Soft Computing (2019) 23:4777–4788
https://doi.org/10.1007/s00500-018-3130-0

METHODOLOGIES AND APPLICATION

A novel approach for protein structure prediction based on an estimation of distribution algorithm

Amir Morshedian[1] · Jafar Razmara[1] · Shahriar Lotfi[1]

Published online: 23 March 2018

Abstract
Protein structure prediction is one of the major challenges in structural biology and has wide potential applications in biotechnology. However, the problem is faced with a difficult optimization requirement with particularly complex energy landscapes. The current article aims to present a novel approach namely AHEDA as an evolutionary-based solution to overcome the problem. AHEDA uses the hydrophobic-polar model to develop a robust and efficient evolutionary-based algorithm for protein structure prediction. The method utilizes an integrated estimation of distribution algorithm that attempts to optimize the search process and prevent the destruction of structural blocks. It also uses a stochastic local search to improve its accuracy. Based on a comprehensive comparison with other existing methods on 24 widely used benchmarks, AHEDA was shown to generate highly accurate predictions compared to the other similar methods.

Keywords Estimation of distribution algorithm (EDA) · Protein structure prediction (PSP) · HP model · Protein folding · Stochastic local search (SLS)

1 Introduction

Proteins are known as the basic macromolecules that regulate all biological activities in cellular organisms. Proteins are synthesized within live cells and then immediately fold into a three-dimensional structure. Such a structure is uniquely specified by its constituent sequence of amino acids (Anfinsen 1973). In other words, the primary structure of a protein determines its three-dimensional structure, and in turn, this structure assists to understand its functional behavior. The protein structure prediction and structural comparison of proteins are two vital steps in structural biology. For quite a long period of time, the 3D protein structures derived with computational approaches were not trusted very much by most experimental scientists. Actually, they only trusted the 3D protein structure determined by X-ray and NMR techniques and thought computational structures were unreliable. Although X-ray crystallography is a powerful tool in determining protein 3D structures, it is time-consuming and expensive, and not all proteins can be successfully crystallized. Membrane proteins are difficult to crystallize, and most of them will not dissolve in normal solvents. Therefore, so far very few membrane protein structures have been determined. NMR is indeed a very powerful tool in determining the 3D structures of membrane proteins, but it is also time-consuming and costly (Guntert 2004).

The complexity of the problem has led researchers to utilize heuristic algorithms to achieve the best performance (Razmara et al. 2013). Several highly accurate methods have been previously proposed using homology model building methods based on the similarity of a new protein sequence to other structurally known proteins. Nevertheless, the full accurate (100%) prediction of protein structure is not theoretically attainable yet. Further, accurate de novo prediction of protein structure is hard to compute where no homologue protein is known. Consequently, derivation of high accurate structure prediction methods is still an active research area. The interest is essentially increased with continuous growth of the protein sequence databases. The techniques developed for protein secondary structure prediction involve a diverse

Communicated by V. Loia.

✉ Jafar Razmara
razmara@tabrizu.ac.ir

Amir Morshedian
a.morshedyan@gmail.com

Shahriar Lotfi
shahriar_lotfi@tabrizu.ac.ir

[1] Department of Computer Science, Faculty of Mathematical Sciences, University of Tabriz, Tabriz, Iran

variety of biotechnological uses including structure-based drug design (Davis and Baker 2009), biofuels, and getting structures from data related to incomplete nuclear magnetic resonance (Shen et al. 2009; Raman et al. 2010). In addition, function of proteins is attributed to sequences of their structures benefiting from the large-scale information introduced in genome projects depending on the ability of predicting the native structure of proteins.

Majority of the methods involved in protein secondary structure prediction are based on the thermodynamics hypothesis. In other words, the conformation that is adopted by proteins under physiological conditions has the lowest rate of the free energy (Anfinsen 1973; Bujnicki 2006). In this regard, the problem could be formulated as a problem of energy minimization and be divided into two subproblems: the former problem is related to defining an appropriate function of energy placing native structures on their global minimum and being able to distinguish the correct and incorrect folds. The later problem is development of an efficient and strong strategy for searching that is able to handle quite a large number of variables and a highly degenerated and complex landscape of energy.

It is a common practice to adopt models with a reduced complexity in protein structure prediction-related studies, e.g., lattice models (Storm and Lyngsø 1999). Lattice models could be implemented in extracting the fundamental principles of folding, making predictions, and unifying knowledge related various features of proteins by relinquishing atomic details, without any need to the related costs of computation. The straightforward and clear definition of energy function for lattice models is a facilitating factor for development of a fast and strong optimization method. When dealing with complex atomistic models, the success of a methodology is subject to inaccuracies in the energy models. Therefore, reliability is of a great concern here.

1.1 A hydrophobic-polar model for the prediction of protein structures

In 1985, Dill was the first person to propose this model based on the impact of hydrophobic interactions between amino acids on the manner of protein folding (Dill et al. 1993). Amino acids within proteins are divided into two classes: hydrophobic and hydrophilic. Folding proteins by the application of hydrophobic–hydrophilic is a very simple model that attempts to reflect the fundamental and general characteristics of protein and determines the structure of protein is space (Storm and Lyngsø 1999). In this model, the protein string is folded on a square two-dimensional network or on a cubic three-dimensional network. The folding process in hydrophobic-polar model (HP) has some behavioral similarities with the folding of the real protein systems (Dobson et al. 1998; De Araújo 1999).

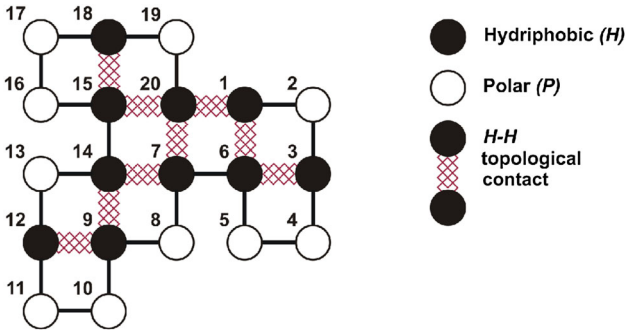

Fig. 1 The optimal structure for the sequence $HPHP_2H_2PHP_2HPH_2P_2HP$ in the 2D mode. Black and white nodes represent hydrophilic and hydrophobic residues, respectively . (Reproduced with the permission from Garza-Fabre et al. 2015)

