A Memetic Algorithm for Protein Structure Prediction based on Conformational Preferences of Aminoacid Residues

Mario Inostroza-Ponta
Center of Biotec. and Bioeng.
DIINF-USACH
Santiago, Chile
mario.inostroza@usach.cl

Camilo Farfán
DIINF-USACH
Santiago, Chile
camilo.farfan@usach.cl

Márcio Dorn
INF-UFRGS
Porto Alegre, Brazil
mdorn@inf.ufrgs.br

ABSTRACT

Memetic Algorithms are intrinsically concerned with exploiting all available knowledge about the problem under study. The incorporation of problem domain knowledge is not an optional mechanism, but a fundamental feature of the MAs. In this paper we present a Memetic Algorithm to tackle the three dimension protein structure prediction problem. The algorithm uses a structured population and incorporates a simulated annealing algorithm as a local search strategy. The algorithms takes advantage of the knowledge stored in the PDB. The algorithm was tested on six proteins and results show its ability to predict good tertiary structures.

Keywords

Memetic Algorithms, Protein Structure Prediction

1. INTRODUCTION

Proteins are large biological macromolecules composed by one or more chains of amino acid residues [3]. The three-dimensional structure of a protein gives researchers very important information about the function of the protein in the cell [3]. The prediction of the 3-D structure of polypeptides based only on the amino acid sequence is NP-complete [2], and it is a problem that has challenged biochemists, biologists, computer scientists and mathematicians over the last 40 years [1]. In this work, we propose a memetic algorithm that uses a population of thirteen agents with a hierarchical structure. It also incorporates the information stored in the Protein Data Bank to reduce the search space. The performance of the algorithm is evaluated using six target protein sequences of amino acids residues.

2. PROPOSED MEMETIC ALGORITHM

In order to tackle the 3D-PSP problem we designed a memetic algorithm [4] that uses a structured population and incorporates a Simulates Annealing implementation of the local search strategy (Alg.1).

Algorithm 1 Pseudocode of the Memetic algorithm

Input: $time_{max}$: time to run, S: aminoacid sequence
Output: sol_{best}: best solution found

```
    //Generate random initial population
 1: pop ← initialPopulation(S)
 2: pop ← updatePopulation(pop)
 3: sol_best ← pop[0].pock[0]; gen ← 0
 4: count ← 0; radio ← 90
 5: repeat
 6:    for each agent_i, i = 1 : 12 do
 7:       par_1 = pop[⌊(i − 1)/3⌋].pock[rand(1 : 5)]
 8:       par_2 = pop[i].pock[rand(1 : 5)];
 9:       pop[i].cur ← crossover(par_1, par_2)
10:    end for
11:    for i = 1 : 12 do
12:       pop[i].cur ← localSearchSA(pop[i].cur, radio)
13:       pop[i].cur ← mutation(pop[i].cur)
14:    end for
15:    radio = radio − 1
16:    pop ← updatePopulation(pop) //Restart control
17:    if sol_best >= pop[0].pock[0] then
18:       count + +
19:    else
20:       count ← 0
21:       sol_best ← pop[0].pock[0]
22:    end if
23:    if count == noImprovement then
24:       pop ← restartPopulation(pop)
25:    end if
26: until max_time
27: return sol_best
```

3. COMPUTATIONAL EXPERIMENTS

The memetic algorithm was coded using AmberTools 14. We used the energy AMBER potential energy function. The algorithm was ran ten times for 24 hours for each one of six protein sequences: 2EVQ (12 res.), 1K43 (14 res.), 1DEP (15 res.), 1E0Q (17 res.), 1RPV (19 res.) and 1L2Y (20 res.). The algorithm reaches low energy values while at the same time reaches good solutions in terms of RMSD. Figure 1 shows an example of the convergence of the algorithm for the best solution and the average of the population. It is

GECCO '15 July 11-15, 2015, Madrid, Spain

ACM ISBN 978-1-4503-3488-4/15/07.
DOI: http://dx.doi.org/10.1145/2739482.2764682

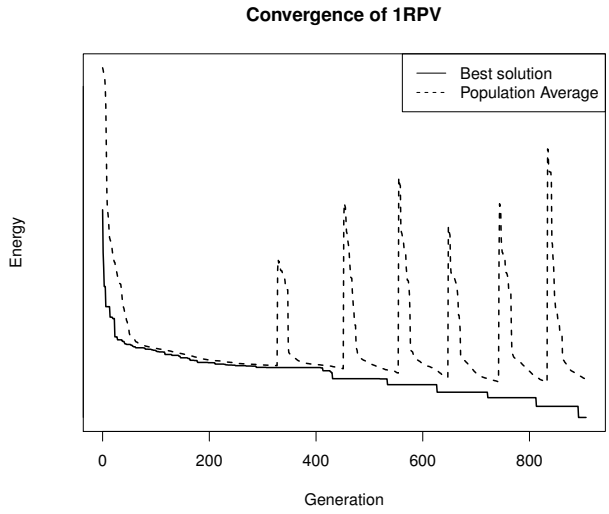

Figure 1: Convergence of the MA for one run on protein 1RPV. Picks show a restart of the population

possible to see every time that a restart procedure is performed and its contribution to improve the overall solution quality. The same behaviour occurs when the APL is not used. In order to measure the contribution of the use of the APL, we modify the MA to instead of selecting angles from the APL, they are randomly generated. The results are shown in Table 1 and Table 2 . Both algorithm converge to similar energies, but the algorithm using APL also reaches solutions with better RMSD. It shows that selecting angles from the APL guides the algorithm to a more correct solutions. The *p-values* computed using a Wilcoxon test, indicates that the difference between the results of both algorithms are statistically significant (Table 2).

