Scatter Search for Homology Modeling

Mouses Stamboulian and Nashat Mansour(✉)

Department of Computer Science and Mathematics,
Lebanese American University, Beirut, Lebanon
{mouses.stamboulian,nmansour}@lau.edu.lb

Abstract. Homology modeling is an effective technique in protein structure prediction (PSP). However this technique suffers from poor initial target-template alignments. To improve homology based PSP, we propose a scatter search (SS) metaheuristic algorithm. SS is an evolutionary approach that is based on a population of candidate solutions. These candidates undergo evolutionary operations that combine search intensification and diversification over a number of iterations. The metaheuristic optimizes the initial poor alignments and uses fitness functions. We assess our algorithm on a number of proteins whose structures are present in the Protein Data Bank and which have been used in previous literature. Results obtained by our SS algorithm are compared with other approaches. The 3D models predicted by our algorithm show improved root mean standard deviations with respect to the native structures.

Keywords: Comparative modeling · Homology modeling · Protein structure prediction · Scatter search · Metaheuristics

1 Introduction

Proteins are considered to be large biological molecules that are composed of a specific sequence of amino acids (AA). A protein is distinguished from another by the sequence of AAs. These sequences of AAs fold and take three dimensional shapes (tertiary structure) forming complex structures of proteins. The function of a protein is decided by the structure of the protein and the way it folds after it gets transcribed [1]. Many diseases in humans result from misfolded proteins; examples are Cystic Fibrosis and Parkinson's disease [2].

The protein Structure Prediction (PSP) problem refers to determining the 3D conformation of proteins, given the initial sequence of AAs. Knowing their structure allows designing drugs and implementing personalized medicine. Next generation sequencing technologies have resulted in a huge explosion in genomic data and in the number of protein sequences. This has led to a growing gap between protein sequences and the structures discovered so far, thus encouraging the development of faster and more effective methods. Accurate wet lab methods exist for the PSP. These techniques are very time consuming and could still be error prone. Therefore, computational methods for PSP constitute a significant alternative.

Computational determination of a protein's 3D conformation is an intractable problem [3]. Hence, heuristic approaches are needed and have been developed to tackle

Y. Tan et al. (Eds.): ICSI 2016, Part I, LNCS 9712, pp. 66–73, 2016.
DOI: 10.1007/978-3-319-41000-5_7

this problem. Computational methods could be categorized into three groups. Ab-Initio methods predict protein structures using only their sequence of AAs, guided by energy functions [4]. Protein Threading, or fold recognition, assumes a limited number of distinct protein folds and uses fold libraries to map protein folds to sequences [5]. Homology modeling is highly based on previous knowledge. It determines the structure of a target protein by using template protein structures [6]. Homology modeling predicts accurate models, provided that suitable template structures exist for the prediction. The rationale for this approach is that protein structures are more conserved than their respective sequences during evolution [7]. Homology protein modeling takes advantage of this fact and aims to predict the structure of a certain target sequence using 3D structures of known homologous proteins. Whenever the sequence similarity falls below 25 %, homology modeling suffers from serious misalignments, resulting in poor comparative models. Three heuristic approaches have been reported for homology modeling: the Genetic Algorithm (GA) [8], Tabu Search (TS), and Particle Swarm Optimization (PSO) [9].

In this paper we focus on homology modeling for predicting 3D protein structures. We use, for the first time, the Scatter Search (SS) metaheuristic to explore the search space for new and better target-template alignments. SS is used within the framework of homology modeling by satisfying spatial restraints [10]. Experiments are conducted using our proposed approach and the results are compared with results from previous literature.

2 SS on Homology Modeling

2.1 Preliminaries

In general, the steps for homology modeling start with a search for appropriate template (s) (given the target sequence), aligning the target with the discovered templates, proposing a 3D structure for the target sequence based on this alignment, and finally evaluating the predicted model's accuracy. To perform the first step, online search tools are used to search for templates in databases of known protein structures. Popular searching tools are BLAST [11] and FASTA [12]. After choosing appropriate template structures, the next step is to perform pairwise or multiple sequence alignment between the target and the template(s) [13]. After aligning the sequences, the actual building of the model is done. One common technique used to build a model is 'modeling by satisfying spatial restraints'. Based on the target-template alignment, the latter technique constrains the possible structure for the target protein based on the restraints extracted from the template structures taking into account their sequence similarity [6]. Stereochemical restraints are added to the extracted constraints, such as the lengths of the bonds, their angles and molecular mechanics. Information from the two sources are combined together into an objective function. The most probable comparative model is suggested by optimizing this objective function.