This model is based on the fact that hydrophobic amino acids are not inclining so much to collide with molecules of water and mostly gather within then network. On the other hand, hydrophilic amino acids have a great inclination to make bonds with water molecules and are moved to the surface of proteins (Lau and Dill 1989). In this model, in order for the proteins to reach their most stable state and the level of energy gets its minimum, the hydrogen bonds between hydrophobic amino acids must be heightened. Hydrophobic amino acids are shown with letter H and hydrophilic amino acids are shown with letter P in this model. In order to obtain the energy level of hydrophobic pairs of amino acids that are not present in sequential protein strings but are in neighboring to each other within the network, the pairs are counted. Such pairs are inclined to establish a core and are shown in the form of (H–H). If these pairs are multiplied by a negative, the level of energy could be obtained. Other pairs present in the structure are shown in the form of H–P or P–P. Such pairs do not have any impact on the level of energy. For instance, if the sequence in the form of $r = r_1 r_2, \ldots, r_{20}$ portrays a protein, it could be shown in this model as $HPHP_2H_2PHP_2HPH_2P_2HP$. In Fig. 1, the manner of folding this sequence within a two-dimensional network is shown. This folding has the level of energy equal to -9 since the number of H–H bonds within it is equal to 9.

Despite simplicity of the model, obtaining an optimal structure on a cubic lattice is considered as an NP-hard problem (Crescenzi et al. 1998; Berger and Leighton 1998). Therefore, several metaheuristics have been implemented inspired from nature including differential evolution (Santos and Diéguez 2011), immune algorithms (Cutello et al. 2007), evolutionary algorithms (Mansour et al. 2010), particle swarm optimization (Kanj et al. 2009), and ant colony optimization (Shmygelska and Hoos 2005). The presented method in this paper is an evolutionary-based approach particularly the estimation of distribution algorithm (EDA). In EDAs, instead of using crossover and mutation operators,

new generations are created by the probability distribution of medial populations selected from the previous generations. Contrary to the genetic algorithms where maintenance of the building blocks over differing generations is implicit, in EDAs, it is done explicitly through joint probability distribution of the selected medial generation.

1.2 Literature review

Unger and Moult (1993) presented a hybrid method based on genetic algorithms (GA) and Monte Carlo (MC). The method simulates some folding pathway by searching for the global minimum state where genetic operators are repeated until a proper confirmation is attained. Another scheme based on standard GA was presented by Patton et al. (1995). In this approach, it is probable to create invalid conformations, and therefore, a penalty on the rate of collision is determined. Furthermore, Khimasia and Coveney (1997) proposed a method based on a simple genetic algorithm using elitism for the purpose of higher efficiency of the algorithm. In their method, invalid conformations were accepted through opening a penalty function. Findings of the tests (Patton et al. 1995; Yue et al. 1995) showed that this method has a similar performance with GA/MC on proteins having the length of 27, while it has better performance for proteins having the length of 64 and requires fewer stages than GA/MC in order to reach the final answer. Lima et al. (Custódio et al. 2014) presented another evolutionary algorithm based on HP lattice model. Their algorithm consisted of multipoint crossover, segment mutation, exhaustive search mutation, local move, and loop move. In that algorithm, the creation of invalid conformations was denied. Findings of this algorithm (based on comparing them to other similar works (Garza-Fabre et al. 2015; Patton et al. 1995; Yue et al. 1995) showed an improvement over GA/MC and SGA methods. Some other evolutionary algorithms could be used for structure prediction problem, e.g., multiobjective fitness function (Garza-Fabre et al. 2015) where each one of the constraints is considered as a single objective and it is attempted to enhance the results. Further, several non-evolutionary approaches have been applied to protein structure prediction. These methods follow heuristics for predicting structures. Heuristics may consist of cooperativity effects, existence of a hydrophobic core, etc. The HZ is based on the hypothesis that hydrophobic contacts are considered as constraints bringing other contacts into spatial proximity. Then, this constrains and zips up the next contacts more, and the process goes on. This algorithm is an attempt to collect hydrophobic contacts and create a structure having a hydrophobic core (Dill et al. 1993). CI (Toma and Toma 1996) applies the hypothesis implemented in HZ. The algorithm uses the Monte Carlo method for search procedure. Based on the results (Garza-Fabre et al. 2015), this algorithm is better and more efficient than GA or HZ algorithms. Spencer et al. (2015) proposed the first deep learning based PSP method, called DNSS. The method uses a deep belief network (DBN) based on restricted Boltzmann machine (RBM) and trained using contrastive divergence46 in an unsupervised manner. It also uses PSSM generated by PSI-BLAST to train deep learning network. Due to the existing difficulties in training of the deep learning network, large amount of training samples, and heavy calculations, they applied graphical processing units (GPU) and CUDA software to optimize the model. Kanj et al. (2009) presented another method based on the PSO algorithm. They turned invalid conformations to valid ones through a repairing process. Regarding the wide application of the methods, several approaches have been proposed including MA (Bazzoli and Tettamanzi 2004), ACO (Shmygelska and Hoos 2005; Do 2017), and DCSaDE-LS (Sudha et al. 2015) methods. Such methods make use of a local search for the enhancement of results. In addition, Tabu search (Liu et al. 2013), neural networks (Babaei et al. 2010), harmony search (Jana et al. 2017), and other forms of machine learning such as HMM (Lee et al. 2009), imperialist competitive (Khaji et al. 2016), fuzzy clustering (Shen et al. 2005), fuzzy support vector machine (Xie et al. 2018), and deep neural network (Wang et al. 2017) have been put forward.

In the current study, a method is presented based on estimation of distribution algorithms (AHEDA) by the application of HP model in order to predict the structure from the three-dimensional lattice. AHEDA makes use of local search for improving its efficiency and exploitation and efficiently searching for structures having minimal energy and higher HH contacts. The rest of this article has been made up of the following sections: Sect. 2 explains the manner of operation for the proposed algorithm. In Sect. 3, the algorithm is evaluated and compared with other methods, and Sect. 4 presents the conclusion of study.

2 Materials and methods

AHEDA is presented in detail in this section. A summary of the proposed method is represented in Fig. 2. In each iteration, the algorithm selects N chromosomes from M existing chromosomes using tournament selection and creates the intermediate population. Then, AHEDA applies a stochastic local search (SLS) on each selected chromosome with a calculated probability and, finally, updates probability vector of the intermediate population and reinitializes the population with new probability.

