Solving the Protein Structure Prediction Problem Through a Multiobjective Genetic Algorithm

R. Day, J. Zydallis, and G. Lamont

Department of Electrical and Computer Engineering
Graduate School of Engineering and Management
Air Force Institute of Technology, Wright-Patterson AFB, OH 45433, USA
Richard.Day, Jesse.Zydallis, Gary.Lamont@afit.edu

ABSTRACT

[1]The Protein Structure Prediction problem, which involves correctly predicting the geometrical conformation of a fully folded protein, is extremely difficult to solve and there currently is no "best" method of generating solutions. This paper focuses on an energy minimization technique and the use of a multiobjective genetic algorithm, the multiobjective fast messy genetic algorithm (fmGA) to obtain solutions to this problem. We extend the fmGA to generate solutions to the PSP problem as a multiobjective problem using the CHARMm energy function. Further, the results of the multiobjective fmGA formulation compare very favorably to our previous results from the single objective fmGA formulation.

Keywords: Multiobjective Optimization, Energy Minimization, Protein Structure Prediction Problem, Fast Messy Genetic Algorithm.

1 INTRODUCTION

Interest in discovering a methodology for solving the protein structure prediction problem extends into many fields of study including biochemistry, medicine, biology, engineering, and scientific disciplines [3]. Yet, the panacea procedure to solving this Grand Challenge Problem has eluded researchers. Approaches for finding the final resting nature (tertiary conformation) of a protein range from empirical, using x-ray crystallographic studies, to mathematical modelling, such as minimum energy models. Although structures are accurately found using empirical methods, these methods can take months before results are achieved. It is for this reason that minimum energy models are utilized; however, there has been no computational breakthrough in regards to accurately forecasting the final folded state of a protein using these modelling approaches.

In the past, Evolutionary Algorithms have been shown to be well suited for optimization problems with single and multiple objectives [6]. In fact, many single objective Evolutionary Algorithm approaches have been successfully applied to the Protein Structure Prediction problem using the energy minimization model. Our previous research has obtained "good" results, for small peptides, using the single objective fast messy genetic algorithm (fmGA) [4], [5]. The extension of this work to the multiobjective realm is done to increase the effectiveness of the algorithm. Previous research by Zydallis [6] has shown a multiobjective version of the fmGA to be effective and efficient when applied to multiobjective optimization problems. These results have motivated the extension of our PSP work to the multiobjective domain. The two proteins analyzed here are [Met]-Enkephelin, which consists of 5 residues and 24 dihedral angles (Tyr-Gly-Gly-Phe-Met amino acids), and Polyalanine which consists of 14 residues and 56 dihedral angles ($ALA_1, \ldots, ALA_{14}$ amino acids). Each of the dihedral angles is represented by a binary string of 10 bits yielding a landscape size of 24^{1024} and 56^{1024} respectively.

The purpose of this paper is to provide, for the first time, an analysis of a multiobjective approach utilizing the fmGA as the searching mechanism applied to the Protein Structure Prediction problem. This paper covers the analysis of two proteins: [Met]-enkephlan and Polyalanine and presents the modified formulation of the fitness function to extend it to a multiobjective formulation. A description of the fmGA algorithm, testing, results, analysis and conclusions follow.

2 MULTIOBJECTIVE FORMULATION

To extend the fmGA to the multiobjective arena, the single objective CHARMm protein energy fitness function must be modified to generate multiple independent fitness functions. In the single objective implementation, the CHARMm energy function is utilized and consists of a summation of ten major terms. To utilize a multiobjective approach, the objectives are drawn from the set of terms within the CHARMm energy function. Specifically, the energy function is decomposed into the connected (bonded) and non-connected (nonbonded) atom energies. These two objectives are further divided into:(bonded) stretching, bending, torsion, and (non-bonded) electrostatic, and van-der-Waals en-

[1] The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

ergy terms. Each term represents a function separately targeted for minimization. The decision variables are the dihedral angles for the protein being solved. The decision maker is provided with the set of solutions, the Pareto Front, that the algorithm has found and makes a decision as to which area of the front they prefer. Ultimately, the multiobjective approach is anticipated to yield "better" results than the single objective approach as each of the functions are simultaneously optimized.