The fitness functions used to guide the SS metaheuristic is the DOPE score [14] and MODELLER [10] is employed to develop 3D models. DOPE stands for Discrete

Optimized Protein Energy. It aims to obtain a global optimum value of a scoring system, which would be a direct indication of a native structure for that particular protein sequence. This is achieved through using Joint Probability Density Function of the positions, in 3D space, for all the atoms present in the protein molecule. Details are found in [14].

MODELLER is an automated comparative modeling tool for protein structure prediction, whose aim is to find the most likely 3D conformation for a given target sequence of amino acids. This process is initiated through the alignment of the target sequence with at least one or more template sequence(s) of known structures. The 3D comparative model of a sequence X of unknown structure is predicted by comparing it with structure(s) of one or more close homologues. If there are more than one template structure, then these structures are first compared with each other and spatial features get extracted from them. After that, the extracted features are sent to the target sequence, and hence a group of spatial restraints about the structure to-be- predicted is obtained. The final predicted 3D model is optimized by maximizing the spatial restraint satisfaction as much as possible. Details are found in [10].

Scatter Search (SS) is a population based evolutionary search strategy and has been adapted for solving a number of intractable optimization problems [15]. Starting with an initial random or controlled-random population, SS maintains a small population set, the reference-set; then, the algorithm combines' solutions and updates the reference-set. The basic steps representing the template for Scatter Search are the following:

- *Diversification-Generation Method:* an initial population of candidate solutions are created completely randomly or with controlled sampling;
- *Solution-Improvement-Method;*
- *Reference-Set-Update-Method:* a small group of candidate solutions are maintained in the 'reference set'. The best-quality and most-diverse solutions obtained after the improvement phase get admitted to the reference set, hence allowing SS to combine both properties of diversification and intensification;
- *Subset-generation-method:* subsets of candidate solutions of defined sizes are created;
- *Solution-Combination-Method:* the subsets of candidate solutions get combined to create new solutions.

In the next sections, we explain how the SS metaheuristic was adapted for the PSP problem.

2.2 Solution Representation

Candidate solutions are represented by objects that refer to the target protein. For practical purposes, they also contain the sequence for the template proteins. These sequences are represented by arrays composed of single letter amino acid representations and gaps, which are manipulated by the SS algorithm.

2.3 Diversification Generation Method (DGM)

This method creates an initial random population of candidate solutions, and enables SS to explore wider ranges of the solution space. This method takes an input which is an initial target-template alignment. The initial alignment is generated using a dynamic programing algorithm for local sequence alignment, with affine gap penalties. Based on this initial alignment, randomly generated alignments are created whose number is specified by *PSIZE* (population size), which is set to 100. In generating the initial population, the following rules are upheld: The length of the initial seed target-template alignment is respected; the number of gaps within the sequence alignments (both in the template and target) remain the same; the order of residues in which they appear in the initial seed alignment is respected. Restricting the process to these rules ensures the generation of feasible solutions. The DGM is called once at the beginning of the algorithm.

2.4 Solution Improvement Method (SIM)

This method aims to improve the quality of the solutions produced by DGM or by the solution combination method. For this purpose, we employ hill climbing for locally improve the solutions. The improvement is done in two phases. First, the method takes the target-template alignment and, using the template sequence, it reshuffles the gap positions randomly. This is done by choosing the gap position and its length randomly. This process is repeated 5 times. Each time the alignment is altered, the respective comparative model is calculated by the program MODELLER and the obtained model gets assessed by the DOPE function. Whenever the score of the new model is better than the old one, the new alignment replaces the old one. If by the end of the specified number of attempts no improvements are made, the original model is kept. The same process is repeated in the second phase using the target sequence instead of the template sequence.

2.5 Reference Set Creation Method (RSCM)

In this method, the initial reference set (RSet) is created. RSet is divided into two sets, the high quality solution set, *HQRefSet,* and the diverse solution set, *DivRefSet.* Usually these two subsets contain equal number of solutions ($|RSet| = 20$, $|HQRefSet| = 10$ and $|DivRefSet| = 10$). To create the *RSet* for the first time, the population resulting from the SIM method is sorted from best to worst solution. After sorting, the first $|HQRefSet|$ solutions are chosen to form the *HQRefSet*. Concerning the *DivRefSet,* the set of most diverse solutions are admitted to this set. The diversity of a candidate solution is characterized as the Levenshtein distance (LD) between the sequence of the considered candidate solution in the population and the solutions already present in the *RSet.* For each candidate solution in the population, the LD between that solution and each of the solutions present in the *RSet* is calculated and the minimum of these distances is recorded for each solution. Then, the solutions with the maximum minimum-distance are added to the *DivRefSet*.

2.6 Subset Generation Method and Solution Combination Method (SCM)

This Subset Generation Method generates subsets using the *RSet*. For computational reasons, we limit this method in enumerating all the subsets of size two.