2.1 Structure representation

In the HP model, sites are located on a 3D space on a lattice. Representation of chromosomes in the form of x, y

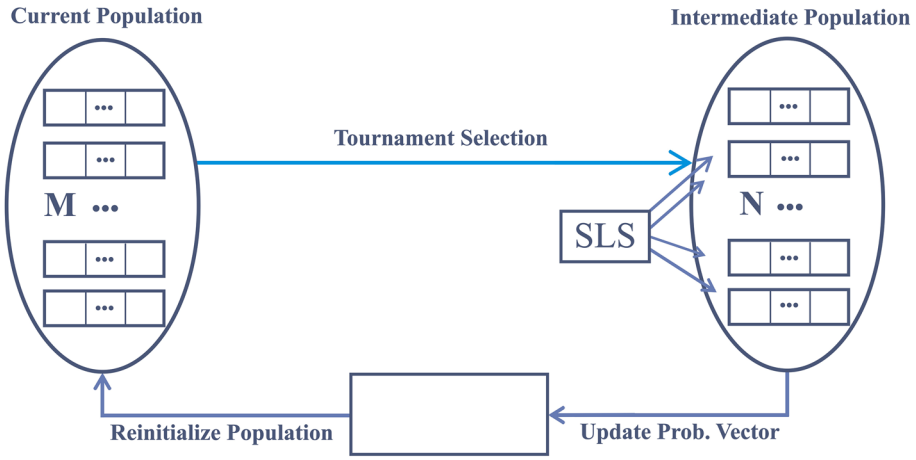

Fig. 2 The proposed method steps in a brief diagram

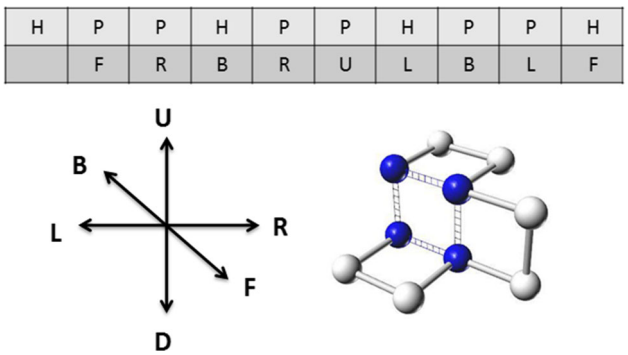

Fig. 3 Encoding scheme used in the genetic algorithm applied to the HP model in a cubic lattice. The sequence [HPPHPPHPPHPPH] consisting of ten monomers is encoded into a chromosome containing the absolute directions [FRBRULBLF]. Such chromosome translates in a 3D structure with four hydrophobic contacts. The H (P) monomers are the blue (white) beads. (Reproduced with the permission from Custódio et al. 2014)

Table 1 Bond directions describing lattice conformations

Direction	Δr
(U)P	$r_z = r_z + 1$
(L)EFT	$r_x = r_x - 1$
(F)RONT	$r_y = r_y + 1$
(B)ACK	$r_y = r_y - 1$
(R)IGHT	$r_x = r_x + 1$
(D)OWN	$r_z = r_z - 1$

A bond direction corresponds to a change, Δr, in one of the Cartesian coordinates of the successive monomer, keeping all other coordinates the same as the previous monomer (Khimasia and Coveney 1997)

and z coordinates is quite cumbersome for each site and needs a high level of statistical sophistication. Therefore, in order to represent chromosomes for an input sequence with n residues, a sequence of n-1 directions is created. In this sequence, the first amino acid is considered as the fundamental element of the structure without direction and located on (0, 0, 0) coordinates. Moreover, each residue in the sequence is encoded into one of the six directions {L–R–U–D–B–F} as shown in Fig. 3. Also, Table 1 shows the encoding scheme for each direction. In other words, for each one of {L–R–U–D–B–F} directions in time of appearance, changes happen in the coordinates of new sites according to Table 1 and coordinates of (x, y, z) are obtained. This encoding scheme has access to the whole structure on the lattice; however, looking for the best conformation among this lattice is difficult and needs a heuristic algorithm. Regarding the nature of the estimation of distribution algorithms that use a binary encoding scheme, three binary bits are considered for each one of n-1 conformations to cover six different directions. Therefore, with increasing population, a binary number is produced and the corresponding direction is included within the sequence.

2.2 Initial population and invalid conformations

Random key encoding is a solution that is used EDA in order to produce the initial population. The proposed method creates a sequence of sixfold direction in a random way for the first-generation chromosomes. It is important to mention that the algorithm has a random nature and there is a possibility for the production of invalid conformations in each stage. Invalid conformations are those structures with two or more amino acids located on a common site on the lattice. In other words, a conformation is valid only when each site is occupied by at most one residue. For each invalid conformation, the method defines a penalty corresponding to the rate of interference that is added to their fitness function. Naturally, such manipulation in the value of fitness function reduces the chance for selection of that structure.

2.3 Fitness function

As mentioned in the introduction section, a lower level of energy shows a closer structure of a protein to its natural form. The proposed method in this study uses the hydrophobic-polar (HP) model whereas multiplying the number of H bonds in -1 gives the energy level of the protein. This multiplication in the HP model is considered as the fitness function in the proposed method.

2.4 Selection strategy

In the first step, the algorithm selects a number of individuals based on the principles of fitness and variety in order to produce an intermediate population. These individuals will be used to update the probabilistic vector in each generation of evolutionary section. Then, a repetitive process of producing offspring is performed based on the produced population. Selection of this medial population, which is the processing population of the algorithm and increases its efficiency, is conducted by the selection operator called tournament. The tournament operator randomly selects k chromosomes from the current population and enters the best chromosome into the intermediate population. This operation is repeated n times, i.e., the size of medial population.

2.5 Probability distribution

The proposed method uses a univariate probability model that could be shown by a vector of possibilities with $3(n-1)$ elements (corresponding to the primary structure of a protein). What is meant by a univariate model is one in which no relationship exists between the variables (here, the directions). The vector of possibilities is initialized with 0.5 based on the Bernoulli distribution principles. For each iteration, the vector is updated according to the selected medial population. The model is factorized as a product of independent univariate marginal distributions. Below formula (Larranaga 2002) shows the update formula for the vector of possibilities.

$$p_l(x) = p\left(x|D_{l-1}^{S_e}\right) = \prod_{i=1}^{n} p_l(x_i)$$

Each univariate marginal distribution is estimated from marginal frequencies:

$$\prod_{i=1}^{n} p_l(x_i) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{N} \delta_j\left(X_i = x_i|D_{l-1}^{S_e}\right)}{N}$$

Table 2 The selected individuals, $D_0^{S_e}$, from the initial population and their joint probability

X_1	X_2	X_3
0	1	0
1	1	1
0	0	1
$p\left(X_1 = 1|D_0^{S_e}\right) = \frac{1}{3}$	$p\left(X_2 = 1|D_0^{S_e}\right) = \frac{2}{3}$	$p\left(X_3 = 1|D_0^{S_e}\right) = \frac{2}{3}$

For example, assume that a number of individuals are selected from D_0 (initial population) using a selection method. $D_0^{S_e}$ denotes the data file containing the selected individuals. The selected individuals have been characterized by means of the joint probability distribution (see Table 2). Thus, only three parameters are required to specify the model. Each parameter, $p\left(x_i|D_0^{S_e}\right)\ i = 1, 2, 3$, will be estimated from the data file $D_0^{S_e}$ by means of its corresponding relative frequency, $p\left(x_i = 1|D_0^{S_e}\right)$.

