Strategies

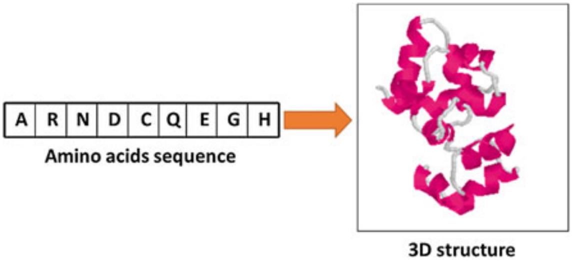
Many computational methodologies and algorithms have been proposed as a solution to the PFP. Strategies used in these algorithms can be classified in three categories: ab initio, homology, and threading. The main difference between these strategies is the information they need to address the problem.

2.1 *Ab Initio Approach*

Ab initio strategy is perhaps the most difficult approach for protein structure prediction. As is shown in Fig. 1, ab initio looks for the three-dimensional structure using only the amino acids' sequence and it does not require other information of the target protein. Ab initio methods are based on basic physics and quantum mechanics, this is on the thermodynamic hypothesis which points out that the NS of a protein is the one for which the free energy achieves the global minimum [9].

Ab initio methods provide a natural approach to obtain structures from protein sequences without referring any information or any appropriate templates. This strategy is clearly the most difficult, but the most useful approach. As any other strategy, ab initio presents some advantages and disadvantages. Ab initio methods are useful when appropriate templates cannot be consulted, that is, when sufficiently homologous proteins have not been found or when the template does not provide an appropriate structure. New folds can be predicted by this strategy, since there are

Fig. 1 Ab initio approach



still proteins whose native structures have not been solved, an ab initio method does not need templates from any library. This strategy requires a lot of computational processing time because of the complexity of the problem. In addition, because PFP is an NP-hard problem [7], heuristic algorithms are currently considered as the best alternative; however, these algorithms do not guarantee to achieve exactly the optimal solution. As a consequence, the research of ab initio algorithms is focused on peptides and proteins with a limited number of amino acids (60–150). However, to study small proteins could lead to finding general algorithm solutions for solving the real challenge that is PFP.

For ab initio strategy, PFP is considered as an optimization problem, where the goal is identifying the values of the variables (angles) which describe the minimum energy of the protein. Ab initio methods simulate the protein conformational space using an energy function, which describes the internal energy of the protein and its interactions with the environment. An ab initio algorithm consists of three components: (1) a geometric representation, (2) an energy function, and (3) a searching technique.

2.2 Homology Approach

Known as comparative modeling or template-based modeling, this strategy is based on the understanding of protein evolution, mainly in two facts: (a) proteins that have a homologous sequence, will have similar three-dimensional structures, and (b) proteins structures are more conserved than their sequences. Many proteins can be solved by this approach. Figure 2 illustrates this strategy. Homology process starts with the identification phase, in which an identification of homologous proteins should be done from PDB (Protein Data Bank), phase two is an alignment, which is carried out between the target protein and its homologous (template), and next, a method for modifying the structure should be applied for optimizing the model and get to the final three-dimensional structure of the target protein.

Comparative modeling exploits the fact that evolutionarily related proteins with similar sequences, as measured by the percentage of identical residues at each

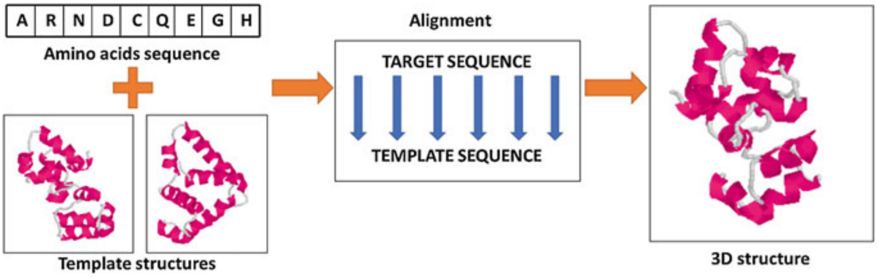


Fig. 2 Homology approach

position based on an optimal structural superposition, often have similar structures. The complexity of the problem becomes smaller than other strategies, since this approach takes advantage of the reduction of the conformational search space, because the process uses a template of a protein whose three-dimensional structure has already been found. When a homolog protein is found, this method is applicable to almost all proteins [14]. If the homology between proteins is high (bigger than 35%) the three-dimensional structure can be found in many cases [15]; however, the use of templates and heuristic algorithms may obtain the NS in almost of the cases.

