Genetic Algorithms with Local Search Optimization for Protein Structure Prediction Problem

Igor Berenboym
Computer Science Dept.
The Open University,
108, Ravutski St., Raanana, Israel
(+972) 50-8314438
igor.berenboym@gmail.com

Mireille Avigal
Computer Science Dept.
The Open University,
108, Ravutski St., Raanana, Israel
(+972) 9-7781232
miray@openu.ac.il

ABSTRACT

This paper presents a new Genetic Algorithm for Protein Structure Prediction problem in both 2D and 3D hydrophobic-hydrophilic lattice models, introduced in [1]. Our algorithm evolves a new local-search genetic operation (called Pull-Move and well described in [2]), into the standard GA[1] ([3,4]). The experiments show that performing a set of Pull-Moves in addition to standard genetic operations in GA (such as crossover and mutation) leads to significant energy improvements. The paper also introduces the Global Energy as fitness function and explains the advantages of utilizing it rather than the standard Free Energy. The experimental results are even more impressive when using the Global Energy as fitness function in GA.

ACM Categories and Subject Descriptors

I.2.8 [Artificial Intelligence]: *Problem Solving, Control Methods and Search*

General Terms: Algorithms, Experimentation

Keywords: Genetic Algorithms, Protein Folding, Lattice HP Model, Global Energy, Pull Move, Local Search.

1. OUR METHOD

Despite providing relatively good results, the standard GA suffers from incapability to perform minor local changes in the structure, such that may improve its overall energy by positioning several residues in better way. The main cause for such a lack is the inflexibility of crossover and mutation operations, as defined in [4]. In other words, it is hard to slightly change the protein structure in the middle using the above operations only.

Therefore, our algorithm, called PMGA, performs a set of Pull-Moves in addition to the mutations before each reproduction. Each residue is a subject for a possible PM, and in case it fits the definition in [2], PM operation is executed with the same probability as the mutation. Intuitively, the operation "pulls" a single residue diagonally, causing some other residues and connections between them to move the same way in order to maintain the validness of the structure. In most cases, the amount of moved connections is small. Thus, the Pull Move operation undoubtedly fits our request for being fluent with local changes.

[1]Abbreviations used: GA-Genetic Algorithm, HP-Hydrophobic-Hydrophilic, FE/GE-Free/Global Energy, PM-Pull Move, PMGA-Pull Moves with Genetic Algorithm, PSP-Protein Structure Prediction.

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-130-9/08/07.

2. THE GLOBAL ENERGY

The Free Energy (FE), as defined in [4], is the negative value of the number of hydrophobic (H-kind) neighbors in the space, which are not consecutive in given sequence. The main disadvantage of this energy definition is that it does not count interactions of H-residues that are not directly adjusted in the space. For instance, two H-residues located on the opposite vertices of the same lattice square do not contribute even miserable value to the overall energy, although they are located reasonably near to each other.

To illustrate this point let us assume that we have a protein with $(HP)^k$ -like representation in 2D Lattice HP Model, i.e. H and P-residues appear alternately in the chain. If we imagine the space as a chessboard, it is obvious that at *any* given folding of such a chain, all H-residues will stand on the fields of the same color. Thus, no two H-residues will be neighbored in the space, meaning that *any* structure obtained from this kind of initial chain will have zero Free Energy. That is quite unfair situation, since a stable compact structure with many hydrophobic interactions is given the same fitness function as the completely extended one.

To confront this problem, a new energy function, called Global Energy (GE), is defined. The formal definition of both the energies is given in the equations on Figure 1. A_i represents the 'H' or 'P' type of *i*-th residue of the sequence, w_{ij} is the weight of an interaction between residues *i* and *j* and (x_i, y_i, z_i) are the coordinates of *i*-th residue of structure S in 3D space.

$$FE(S) = -\sum_{i=1}^{n-2}\sum_{j=i+2}^{n} w_{ij},\ GE(S) = -\sum_{i=1}^{n-2}\sum_{j=i+2}^{n} w'_{ij}$$

when

$$w_{ij} = \begin{cases} 1, \text{ when } A_i = A_j = 'H' \text{ and } |x_i - x_j| + |y_i - y_j| + |z_i - z_j| = 1 \\ 0, \text{ otherwise} \end{cases},$$

$$w'_{ij} = \begin{cases} \dfrac{1}{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}, \text{ when } A_i = A_j = 'H' \\ 0, \text{ otherwise} \end{cases}$$

Figure 1: The Global Energy vs. the Free Energy

3. EXPERIMENTAL RESULTS

12 HP sequences with various lengths (first 9 are taken from [4] and the last three are the largest ones from [2]) are observed in 2D model under both the energies, while 2 additional $(HP)^k$ - kind ones (of lengths 50 and 100) are tested under GE only (since FE=0 for any folding in this case). In 3D model, ten 64-residues long sequences (taken from [3]) are observed under both the energies, while additional $(HP)^k$ - kind sequence with the above length is tested under GE.