2.6 The stochastic local search

Population-based algorithms are programmed to direct the population toward overcoming limitations with two features: exploration and exploitation. Regarding the selection of new values for the vector of possibilities, exploration is in the nature of EDA. However, these algorithms are weak due to their exploitation property and are not efficient in the local search for chromosomes. To overcome this problem, the algorithms benefit from a local search in order to create exploitation for finding efficient solutions. In local search, after selection of n chromosomes and creation of an intermediate population, it searches for the neighboring chromosomes with the probability of *wp*. The *wp* parameter provides randomness in the algorithm and reduces the processing time. Without this parameter, the algorithm needs to analyze and investigate each chromosome in the peripheral regions that is a time-consuming procedure regarding the nature of EDA. For the case of this study, the *wp* parameter was set to 0.01. For a better neighboring chromosome, it is replaced in the medial population and the vector of possibilities is updated. Algorithm 1 shows the pseudo-code implemented for local search where *maxiter* is the size of intermediate population with a size less than M and *PSP instance* is protein structure in form of direction.

Algorithm 1 Stochastic Local Search

Require: *a PSP instance, maxiter, wp*
Ensure: *a neighbour PSP solution*
for *i = 1* **to** *maxiter* **do**
  r = random number between 0 and 1
  if *r < wp* **then**
    if *Solution has collision* **then**
      Repair collision
    else
      Change some direction from chromosome
    end if
  end if
end for
return *the neighbour solution*

2.7 Replacement strategy

After updating the vector of possibilities, a new population is created using a new vector and new chromosomes are copied over the chromosomes of the previous generation. In each iteration, the best chromosomes of the previous generation are selected based on the fitness function. The new chromosomes are copied over the old ones if their fitness function value is lower than the previous generation. Otherwise, the best chromosomes from the previous generation are transferred to a new population. In other words, the algorithm makes use of elitism and it is expected that elitism in the algorithm be shown in the Fitness diagram.

2.8 Pseudo-code of the AHEDA

Concerning the proposed EDA in Algorithm 2, it could be observed that the algorithm attempts to explore the various parts in the search space using a stochastic local search (SLS). The employed selection strategy is a potential factor in the enhancement of the algorithm in the areas such as searching the whole search space and moving away from the local optimal.

As it can be observed in the pseudo-code of the algorithm, the probability value is initially set to 0.5. In the iterative phase, the first set of M individuals is produced based on the values in the probability vector according to random encoding. Then, a subset of N individuals is selected from the previous set and created an intermediate population and random local search is performed on each selected individuals in the intermediate population. In each stage of the local search, V individual ($V \leq N$) selected if the neighboring individual leads to an enhancement of the fitness function, it is replaced in the individual from the intermediate population. Finally, the algorithm attempts to update the probability curve in the last repetition. The estimation of distribution procedure is run until the termination condition is satisfied. A summary of the algorithm is shown in Algorithm 2.

3 Results

The proposed EDA has been implemented within Visual Studio 2008 and C# programming language in a personal computer having a 2.5 GHz dual-core processor and a 3 GB RAM. To examine the direction and correctness of the proposed method and present its efficiency and strength, we arranged a set of assessments that are represented in this section. Generally, performance of the evolutionary solutions developed for optimization problems is verified using certain types of assessment methods. Accordingly, a number of these standard assessment tools have been applied to analyze and examine the proposed solution. Furthermore, the performance of the proposed algorithm was assessed using different sets of benchmarks.

Algorithm 2 AHEDA for PSP

Require: *a protein primary structure*
Ensure: *a sequence of direction that minimize energy*
Initialize the probability vector $P_0(x)$
  $P_0(x) = P_0(x_1, x_2, \ldots, x_{3(n-1)}) = (P_0(x_1), P_0(x_2), \ldots, P_0(x_{3(n-1)})) = (0.5, 0.5, \ldots, 0.5)$
while *Termination condition satisfied* **do**
  Using P(x) obtain M individual $x_1, x_2, \ldots, x_m$ *according to the RK encoding*
  Select the N ($N \leq M$*) according to the selection strategy from M*
  repeat
    Select an individual V from N
    Apply SLS on V
    if *Fitness (V) improved* **then**
      Replace improved V in N according to the V position
    end if
  until *All the N individuals are examined*
  Update the probability vector P(x) using the N individual
end while
return *the best individual solution found*

3.1 Convergence analysis

Convergence analysis involves analyzing convergence circumstance of AHEDA to an optimal solution. Figures 4 and 5 illustrate how AHEDA converges to an optimal solution of six selected test cases that have been selected from (Garza-Fabre et al. 2015; Patton et al. 1995; Yue et al. 1995). In this figures two criterions are considered: BestFitness which means best fitness value in generation and Fitness which means average of fitness value in the generation. Blue curve showed average fitness of generation, and black steps curve is related to best fitness value in the generation. Also d48.2 means second version of some peptide of length 48 amino acids. (All of these peptides with different lengths are presented in "Appendix" section.) As shown in the figures, fitness curve of the AHEDA is descending