One of the main disadvantages of this strategy is that only structures of proteins with known homologous sequences can be predicted. If the degree of homology is low, the method must use a more powerful algorithm to be able to find the three-dimensional structure, since with a lower homology the quality of the model will be smaller.

2.3 Threading Approach

Known as fold recognition, this strategy construct protein from known templates even if there is no homologous protein deposited in the Protein Data Bank. Threading models the protein with experimental structures as templates, is a different approach from the homology in terms of the methodology. In Fig. 3, this strategy is shown. The term threading is stand for the process of aligning a protein sequence into a backbone structure and evaluate the compatibility with a set of potential scores or energy functions. Threading is based on the observation that the number of unique protein folds in nature is much smaller than number of proteins.

During the process of threading, the target protein is placed, following the sequential order, into structural positions of a template three-dimensional structure in an optimal way. This process consists of two phases: (1) select a structural template from a library, and (2) find the correct replacement between the target protein against the structural models in the space of possible replacements. Threading has some advantages; it uses known protein structures as templates for

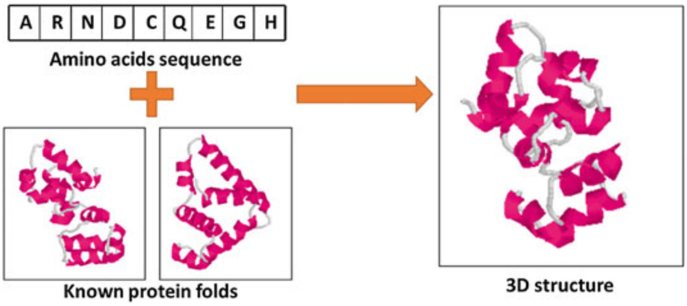


Fig. 3 Threading approach

sequences of unknown structures. Threading finds the most similar conformation to the NS that can be uses as an initial solution with other methods. Threading presents some limitations; these methods are computationally expensive. Identifying appropriate templates for a given protein is also a problem classified as NP-Hard [16]. In addition, the NS found with this approach could not be present in the space of possible conformations.

Many algorithms implement different metaheuristics to provide near optimal solutions for PFP, considering the limitations and the advantages of the approaches for protein structure prediction methods, researchers have developed hybrid methods in their algorithms, which combine principles of the three strategies presented in this paper.

3 Methods

There is a biannual competition named CASP (Critical Assessment of protein Structure Prediction), in this competition researchers test their structure prediction methods. Targets proteins for structure prediction are structures solved, but they are kept on hold by the Protein Data Bank. Here are presented three protein structure prediction methods, which use different approaches and different strategies for constructing three-dimensional protein models. These methods have been presented and tested in CASP, obtaining good results, so that they have obtained first places in lasts competitions.

I-TASSER

I-TASSER is a server for protein structure predictions, built by Zhang Lab. This server was ranked as the number one server for protein structure prediction in CASP7, CASP8, CASP9, CASP10, CASP11, and CASP12 experiments. The I-TASSER method is divided in three phases: threading, assembling, and refinement. In the first phase, I-TASSER identifies templates from the PDB (Protein Data Bank) by a threading approach using LOMETS (which combines algorithms to generate models by collecting their target-template alignments). In the second phase, fragments of the threading aligned regions are extracted from the template structures, and are used to assemble new structural conformations, while ab initio approach processes the unaligned regions. The assembly is performed by a replica-exchange Monte Carlo (REMC) Simulation. The low free-energy states are identified by SPICKER (algorithm to identify the near-native models) through clustering. In the third phase, a second assembly is performed, the purpose of the second iteration is to refine the global topology of the cluster centroids. The lowest energy structures are selected, and the final full-atomic models are obtained by REMO, and fragment-guided molecular dynamics [10, 17].