The effect of evolving Pull-Move operations into GA is observed in both 2D and 3D Lattice HP models, while each one of them is tested separately under Free and Global Energy (4 test cases totally). Each test case was performed 10 times, and the best solution was saved. The results on Table 1 compare the performance of PMGA relatively to GA, while the ones in 2D under FE are also compared to the optimal known energy values, according to [4].

2D / FE	2D-1	2D-2	2D-3	2D-4	2D-5	2D-6
GA	-9	-12	-9	-8	-14	-21
PMGA	-9	-12	-9	-8	-14	**-22**
Optimal	-9	-12	-9	-8	-14	-23
	2D-7	**2D-8**	**2D-9**	**2D-10**	**2D-11**	**2D-12**
GA	-21	-34	-37	-47	-47	-45
PMGA	-21	-34	**-38**	**-50**	**-51**	-45
Optimal	-21	-34	-42	-51	-53	-48

2D / GE	2D1	2D2	2D3	2D4	2D5	2D6	HP50
GA	-17.3	-29.7	-16.6	-15.4	-33	-60.4	-42.7
PMGA	-17.3	-29.7	**-17.1**	-15.4	**-35.6**	**-66.9**	**-42.8**
	2D7	2D8	2D9	2D10	2D11	2D12	hp100
GA	-50.4	**-136**	-112	-198	-194	-175	-107
PMGA	**-53.6**	-135	**-132**	**-205**	**-210**	**-176**	**-109**

3D / FE	3D-1	3D-2	3D-3	3D-4	3D-5
GA	-30	-35	-41	-38	-39
PMGA	-30	**-36**	-41	-38	-39
	3D-6	3D-7	3D-8	3D-9	3D-10
GA	-32	-27	-35	-36	-30
PMGA	**-33**	-27	-35	**-37**	-30

3D / GE	3D-1	3D-2	3D-3	3D-4	3D-5	HP64
GA	-74.4	-92	-144.9	-95	-110.3	-102.6
PMGA	**-77.3**	**-100**	**-145.4**	**-101.9**	**-120.9**	**-105.2**
	3D-6	3D-7	3D-8	3D-9	3D-10	
GA	-86.1	-76.6	-99.8	-104.2	-72.5	
PMGA	**-93.6**	**-80.1**	**-108.6**	**-109.5**	**-80.1**	

Table 1: Experimental Results – PMGA vs. GA

4. CONCLUSIONS

The results show that in general, PMGA performs better than the standard GA in both 2D and 3D Lattice HP models and under both Free and Global energies. On the other hand, the results produced by PMGA are in average somewhere in the middle between the ones produced by the standard GA and the optimal ones, meaning that relatively simple local operation shortened the distance to the optimum by half.

The positive effect of the local search on GA is undoubtedly shown. Therefore, we reasonably believe that in order to produce excellent results for the PSP problem using GA, some kind of the local search is more than essential. It can be other local operation, any local searching method combined with GA or even some real-time human local interference, as is done in [2]. We found this direction very important and potential for the relevant further research.

5. REFERENCES

[1] K. F. Lau and K. A. Dill. Theory for Protein Mutability and Biogenesis. *Proc. Nat. Acad. Sci., U.S.A. (1990), vol. 87, pp.638-642.*

[2] N. Lesh, M. Mitzenmacher and S. Whitesize. A Complete and Effective Move Set for Simplified Protein Folding. *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (2003), pp.188-195.*

[3] R. Unger and J. Moult. Genetic Algorithm for 3D Protein Folding Simulations. *Proceedings of the 5th International Conference on Genetic Algorithms (1993), pp. 581-588.*

[4] R. Unger and J. Moult. Genetic Algorithm for Protein Folding Simulations. *Journal of Molecular Biology (1993) **231**, pp.75-81.*