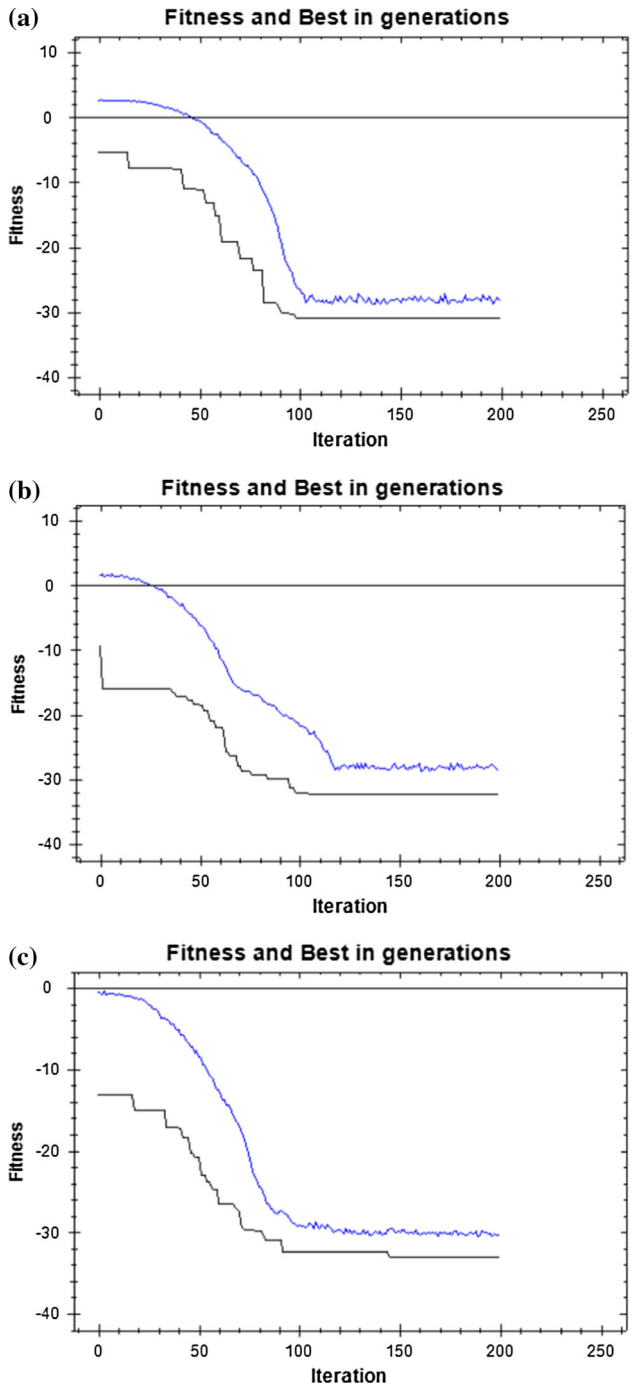

Fig. 4 AHEDA convergence circumstance to an optimal solution for **a** d64.9, **b** d48.2, **c** d48.4. The blue curve shows the average fitness of generation, and the black steps curve represents the best fitness value in the generation (color figure online)

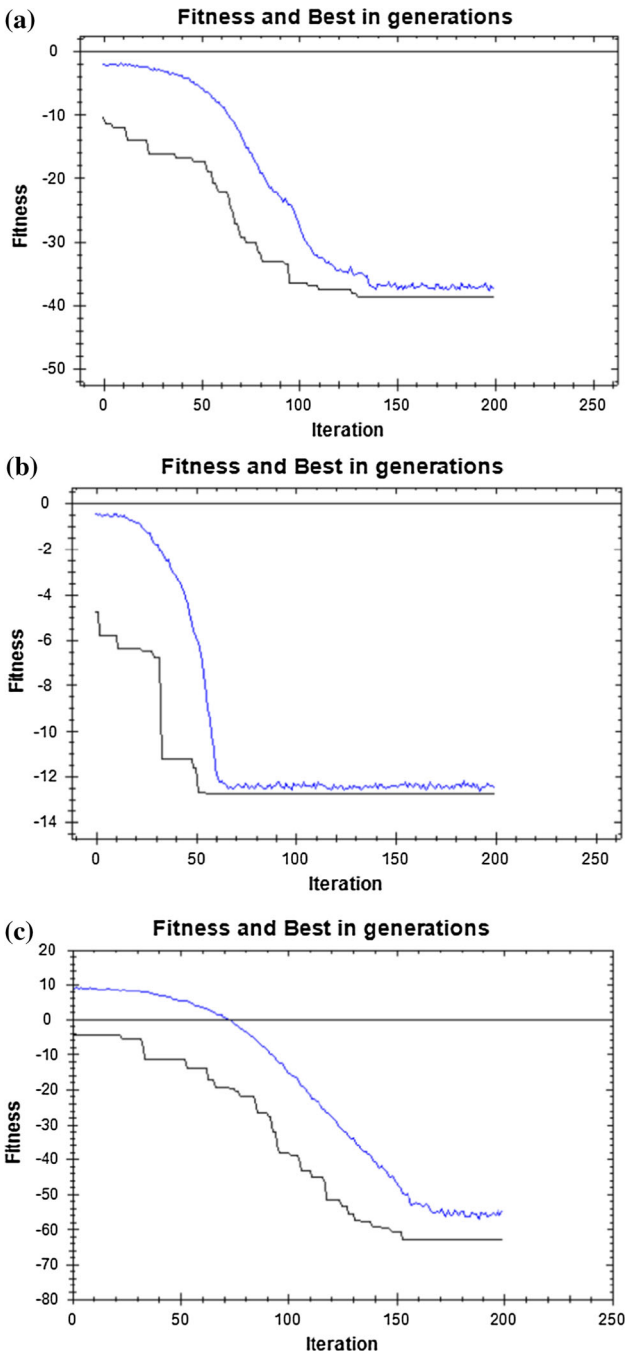

Fig. 5 AHEDA convergence circumstance to an optimal solution for **a** d64.4, **b** d24, **c** d124. The blue curve shows the average fitness of generation, and the black steps curve represents the best fitness value in the generation (color figure online)

and this is consistent with the protein structure prediction problems which are a minimization task. It is expected that elite curve be stairs and descending in this kind of problems where AHEDA also produces such curves for selected benchmarks.

3.2 Stability analysis

Usually, statistical procedures such as *t* test or the Wilcoxon test are applied for the purpose of analyzing the stability of evolutionary algorithms. The one-sample *t* test is applied when one intends to make a comparison between the mean

Table 3 Normality test for 6 different instances in 50 independent executions using SPSS

	Kolmogorov–Smirnov			Shapiro–Wilk		
	Statistic	*df*	Sig.	Statistic	*df*	Sig.
d48.2	0.245	50	0.000	0.854	50	0.000
d48.4	0.210	50	0.000	0.846	50	0.000
d64.4	0.197	50	0.000	0.895	50	0.000
d64.9	0.183	50	0.000	0.902	50	0.001
d24	0.233	50	0.000	0.802	50	0.000
d124	0.159	50	0.003	0.930	50	0.006

score of a sample to the known value (usually the population means or the average for the outcome of some population of interest). The test is primarily means that a comparison is made between the average of the sample (the observed average) and the population (the expected average). In addition, an adjustment is made for the number of cases in the sample and the standard deviation of the average. In this test, the null hypothesis is that the difference between the observed mean of AHEDA results and the expected mean of optimum value equals zero. On the other hand, the alternative hypothesis states that the difference between the observed mean of AHEDA and the expected mean of optimum does not equal zero. The p value was set at $p < 0.05$, indicating that the conditions for the rejection of the null hypothesis are met.