Although the use of energy as a fitness function is a common place in the 3D-PSP, it does not guarantee that the best energy solution will have the best RMSD when compare with the experimental protein. The quality of the predicted structures were evaluated by similarity comparisons with the structures of the experimental proteins obtained from the PDB. Quality measurements have been made in terms of the root mean square deviation (RMSD) between the position of the C_α atoms of the predicted and the experimental structures. In Figure 2 we show the ribbons representation of three resulting proteins.

Table 1: Best solution found by the memetic algorithm using APL.

Protein	APL	
	Energy	RMSD
2EVQ	-94.2 (-70.2 ±14.7)	3.58 (2.87 ±0.84)
1K43	-558.6 (-515.2 ±36.4)	2.50 (2.71 ±0.83)
1DEP	-304.2 (-272.7 ±24.4)	1.43 (1.03 ±0.44)
1E0Q	-280.9 (-236.7 ±32.2)	7.08 (4.77 ±2.10)
1RPV	-1027.9 (-937.3 ±76.9)	2.15 (1.88 ±0.46)
1L2Y	-261.9 (-225.7 ±32.4)	5.43 (4.04 ±1.10)

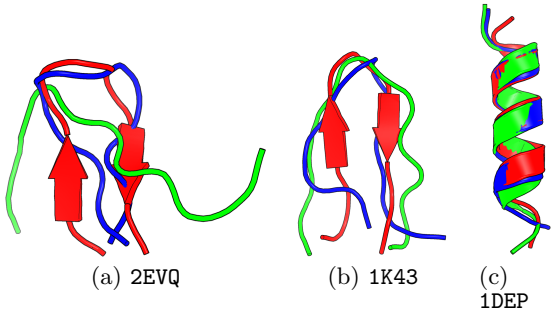

Figure 2: Ribbon representation of the experimental (red), lowest RMSD (blue) and predicted (green) 3-D structures. The C_α of the experimental and the predicted 3-D structure are fitted. Amino acid side chains are not shown for clarity. Graphic representation was prepared with PYMOL.

Table 2: Best solution found by the memetic algorithm without the use of APL

Protein	No APL		*p-value*	
	Energy	RMSD	Energy	RMSD
2EVQ	-92.6 (-40.8 ±39.7)	3.76 (2.76 ±1.05)	8.398^{-2}	4.922^{-1}
1K43	-447.8 (-405.0 ±29.1)	5.05 (4.77 ±1.23)	$\mathbf{1.953^{-3}}$	$\mathbf{1.953^{-3}}$
1DEP	-377.3 (-239.2 ±102.3)	4.12 (4.28 ±0.46)	3.750^{-1}	$\mathbf{5.889^{-3}}$
1E0Q	-141.2 (-49.4 ±59.0)	5.04 (5.41 ±1.16)	$\mathbf{1.953^{-3}}$	4.922^{-1}
1RPV	-1075.1 (-947.3 ±58.4)	5.66 (5.66 ±0.56)	5.566^{-1}	$\mathbf{1.953^{-3}}$
1L2Y	-187.4 (-23.8 ±86.0)	5.01 (5.39 ±0.59)	$\mathbf{1.953^{-3}}$	$\mathbf{9.766^{-3}}$

4. CONCLUSION

In this paper, we introduced a novel search strategy for the PSP problem. The search strategy is based on a memetic algorithm that incorporates in the search process the information extracted from the PDB data, by a Angles Probablity List. Results showed that the proposed algorithm is able to find good solutions in terms of Energy and also in terms of RMSD when compared to the experimental structure. Additionally, results of the computational experiments show the contribution of the incorporation of the APL in the algorithm, since it allows the algorithm to reach solutions more similar to the experimental ones.

ACKNOWLEDGEMENTS

This work was partially supported by grants from FONDECYT 11121288 and Basal Funds FB0001, Chile and FAPERGS (002021-25.51/13) and MCT/CNPq (473692/2013-9), Brazil.

5. REFERENCES

[1] Baxevanis, A., and Quellette, B. *Bioinformatics: A practical guide to the analysis of genes and proteins*, 2 ed. John Wiley and Sons, Inc., New York, USA, 1990.

[2] Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., and Yannakakis, M. On the complexity of protein folding. *J. Comput. Biol. 5*, 3 (1998), 423–466.

[3] Lesk, A. M. *Introduction to Bioinformatics*, 1 ed. Oxford University Press Inc., New York, USA, 2002.

[4] Moscato, P. On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. Tech. Rep. Caltech Concurrent Computation Program, Report. 826, Pasadena, California, USA, 1989.