The CHARMm energy function consists of ten major terms. This yields a multiobjective formulation consisting of anywhere from two to ten objective functions. The ten terms of the CHARMm energy model are:

1. E_1 = Fixed Energy
2. E_2 = Non-Bonded Energy
3. E_3 = Non-Bonded Energy One-Four
4. E_4 = Dependent Bond Energy
5. E_5 = Independent Bond Energy
6. E_6 = Dependent Angle Energy
7. E_7 = Independent Angle Energy
8. E_8 = Dependent Dihedral Energy
9. E_9 = Independent Dihedral Energy
10. E_{10} = Independent Improper Dihedral Energy

In our previous single objective research and the multiobjective testing presented in this paper, only nine of the CHARMm energy terms are utilized. The last term, *Independent Improper Dihedral Energy*, is not utilized as it does not provide a significant impact to the overall energy value. The multiobjective formulation presented in this paper consists of the following two objectives, built from the nine available objectives listed above. The terms of the objectives are grouped this was to give us valuable insight into tertiary structures that may not be of a unique minimum energy and others that are stable structures.

- *Objective 1*

$$\mathcal{F}_1 = \sum_{k=1}^{3} E_k \tag{1}$$

- *Objective 2*

$$\mathcal{F}_2 = \sum_{k=4}^{9} E_k \tag{2}$$

3 FAST MESSY GENETIC ALGORITHM

The fmGA is a binary, stochastic, variable string length, population based approach to solving optimization problems. The fmGA was developed by Goldberg, Deb and Kargupta [1] and later applied to the PSP problem by Merkle, Gates, Lamont and Pachter [2]. The main difference between the fmGA and other genetic approaches is the ability of the fmGA to explicitly manipulate building blocks (BBs) of genetic material in order to obtain "good" solutions and potentially the global optimum. Most other single and multi objective approaches implicitly manipulate these BBs to obtain solutions. The fmGA contains three phases of operation: *the initialization phase, the building block filtering (BBF) phase, and the juxtapositional phase*, which includes various computational parameters. Prior to execution of the algorithm, the user specifies which BB sizes to execute. During execution of the fmGA, the initial BB size is run to completion (through all three phases), the BB size is incremented and the subsequent BB size is run to completion and so on until the last BB size completes. At this point the algorithm terminates with a final solution for the user.

In the initialization phase of the fmGA, the population size is determined by an equation derived to overcome the noise present in the BBF process. Once the population size is determined, each of the population members are randomly generated and their corresponding fitness values are calculated through the use of the CHARMm energy model. These population members are referred to as fully specified since all of the associated loci of the population member contain specified allelic values.

The fully specified population members from the initialization phase are then systematically reduced in length to the user specified BB size through the use of a BBF schedule. The BBF process randomly deletes a specified number of bits from each population member over a number of generations specified in the schedule. This deletion of bits is alternated with tournament selection so that only the "best" partial strings are kept for processing in the subsequent generations. At the end of the BBF process, the entire population consists of population members of the user specified BB length. These population members are referred to as underspecified strings strings since there are some bit positions that do not have an allelic value specified. A Competitive Template (CT) is used to evaluate these partial strings.

The juxtapositional phase takes the "good" BBs found from the filtering process and combines them together through a cut-and-splice operator. This operator randomly chooses two strings and based on the probabilities of cut and splice, cuts the strings and splices them together accomplishing the goal of crossing over information between the strings. This process is alternated with tournament selection so that only the best strings are kept from generation to generation. At the conclusion of this phase, fully specified strings exist in the population and the next BB size is evaluated via an outer loop over these three phases

The competitive templates are an extremely important part of the fmGA. To evaluate an underspecified population member, the CT is copied into a temporary location and the bits that are specified in the population member replace the bits of the CT within this temporary location. Once this is accomplished, the temporary string is evaluated and the resulting fitness is associated with the underspecified population member. In the case of an overspecified population member, which may occur when the cut-and-splice procedure causes a member to have multiple occurrences of a particular bit, a left-to-right method is employed. In this method, the first allelic value encountered for a particular loci is recorded as the value present for evaluation purposes.