QUARK

QUARK is an ab initio structure prediction built by Zhang Lab, which construct 3D structures models. QUARK was ranked as the No 1 server in free-modeling in

CASP9 and CASP10 experiments. QUARK models are built from small fragments (1–20 residues long) by replica-exchange Monte Carlo simulation. This procedure can be divided into three steps. The first step is for multiple feature predictions and fragment generation starting from one query sequence. QUARK first predicts a variety of selected structural features by neural network (NN). In the second step, the global fold is generated by replica-exchange Monte Carlo (REMC) simulations by assembling the small fragments, these fragments in QUARK have multiple sizes from 1 to 20 residues. The third step is full-atomic refinement. QUARK simulations perform movements of free-chain constructions and fragment substitutions between decoy and fragment structures. These techniques have increased the efficiency of conformational search while taking the advantage of the reduction of the conformational search owing to fragment assembly [10, 18].

ROSETTA

ROSETTA is a fragment-based method for the three-dimensional protein structure prediction problem developed by Baker Lab. Is one of the best-established ab initio protein folding methods as demonstrated in the last CASP experiments. ROSETTA uses an assembly strategy to combine native-like structures of fragments of unrelated protein structures with similar local sequences using Bayesian scoring functions. The main goal of ROSETTA scoring function is to search for the most probable structure of a protein given the amino acid sequence. This algorithm predicts protein structures based on a library of residue fragments. The fragments are selected according to their sequence similarity with the target protein. The Rosetta method assumes that short sequence segments have strong local structural biases. In the first step, fragment libraries for each 3- and 9-residue segment of the target protein are extracted from the protein structure database. Then, tertiary structures are generated using a Monte Carlo search of the possible combinations, minimizing a scoring function [10, 19].

4 Protein Folding Problem

Protein folding problem is the process of finding the three-dimensional native structure of a protein, this structure is usually named Native Structure (NS). NS is the conformation in which the protein performs its biological role. As mentioned earlier, the PFP since the ab initio approach can be considered as an optimization problem, where the goal is identifying the set of values of the variables that satisfy an objective function, that in this case is the energy function. PFP is an enormous challenge because the space of possible conformations a protein can take is extremely large [7]. For an ab initio approach PFP can be defined as follows:

- A sequence of n amino acids; $a_1, a_2, a_3, \ldots, a_n$, that represents the primary structure of a protein, with a set of dihedral angles $\sigma^m = \sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_m$,
- An energy function $f(\sigma_1 \sigma_2 \ldots \sigma_m)$ that represents the free energy.

The solution to this problem is to find the native structure such that $f^*(\sigma_1\sigma_2\ldots\sigma_m)$ represents the minimum energy value, where the optimal solution $\sigma^* = \sigma_1\sigma_2\ldots\sigma_m$ defines the best three-dimensional configuration.

The atoms of a protein are represented in three-dimensional Cartesian coordinates. There are four types of torsion angles or dihedral angles presented in Fig. 4, and defined below:

- Phi (ϕ) is the angle between the amino group and the alpha carbon. Represents the angle between the amino group (or NH_2) of the amino acid i, and the alpha Carbon C_i in the sequence; it represents the bond angle between the N_i atom of amino group and the alpha carbon (αC_i).
- Psi (ψ) is the dihedral angle between the alpha carbon and the carboxyl group. Psi represents the angle between the carboxyl $(COOH_i)$ group of the amino acid i, and the alpha carbon i (C_i) of the same amino acid. Psi measures the angle of the covalent bond between the C_i of the carboxyl group, and the alpha carbon (αC_i).
- Omega (ω) is defined for each two consecutive amino acids; it is the angle of the covalent bond between the atom N_i of amino acid i, and carbon $C_{(i-1)}$ of the carboxyl group of the amino acid $(i-1)$.
- And, Chi (χ) is defined between the two planes conformed by two consecutive carbon atoms in the radical group.