In case that the findings of multiple operations related to an evolutionary algorithm on a certain measure do are in the normal distribution, the examination of its stability should make use of the t test. If that was not the case, the Wilcoxon test should be applied for the purpose of investigating the stability of the evolutionary solution.

The results of Kolmogorov–Smirnov and Shapiro–Wilk tests for the AHEDA are presented in Table 3. The p value (Sig.) represents the dissimilarity of the sample of results regarding the normal shape. Thus, lower p values ($p < 0.05$) indicate the existence of non-normal distributions. According to Table 3, it can be concluded that the findings are not in a normal distribution and the Wilcoxon signed-rank test must be applied.

Both Wilcoxon signed-rank and the rank-sum tests for the comparison of two independent samples were proposed by Frank Wilcoxon. In his famous book on nonparametric statistics, Siegel popularized these tests. As a nonparametric statistical hypothesis test, the Wilcoxon signed-rank test is used for the comparison pf two related samples, matched samples, or repeated measurements on a simple sample. The test is used to assess if their population mean ranks differ or not. In other words, it is a paired-difference test (Ramezani and Lotfi 2013).

The null hypothesis tests if the difference ($z = x - y$) between the members of each pair (x, y) has the median value equal to zero ($H_0 : \theta = 0$). To be more precise, x and y have similar distributions and the alternative hypothesis tests whether $H_1 : \theta$ is not equal to 0. Table 4 provides the results of the Wilcoxon signed-rank test used for the purpose of stability investigation. For each benchmark, if the level of significance is lower than 0.05 ($p < 0.05$), then the proposed method is not stable. Nevertheless, the results provided in Table 4 indicate that the AHEDA is stable.

Table 5 Benchmark sets with different length used for comparative study

The set number	Number of proteins	Length
1 (Yue et al. 1995)	10	48
2 (Patton et al. 1995)	9	64
3 (Garza-Fabre et al. 2015)	5	Different length

3.3 Comparative study of the structure prediction methods

To analyze performance of the proposed algorithm for protein structure prediction, various benchmark sets were used. Each set contains proteins with almost the same length; however, different sets include proteins with different peptide length. Table 5 shows a summary of the benchmark sets in our tests. The common feature between all these benchmarks is difficulty in the prediction of their 3D structure because of the existence of hydrophobic amino acids in a way to form a dense hydrophobic core and high probability of being stuck in a local optimal. The algorithm was executed 50 times independently to choose the best and average results. We compared the results of our proposed algorithm with other evolutionary-based methods including GAHP (Custódio et al. 2014), HZ (Dill et al. 1993), SGA (Khimasia and Coveney 1997), and CI (Toma and Toma 1996).

Table 6 shows the results obtained by AHEDA and five other methods applied on the set number 1 from Table 4 having length of 48 residues. Based on the results in Table 6,

Table 4 Wilcoxon signed-rank test statistics in 50 independent executions using SPSS

	d48.2_2 – d48.2	d48.4_2 – d48.4	d64.4_2 – d64.4	d64.9_2 – d64.9	d24_2 – d24	d124_2 – d124
Z	− 0.654	− 0.1018	− 0.133	− .1160	− 0.1292	− 0.1323
Asymp. Sig. (2 tailed)	0.519	0.309	0.894	0.246	0.196	0.186

Table 6 Results obtained by different methods on benchmark set number 1

Benchmark	#HH Contacts						
	AHEDA	AHEDA[a]	CI	SGA	HZ	GAHP	GAHP[a]
48.1	31	29.48	32	24	31	32	30.72
48.2	34	32.28	33	24	32	34	31.26
48.3	34	32.38	32	23	31	34	32.08
48.4	33	31.44	32	24	30	33	31.16
48.5	31	29.76	30	28	30	32	30.52
48.6	32	30.1	30	25	29	32	29.86
48.7	32	30.28	30	27	29	32	29.82
48.8	30	28.48	30	26	29	31	29.32
48.9	34	32.3	34	27	31	34	31.92
48.10	31	29.46	33	26	33	33	31.08

The results were taken from Custódio et al. (2014) except for AHEDA, and X[a] show the average of method X from 50 independent executions

Table 7 Results obtained by different methods on benchmark set number 2

Benchmark	#HH Contacts								
	AHEDA	AHEDA[a]	GA/MC	SGA1	SGA2	PSO	GA	GAHP	GAHP[a]
64.1	32	30.44	21	27	27	28	28	31	28.5
64.2	37	34.84	26	30	29	31	32	36	33.18
64.3	40	37.88	30	34	34	36	35	39	36.02
64.4	40	38	28	36	32	38	36	40	37.96
64.5	35	32.92	22	31	29	31	31	33	31.52
64.6	26	24.04	17	25	20	27	25	28	26.7
64.7	35	32.96	28	34	29	35	35	36	33.72
64.8	38	36.32	29	33	32	35	34	38	36.32
64.9	32	29.98	20	26	24	27	27	31	28.9

The results were taken from Custódio et al. (2014) except for AHEDA, and X[a] show the average of method X from 50 independent executions

AHEDA produced better energies (a large number of HH contacts or the same number of HH with better average) than SGA method for all ten sequences and also better than GAHP for six sequences. Among other construction-based methods, AHEDA found better energies than those reported by CI method for six sequences and those generated by the HZ method for eight sequences. Furthermore, the average energies from 50 independent executions were near to the global minimum energy showing a difference less than three HH contacts.

Furthermore, Table 7 represents the results obtained by different methods on benchmark set number 2 with length of 64. As it is evident from the table, the proposed algorithm is able to produce better energy levels in comparison with the other evolutionary-based methods. AHEDA found better energy levels than GA/MC, SGA1 and SGA2 methods. In addition, its results are better for 7 sequences compared to PSO and GAHP methods and also are better for 8 sequences compared to the GA method.

Additionally, the benchmark set number 3 including sequences with varying length has been used for better evaluation of the methods. According to the results in Table 8, the global optimal is known for the first three sequences and AHEDA is able to reach the global optimal through the best average than the other methods. For two last sequences with a longer length, AHEDA produces an average of one unit enhancement.