Population members that contain very few specified bits with respect to the overall string length, as is the case at the end of the BBF process, are highly dependent on the CT. The reverse holds for strings that have the majority of their bits specified (the case at the end of the juxtapositional phase), as they only need to take a few bits from the CT. This illustrates the importance of the CT in the overall execution of the fmGA. Previous research is based on the concept of generating a random CT and periodically updating this template with the best found population member over the course of execution of the algorithm. In this paper we provide an analysis of the multiobjective approach along with a more intelligent choice for the CT in order to increase the effectiveness of the fmGA.

In the multiobjective version of the fmGA, one CT is used per objective function. Each of the CTs are tied to a particular objective and are updated with the best population members per that objective at the end of the juxtapositional phase. Random CTs are a natural starting point since the goal of the fmGA work is to generate a robust algorithm that obtains solutions for various optimization problems. In order to increase the effectiveness of the algorithm problem domain knowledge is incorporated into the fmGA and the number of CTs utilized is increased. The three CT methods used in this paper are:

1. Randomly generate a CT per objective function, then conduct a localized search on these CTs. This memetic approach involves conducting a local search of the competitive templates before each template update.

2. The use of fully specified population members containing a Secondary Structure as the CTs. Seeded CTs are hard coded into the fmGA using known alpha-helix and beta-sheet dihedral angles. The algorithm is expected to achieve better fitness values at a faster rate for proteins having either of these secondary structures through this method.

3. Using more than a single CT per objective function developed via the aforementioned methods. This approach allows for more exploration since each population member is evaluated using a number of templates and therefore has the potential to find a better solution by searching different areas of the landscape.

4 TESTING, RESULTS AND ANALYSIS

Testing of the various CT approaches in the fmGA algorithm was accomplished on a 1.7 GHz Intel P-4 machine with 256 MB of RAM, using the Red Hat 7.1 distribution of the Linux operating system. The code was written in ANSI C. The MO fmGA algorithm was executed ten times for each of the experiments in order to provide statistical results. All of the results presented here are averaged over ten runs and the Pareto Front plots are the combined results over the ten runs. Over all runs the following fmGA parameters were kept constant; cut probability = 0.02, splice probability = 1.00, primordial generations = 200, juxtapositional generations = 100, total generations = 300. An input schedule was used to specify sizes of the building blocks the algorithm uses and during which generations BBF occurs. Tests were conducted using both [Met]-Enkephelin, with 240 bit length strings and BB sizes 6-10, and Polyalanine, with 560 bit length strings and BB sizes 16-20.

Figure 1 presents the Pareto Front found from the [Met]-enkephlan testing. In this figure the Random CT obtained the best distribution of points along the front, all of the points are Pareto Front members if combined with the other methods, as well as the largest cardinality out of the three CT methods tested. This is expected as [Met]-enkephlan does not contain a structure and hence the Alpha or Beta methods do not provide better results than the random generation of the CT.

Figure 2 presents the Pareto Front found from the Polyalanine testing. In this figure the Alpha CT method performed the best in terms of the overall distribution of points along the front as well as the cardinality of the Pareto Front set. This is expected since Polyalanine has a alpha-helix structure and therefore the Alpha CT should provide the best results.

The CT testing produced "good" results and results that are anticipated considering the structure of the proteins analyzed. The multiobjective (MO) implementation of the fmGA compares very favorably to the original fmGA results regarding minimum energies. Since