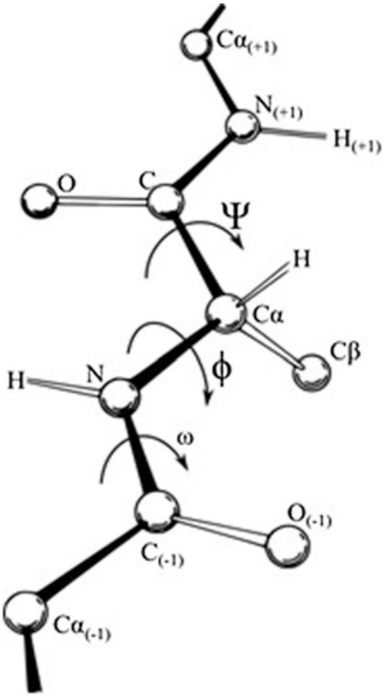


Fig. 4 Representation of the four dihedral angles

The PFP variables are the set of dihedral angles that satisfies the minimum energy value.

4.1 Energy Function

The protein's energy depends on the interaction among their atoms (angles and distance). Force fields are used to measure the energy of a protein; these include many interactions among atoms affecting different energies [20]. A force field includes terms associated with the bond interactions, and terms associated with no-bond interactions. Some of the most popular and successful force fields are CHARMM [21], AMBER [22], ECEPP/2 and ECEPP/3 [23].

One of the most used energy functions for PFP is ECEPP/2, that is a relatively simple force field based on rigid geometry (i.e., constant bond angles and lengths), with conformations thus defined solely by the backbone and side chain dihedral angles. In ECEPP/2 the potential energy is given by the sum of the electrostatic term E_{elect}, Lennard-Jones term E_{LJ}, and hydrogen-bond term E_{HB} for all pairs of atoms in the peptide together with the torsion term E_{tor} for all torsion angles [24]:

$$E_{bonded} = E_{elect} + E_{LJ} + E_{HB} + E_{tor} \tag{1}$$

These terms in Eq. (1) are expressed in Eq. (2) through which energy function ECEPP/2 minimize the energy [24].

$$\begin{aligned} E_{total} = & \sum_{j>i} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + 332 \sum_{j>i} \frac{q_i q_j}{\varepsilon r_{ij}} + \sum_{j>i} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ & + \sum_{n} U_n (1 \pm \cos(k_n \varphi_n)) \end{aligned} \tag{2}$$

where:

- r_{ij} is the distance in Å between the atoms i and j.
- A_{ij}, B_{ij}, C_{ij} and D_{ij} are the parameters of the empirical potentials.
- q_i and q_j are the partial charges on the atoms i and j, respectively.
- ε is the dielectric constant which is usually set to $\varepsilon = 2$.
- 332 is a factor for using the energy units expressed in kcal/mol.
- U_n is the energetic torsion barrier of rotation about the bond n.
- k_n is the multiplicity of the torsion angle φ_n.

The energy function ECEPP/3 is a modify version of ECEPP/2. ECEPP/3 contains updated parameters for proline and oxyproline residues. This energy function is used until recently for PFP.

5 Conclusions

The study of protein folding problem for finding the three-dimensional structure is one of most important research problems in Bioinformatics. Over the last decades, computational methods, and algorithms have been developed for solving PFP. However, there is no method yet that can predict structures without the need of information about templates, this is because of the complexity and high conformational search space, so that the problem still challenges in bioinformatics and computer science. Three strategies were described in this paper, these strategies are now use in algorithms and methods for PFP. Some of these methods are ROSETTA, I- TASSER and QUARK, which have been three of the most successful predictors in the CASP competition. A common characteristic of these methods is that for some of their processes they use a Monte Carlo method.

Acknowledgements The authors would like to acknowledge with appreciation and gratitude to CONACYT. Also, acknowledge to Laboratorio Nacional de Tecnologías de la Información (LaNTI) of the Instituto Tecnológico de Ciudad Madero for the access to the cluster. Fanny Gabriela Maldonado-Nava would like to thank CONACYT for the support in the project 429028.