The SCM combines the information from the candidate solutions present in each subset and gives rise to new solutions. Since we restricted subsets of size two, a crossover operator is used, which crosses over information between two solution points and gives rise to a new solution carrying information from both initial solutions. The two target and template sequences from the two parents grouped together by the SGM undergo crossover separately.

First, the two target sequences from the parent alignments are extracted. The sequential gap indices for the two sequences are extracted into vectors. A location for crossover is chosen among the gap indices where the vectors differ. The vector which ends up with the smaller index gets combined with the vector starting with the larger index. The resulting combined vector specifies the locations of the gaps in the child target sequence. This entire procedure is repeated for the template sequence as well. The respective homology model is built afterwards by MODELELR using the new target-template sequences.

2.7 Reference Set Update Method (RSUM)

This method updates the reference set by using the population resulting from SCM. New solutions are admitted to the *RSET* and replace an existing solution in either of the two cases:

- If the considered solution's fitness score is better than that of the worst Solution in the *HQRefSet,* then it replaces it, and the updated *HQRefSet* is sorted again from best to worst.
- If the solution considered is more diverse than the least diverse Solution in the *DivRefSet.*

This entire process is repeated until all of the candidate solutions generated by the SCM are considered.

3 Experimental Results and Discussion

3.1 Experimental Procedure

The proposed approach is tested by predicting the 3D conformations of proteins of known structures that are stored in PDB. We compare our algorithm to three other approaches (GA; TS; PSO) proposed in [8, 9] that fall under the same framework of satisfying spatial restraints and using the same proteins. A total of eight target-template protein pairs were used. In addition to this, we compare our approach to a pure ab-initio approach [16] and a fragment based approach [17]. This is done by predicting the structures of three proteins that were also used in their experiments.

The results were assessed by calculating the root mean standard deviation (RMSD) between the predicted structures and the native ones found in the protein data bank (PDB). The RMSD values were calculated using PyMOL [18], which is also used to visualize the predicted structures. The algorithm starts with 100 initial random population of solutions. A reference set of size 20 was maintained throughout its execution. Each solution was allowed five iterations of improvement in the solution improvement phase. The termination criterion for the algorithm was that the reference set did not get updated.

3.2 Comparison with GA, TS and PSO

We compared the effectiveness of SS with that of genetic algorithm (GA), tabu search (TS), and particle swarm optimization (PSO). The comparison is based on the best RMSD results of the algorithms. TS and PSO were only tested on two protein pairs, hence the rest of the experiments were carried out by comparing our algorithm to the GA. The results are summarized in Table 1. RMSD values for the first two cases, using 2CCY-1BBH and 2RHE-3HLA target-template protein pairs, indicate that Tabu-Search and PSO perform poorly compared to GA and SS. Similar results were obtained while modeling the protein 3HLA. TS and PSO resulted in poor RMSD values, 15.209 and 15.24 respectively, whereas GA and SS maintained much lower RMSDs, 7.579 and 5.791. Furthermore, for both proteins, SS led to lower RMSD values than those of GA.

For the remaining proteins, SS produced better RMSD values than those of GA for three cases out of the total six and a similar RMSD for one, while it failed on the remaining two. The total count shows that SS yielded competitive results in comparison with GA, TS, and PSO.

Table 1. RMSD values for 8 protein.

Template	Target	Target length(AA)	% identity	% coverage	GA(Å)	TS(Å)	PSO(Å)	SS(Å)
2CCY (5:A-128:A)	1BBH (5:A-131:A)	126	21.3%	97.0%	2.362	3.048	3.762	1.752
2RHE (3-108)	3HLA (4:B-98:B)	94	2.4%	96.0%	7.579	15.209	15.24	5.791
1BOV (2:A-69:A)	1LTS (17:D-102:D)	85	4.4%	83.5%	8.579	N/A	N/A	7.84
1PAZ (3-93)	1AAJ (21-105)	84	27.5%	84.7%	2.1	N/A	N/A	1.339
1EG0 (1-84)	1ABA (1-87)	86	16.9%	100.0%	4.1	N/A	N/A	4.164
2RHE (8-112)	3CD4 (2-100)	98	21.7%	100.0%	6.5	N/A	N/A	3.938
1FLB (15-86)	1HOM (7-57)	50	17.6%	75.0%	1.1	N/A	N/A	5.02
9RNT (2-104)	2SAR (7:A-91:A)	84	13.1%	88.5%	4.8	N/A	N/A	5.472

3.3 Comparison with Ab-Initio Method and Fragment Based Assembly Method

Further experiments were conducted by comparing our algorithm to a pure Ab-initio based method proposed in [16], and to a fragment based assembly method proposed in

[17]. The three methods compared in this section are based on entirely different techniques, making the comparison difficult. The performance of these techniques is directly related to the amount of information known about the predicted proteins. We ran our algorithm on the three proteins that were tested in the other two methods. The results are summarized in Table 2. To make this comparison as fair as possible, we chose the worst hits returned by BLAST as template proteins, provided they are evolutionary related to the target at least. Clearly, the results that the homology-based PSP is a better choice made provided that enough information exists. This is supported by the RMSD results for predicting 1CRN and 1UTG. To make things even more difficult, the template chosen to predict the structure of 1ROP only covers 57 % of the target protein sequence. This means that no spatial restraints could be extracted to constrain the target structure's prediction for the rest of the 43 %. Despite this, our approach returned an RMSD value of 7.033 Å, which is lower than 12.14 Å resulting by the pure ab-initio approach, and somewhat worse than the 5.43 Å returned by the fragment based approach.