Table 8 Results obtained by different methods on benchmark set number 3

Benchmark	#HH Contacts				
	AHEDA	AHEDA[a]	CI (Toma and Toma 1996)	GAHP (Custódio et al. 2014)	GAHP[a]
20	11	10.2	9	–	–
24	13	12.12	9	–	–
25	9	8	8	–	–
103	49	46.04	49	50	46.58
124	65	62.78	58	63	58.12

The results were taken from Custódio et al. (2014) except for AHEDA, and X[a] show the average of method X from 50 independent executions

4 Discussion and conclusion

The main objective in this work was to develop a novel method for protein structure prediction based on the estimation of distribution algorithm. The proposed method uses an integrated EDA to optimize the search process and prevent the destruction of structural blocks. The scheme is relevant to discovery of new sites; however, it obtains a lower performance in local search and looking for neighboring sites. Accordingly, the method uses an additional stochastic local search (SLS) algorithm to increase the accuracy of the prediction. Most of the methods presented in Tables 6, 7 and 8 use the HP model for prediction of a protein structure, and therefore, comparing the proposed algorithm with these methods can be a reasonable criterion. The results obtained from experimental studies indicate the capability of the proposed scheme in precisely prediction of protein structure.

As it is mentioned above, the more the number of HH bonds in the predicted structure, the lower would be the value of fitness function. Furthermore, the lower the value of fitness function, the lower would be the energy level of the structure, and as a result, it will resemble the actual and natural structure of the target protein. The major characteristic of the proposed method in comparison with other structure prediction techniques is its ability to predict a structure very close to the optimal ones through a repetitive search procedure. Therefore, the average accuracy of the predicted structure is higher than other proposed techniques. Compared to the other methods, AHEDA is capable of producing structures with a higher number of HH bonds with a high degree of reliability. Though AHEDA and GA are relatively from the same family, AHEDA obtains a better accuracy than GA. This is due to the fact that AHEDA uses the SLS technique, whereas the faulty features of GA have been eliminated. As mentioned before, the major shortcoming of GA is lack of maintaining building blocks. Not only the AHEDA method overcomes this shortcoming, but also it accommodates a local search technique in its inside. These changes enable the method to produce better results in the exploitation stage and avoid the shortcoming present in the EDA method.

Additionally, brief look at the results presented in Tables 6, 7 and 8 shows that the method provides totally a better average of the energy level for the benchmark sets with different length compared to the other methods. Based on a comprehensive comparison with other existing methods on 24 widely used benchmarks, the proposed method produces highly accurate predictions. We showed that AHEDA finds the optimal solution for proteins with length less than 48 and produces highly accurate structure with higher HH contacts than other approaches for benchmark sets with length 48, 64 and bigger than 100 residues.

The main goal in protein structure prediction is to find an inherent 3D structure for a given amino acid sequence. In this paper, we described a memetic estimation of distribution algorithm, AHEDA, based on the 3D HP model. As shown in a series of recent publications in demonstrating new approaches (Chen et al. 2016), user-friendly and publicly accessible web servers will significantly enhance their impacts, and we shall make efforts in our future work to provide a web server for the new approach presented in this paper. In future studies, we plan to use machine learning and evolutionary approaches such as fuzzy evolutionary and social-based algorithms and merge them to improve the prediction results.

Compliance with ethical standards

Conflict of interest Authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Appendix: TEST DATA

See Tables 9, 10, and 11.

Table 9 48 Peptide length test cases. (Reproduced with the permission from Yue et al. 1995)

Case#	Sequence
d48.1	HPHHPPHHHHPHHHPPHHPPHPHHHPHPHHPPHHPPPHPPPPPPPPHH
d48.2	HHHHPHHPHHHHHPPHPPHHPPHPPPPPPHPPHPPPHPPHHPPHHHPH
d48.3	PHPHHPHHHHHHPPHPHPPHPHHPHPHPPPHPPHHPPHHPPHPHPHPHP
d48.4	PHPHHPPHPHHHPPHHPHHPPPHHHHHPPHPHHPHPHPPPPHPPHPHP
d48.5	PPHPPPHPHHHHPPHHHHPHHPHHHPPHPHPHPPHPPPPPPHHPHHPH
d48.6	HHHPPPHHPHPHHPHHPHHPHPPPPPPPHPHPPHPPPHPPHHHHHHPH
d48.7	PHPPPPHPHHHPHPHHHHPHHPHHPPPHPHPPPHHHPPHHPPHHPPPH
d48.8	PHHPHHHPHHHHPPHHHPPPPPHPHHPPHHPHPPPPHHPHPHPHHPPP
d48.9	PHPHPPPPHPHPHPPHPHHHHHHPPHHHPHPPHPHHPPHPHHHPPPPH
d48.10	PHHPPPPPPHHPPPHHHPHPPHPHHPPHPPHPPHHPPHHHHHHPPHH

Table 10 64 peptide length test cases. (Reproduced with the permission from Patton et al. 1995)

Case#	Sequence
d64.1	$P_2H_5P_3H_2P_5H_2P_3HP_6HPHP_3HP_2HP_2HP_5HP_4H_2PH_2P_2HP_2HP$
d64.2	$P_2HPHP_2HP_2H_3PH_4P_2H_3P_4HPHP_3HPHP_3HPHP_5HPHP_2HPHP_3HP_2HP_2$
d64.3	$HP_2H_2P_2HP_2HPHP_2HP_4HP_6HPHPH_3P_2HPHP_3HPHP_2H_2P_2HP_2HP_2HPH_3PH$
d64.4	$HP_3H_2P_2HPHP_3HP_3HPH_2P_3H_2PHPH_3PHP_2HP_3HP_2HPH_3P_2HP_2HP_2H_3PH_4$
d64.5	$HP_2H_2PH_4P_6H_2P_2HP_4H_2P_3HP_2HPH_2PHP_4H_2P_4HP_5HP_4HPH_2$
d64.6	$P_4HP_3HP_3H_4PH_2P_5HP_2HPH_2PHPHP_5HP_{10}H_4P_4H_2P_2H$
d64.7	$P_3H_3P_2HPHP_2HP_2H_2P_3HP_2HP_2H_2PHP_3HP_7HPH_3PH_5P_2H_2P_3HP_2H$
d64.8	$HP_2HP_2H_3P_4HPHP_3HPH_2PH_5P_4HPHPHP_4HPHP_3H_2PHP_4HP_2H_2PHP$
d64.9	$P_2HP_2HP_2H_3P_3HPHP_2HP_2HP_6HP_2H_3P_2HP_2HP_2HPHP_6H_3P_5HPHP$

Table 11 Different length test cases. (Reproduced with the permission from Garza-Fabre et al. 2015)