References

1. C. Levinthal, Are there pathways for protein folding. J. Chim. Phys. **65**(1), 44–45 (1968)
2. T.K. Chaudhuri, S. Paul, Protein-misfolding diseases and chaperone-based therapeutic approaches. FEBS J. **273**(7), 1331–1349 (2006)
3. K.A. Dill, J.L. MacCallum, The protein-folding problem, 50 years on. Science (80-.) **338** (6110), 1042–1046 (2012)
4. J.S. Valastyan, S. Lindquist, Mechanisms of protein-folding diseases at a glance. Dis. Model. Mech. **7**(1), 9–14 (2014)
5. C. Spiess, A.S. Meyer, S. Reissmann, J. Frydman, Mechanism of the eukaryotic chaperonin: protein folding in the chamber of secrets. Trends Cell Biol. **14**(11), 598–604 (2004)
6. K.A. Dill, S.B. Ozkan, T.R. Weikl, J.D. Chodera, V.A. Voelz, The protein folding problem: when will it be solved? Curr. Opin. Struct. Biol. **17**(3), 342–346 (2007)
7. W.E. Hart, S. Istrail, Robust proofs of NP-hardness for protein folding: general lattices and energy potentials. J. Comput. Biol. **4**(1), 1–22 (1997)
8. J.T. Ngo, J. Marks, M. Karplus, in *The Protein Folding Problem and Tertiary Structure Prediction.* Computational Complexity, Protein Structure Prediction, and the Levinthal Paradox (Birkhäuser Boston, Boston, 1994), pp. 433–506
9. C.B. Anfinsen, Principles that govern the folding of protein chains. Science (80-.) **181**(4096), 223–230 (1973)
10. M. Dorn, M.B. e Silva, L.S. Buriol, L.C. Lamb, Three-dimensional protein structure prediction: methods and computational strategies. Comput. Biol. Chem. **53**, 251–276 (2014)
11. A.A. Yee, A. Savchenko, A. Ignachenko, J. Lukin, X. Xu, T. Skarina, E. Evdokimova, C.S. Liu, A. Semesi, V. Guido, A.M. Edwards, C.H. Arrowsmith, NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins. J. Am. Chem. Soc. **127** (47), 16512–16517 (2005)
12. L.B. Morales, R. Garduño-Juárez, D. Romero, Applications of simulated annealing to the multiple-minima problem in small peptides. J. Biomol. Struct. Dyn. **8**(4), 721–735 (1991)

13. J. Frausto-Solis, J.P. Sánchez-Hernández, M. Sánchez-Pérez, E.L. García, Golden ratio simulated annealing for protein folding problem. Int. J. Comput. Methods **12**(6), 1550037 (2015)
14. Y. Zhang, J. Skolnick, The protein structure prediction problem could be solved using the current PDB library. Proc. Natl. Acad. Sci. **102**(4), 1029–1034 (2005)
15. G. Helles, A comparative study of the reported performance of ab initio protein structure prediction algorithms. J. R. Soc. Interface **5**(21), 387–396 (2008)
16. R.H. Lathrop, The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng. Des. Sel. **7**(9), 1059–1068 (1994)
17. J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER Suite: protein structure and function prediction. Nat. Methods **12**(1), 7–8 (2015)
18. Y. Zhang, Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins Struct. Funct. Bioinforma. **82**(SUPPL. 2), 175–187 (2013)
19. C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, D. Baker, Protein structure prediction using rosetta. Methods Enzymol. **383**, 66–93 (2004)
20. K.A. Dill, Dominant forces in protein folding. Biochemistry **29**(31), 7133–7155 (1990)
21. B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. **4**(2), 187–217 (1983)
22. J.W. Ponder, D.A. Case, Force fields for protein simulations. Adv. Protein Chem. **66**, 27–85 (2003)
23. F.A. Momany, R.F. McGuire, A.W. Burgess, H.A. Scheraga, Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. J. Phys. Chem. **79**(22), 2361–2381 (1975)
24. F. Eisenmenger, U.H.E. Hansmann, S. Hayryan, C.K. Hu, [SMMP] A modern package for simulation of proteins. Comput. Phys. Commun. **138**(2), 192–212 (2001)