Table 2. Comparison with ab-initio method and fragment based modelling.

Template	Target	% identity	% coverage	Mansour et al. (Å)	Fragment based SS (Å)	SS(Å)
1ED0(2-44)	1CRN(1-47)	51.0%	93.0%	9.01	8.05	0.952
1UTR(23:A-90:A)	1UTG(1-70)	57.0%	97.0%	14.78	12.34	1.71
4LCT(271:A-304:A)	1ROP(1-63)	42.0%	57.0%	12.14	5.43	7.033

4 Conclusion

We have presented a homology based protein modeling using SS to predict the 3D folds by satisfying spatial restraints. The heuristic was guided by assessing the resulting structures using two scoring functions, GA341 and DOPE. Our algorithm was evaluated by running it on 11 target-template protein pairs and the results were assessed by measuring the RMSD errors between the native and predicted structures. Comparisons were made between our results and those produced by 3 algorithms from previous literature. Out of 8 protein pairs, our approach resulted in lower RMSD values in 5 cases. Furthermore, we compared our approach to a pure ab-initio and a fragment-based assembly methods, by predicting 3 protein structures. Our algorithm was superior in terms of RMSD values for the first two cases, and returned comparable values in the third case.

References

1. Skolnick, J., Fetrow, J.: From genes to protein structure and function: novel applications of computational approaches in the genomic era. Trends Biotechnol. **18**(1), 34–39 (2000)
2. Welch, W.: Role of quality control pathways in human diseases involving protein misfolding. Semin. Cell Dev. Biol. **15**(1), 31–38 (2004)

3. Guyeux, C., Côté, N., Bahi, J., Bienia, W.: Is protein folding problem really a NP-Complete one? First investigations. J. Bioinf. Comput. Biol. **12**(01), 1350017 (2014). 24 pages
4. Abbass, J., Nebel, J.C., Mansour, N.: Ab Initio protein structure prediction: methods and challenges. In: Elloumi, M., Zomaya, A.Y. (eds.) Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data. IEEE-Wiley, New Jersey (2014)
5. Jones, D.: GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J. Mol. Biol. **287**(4), 797–815 (1999)
6. Kopp, J., Schwede, T.: Automated protein structure homology modeling: a progress report. Pharmacogenomics **5**(4), 405–416 (2004)
7. Chothia, C., Lesk, A.: The relation between the divergence of sequence and structure in proteins. EMBO **5**(4), 823–826 (1986)
8. John, B.: Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res. **31**(14), 3982–3992 (2003)
9. Doong, S.: Protein homology modeling with heuristic search for sequence alignment. In: 40th Annual Hawaii International Conference on System Sciences, p. 128, Waikoloa (2007)
10. Šali, A., Blundell, T.: Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**(3), 779–815 (1993)
11. Altschul, S.: Basic local alignment search tool. J. Mol. Biol. **215**(3), 403–410 (1990)
12. Pearson, W.: Empirical statistical estimates for sequence similarity searches. J. Mol. Biol. **276**(1), 71–84 (1998)
13. Mishra, S., Saxena, A., Sangwan, R.: Fundamentals of homology modeling steps and comparison among important bioinformatics tools: an overview. Sci. Int. **1**(7), 237–252 (2013)
14. Shen, M., Sali, A.: Statistical potential for assessment and prediction of protein structures. Protein Sci. **15**(11), 2507–2524 (2006)
15. Marti, R., Laguna, M.: Scatter Search: Basic Design and Advanced Strategies. Int. Artif., vol. 7, no. 19 (2003)
16. Mansour, N., Ghalayini, I., Rizk, S., El-Sibai, M.: Evolutionary algorithm for predicting all-atom protein structure. In: Proceedings of the ISCA 3rd International Conference on Bioinformatics and Computational Biology, pp. 7–12, New Orleans (2015)
17. Mansour, N., Terzian, M.: Fragment-based computational protein structure prediction. In: The Eighth International Conference on Advanced Engineering Computing and Applications in Sciences, pp. 108–112 (2015)
18. The PyMOL Molecular Graphics System. Schrödinger, LLC