Case#	Sequence
d20	$HPHP_2H_2PHP_2HPH_2P_2HPH$
d24	$H_2P_2HP_2HP_2HP_2HP_2HP_2HP_2H_2$
d25	$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$
d103	$P_2H_2P_5H_2P_2H_2PHP_2HP_7HP_3H_2PH_2P_6HP_2HP\ HP_2HP_5H_3P_4H_2PH_2P_5H_2P_4H_4PHP_8H_5P_2HP_2$
d124	$P_3H_3PHP_4HP_5H_2P_4H_2P_2H_2(P_4H)_2P_2HP_2H_2P_3H_2PHPH_3P_4H_3P_6H_2P_2HP_2HPHP_2HP_7HP_2H_3P_4HP_3H_5P_4H_2(PH)_4$

References

Anfinsen C (1973) Principles that govern the folding of protein chains. Science 181(96):223–230

Babaei S, Geranmayeh A, Seyyedsalehi SA (2010) Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks. Comput Methods Programs Biomed 100(3):237–247

Bazzoli A, Tettamanzi AG (2004) A memetic algorithm for protein structure prediction in a 3D-lattice HP model. Workshops on applications of evolutionary computation. Springer, Berlin, pp 1–10

Berger B, Leighton T (1998) Protein folding in the hydrophobic–hydrophilic (HP) model is NP-complete. J Comput Biol 5(1):27–40

Bujnicki JM (2006) Protein-structure prediction by recombination of fragments. ChemBioChem 7(1):19–27

Chen W, Ding H, Feng P, Lin H, Chou KC (2016) iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget 7(13):16895

Crescenzi P, Goldman D, Papadimitriou C, Piccolboni A, Yannakakis M (1998) On the complexity of protein folding. J Comput Biol 5(3):423–465

Custódio FL, Barbosa HJ, Dardenne LE (2014) A multiple minima genetic algorithm for protein structure prediction. Appl Soft Comput 15:88–99

Cutello V, Nicosia G, Pavone M, Timmis J (2007) An immune algorithm for protein structure prediction on lattice models. IEEE Trans Evol Comput 11(1):101–117

Davis IW, Baker D (2009) RosettaLigand docking with full ligand and receptor flexibility. J Mol Biol 385(2):381–392

De Araújo AFP (1999) Folding protein models with a simple hydrophobic energy function: the fundamental importance of monomer inside/outside segregation. Proc Nat Acad Sci 96(22):12482–12487

Lau KF, Dill KA (1989) A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 22(10):3986–3997

Dill KA, Fiebig KM, Chan HS (1993) Cooperativity in protein-folding kinetics. Proc Nat Acad Sci 90(5):1942–1946

Dobson CM, Šali A, Karplus M (1998) Protein folding: a perspective from theory and experiment. Angew Chem Int Ed 37(7):868–893

Do DD (2017) A novel and efficient ant colony optimization algorithm for protein 3D structure prediction. VNU-UET technical report

Garza-Fabre M, Rodriguez-Tello E, Toscano-Pulido G (2015) Constraint-handling through multi-objective optimization: the hydrophobic-polar model for protein structure prediction. Comput Oper Res 53:128–153

Guntert P (2004) Automated NMR structure calculation with cyana. Protein NMR Tech 278:353–378

Jana ND, Sil J, Das S (2017) An improved harmony search algorithm for protein structure prediction using 3D off-lattice model. International conference on harmony search algorithm. Springer, Singapore, pp 304–314

Kanj F, Mansour N, Khachfe H, Abu-Khzam F (2009) Protein structure prediction in the 3D HP model. In IEEE/ACS international conference on computer systems and applications, 2009. AICCSA 2009. IEEE. pp 732–736

Khaji E, Karami M, Garkani-Nejad Z (2016) 3D protein structure prediction using Imperialist Competitive algorithm and half sphere exposure prediction. J Theor Biol 391:81–87

Khimasia MM, Coveney PV (1997) Protein structure prediction as a hard optimization problem: the genetic algorithm approach. Mol Simul 19(4):205–226

Larranaga P (2002) A review on estimation of distribution algorithms. Estimation of distribution algorithms. Springer, New York, pp 57–100

Lee SY, Lee JY, Jung KS, Ryu KH (2009) A 9-state hidden Markov model using protein secondary structure information for protein fold recognition. Comput Biol Med 39(6):527–534

Liu J, Sun Y, Li G, Song B, Huang W (2013) Heuristic-based tabu search algorithm for folding two-dimensional AB off-lattice model proteins. Comput Biol Chem 47:142–148

Mansour N, Kanj F, Khachfe H (2010) Evolutionary algorithm for protein structure prediction. In: 2010 sixth international conference on natural computation (ICNC), vol 8. IEEE, pp 3974–3977

Patton AL, Punch III WF, Goodman ED (1995) A standard GA approach to native protein conformation prediction. In: ICGA, pp 574–581

Raman S, Huang YJ, Mao B, Rossi P, Aramini JM, Liu G, Montelione GT, Baker D (2010) Accurate automated protein NMR structure determination using unassigned NOESY data. J Am Chem Soc 132(1):202–207

Ramezani F, Lotfi S (2013) Social-based algorithm (SBA). Appl Soft Comput 13(5):2837–2856

Razmara J, Deris SB, Parvizpour S (2013) A rapid protein structure alignment algorithm based on a text modeling technique. Bioinformation 6(9):344

Santos J, Diéguez M (2011) Differential evolution for protein structure prediction using the HP model. International work-conference on the interplay between natural and artificial computation. Springer, Berlin, pp 323–333

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334(2):577–581

Shen Y, Vernon R, Baker D, Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43(2):63–78

Shmygelska A, Hoos HH (2005) An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC Bioinform 6(1):30

Spencer M, Eickholt J, Cheng J (2015) A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinf 12(1):103–112

Storm CN, Lyngsø RB (1999) Protein folding in the 2D HP model. Tech rep, Technical Report RS-99-16 BRICS, University of Aarhus, Denmark

Sudha S, Baskar S, Amali SMJ, Krishnaswamy S (2015) Protein structure prediction using diversity controlled self-adaptive differential evolution with local search. Soft Comput 19(6):1635–1646

Toma L, Toma S (1996) Contact interactions method: a new algorithm for protein folding simulations. Protein Sci 5(1):147–153

Unger R, Moult J (1993) Genetic algorithms for protein folding simulations. J Mol Biol 231(1):75–81

Wang Y, Mao H, Yi Z (2017) Protein secondary structure prediction by using deep learning method. Knowl-Based Syst 118:115–123

Xie S, Li Z, Hu H (2018) Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization. Gene 642:74–83

Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA (1995) A test of lattice protein folding algorithms. Proc Nat Acad Sci 92(1):325–329