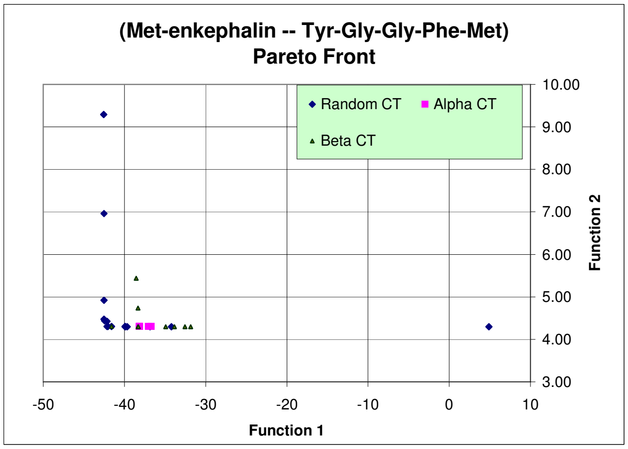

Figure 1: [Met]-enkephlan Pareto Front

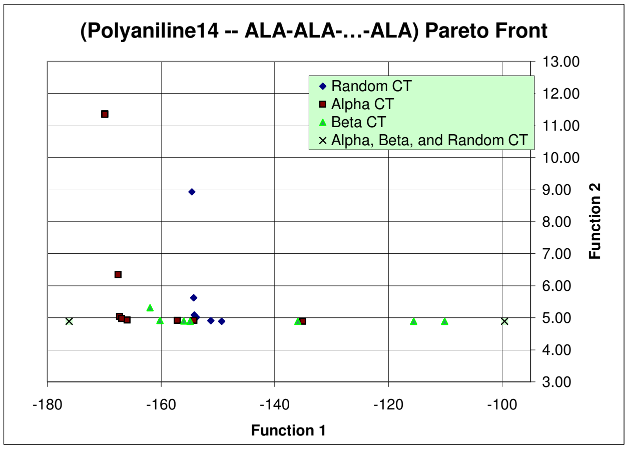

Figure 2: Polyalanine Pareto Front

the MO fmGA implementation involves decomposing the summation of terms used in the original fmGA, one can sum up the two fitnesses and obtain what the single objective value would be and then make a limited comparison to the original fmGA results. Table 1 presents the results of the best found fitness for each of the proteins from the original fmGA testing and the MO fmGA testing. For [Met]-enkephlan the MO fmGA finds the best overall fitness value when compared with the original fmGA. In the Polyalanine analysis, the MO fmGA compares very favorably to the original.

Table 1: Best Fitness Found

	Alpha	Beta	Random	A,R&B
Met	-31.716	-33.191	**-34.114**	-31.834
MO Met	-33.857	-37.287	**-38.047**	N/A
Poly	-163.393	-157.203	-159.105	**-171.760**
MO Poly	-162.246	-156.624	-149.052	**-171.314**

5 CONCLUSIONS

We have presented the results for three different innovative CT generation schemes used in the MO formulation of the PSP problem. The results presented in this paper support our hypothesis that the MO version of the fmGA would produce better results than the original and that the Random CT scheme would perform well in cases where the protein does not contain a structure and a CT method that includes the structure of the protein tested would perform the best there. Future work will look at larger proteins and other protein structures to include TIM barrel, Sandwich, Roll, flavodoxin, and β-lactamase. Also addressed will be the incorporation of a sharing mechanism to provide a better distribution of points along the Pareto Front.

REFERENCES

[1] David E. Goldberg, Kalyanmoy Deb, Hillol Kargupta, and Georges Harik. Rapid, accurate optimization of difficult problems using fast messy genetic algorithms. In Stephanie Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 56–64, San Mateo, CA, July 1993. Morgan Kaufmann Publishers.

[2] Laurence D. Merkle, George H. Gates, Jr., Gary B. Lamont, and Ruth Pachter. Application of the parallel fmga to the protein structure prediction problem. *Proceedings of the Intel Supercomputer Users' Group Users Conference*, pages 189–195, 1994.

[3] Kenneth M. Merz and Scott M. Le Grand, editors. *The Protein Folding Problem and Tertiary Structure Prediction*. Springer, New York, 1994.

[4] Steven R. Michaud. Solving the Protein Structure Prediction Problem with Parallel Messy Genetic Algorithms. Master's thesis, Air Force Institute of Technology, Wright Patterson AFB, March 2001. AFIT/GCS/ENG/01M.

[5] Steven R. Michaud, Jesse B. Zydallis, Gary Lamont, and Ruth Pachter. Scaling a genetic algorithm to medium-sized peptides by detecting secondary structures with an analysis of building blocks. In Matthew Laudon and Bart Romanowicz, editors, *Proceedings of the First International Conference on Computational Nanoscience*, pages 29–32, Hilton Head, SC, March 2001.

[6] Jesse B. Zydallis, David A. Van Veldhuizen, and Gary B. Lamont. A Statistical Comparison of MOEAs Including the MOMGA–II. In Eckart Zitzler, Kalyanmoy Deb, Lothar Thiele, Carlos A. Coello Coello, and David Corne, editors, *First International Conference on Evolutionary Multi-Criterion Optimization*, pages 226–240. Springer-Verlag. Lecture Notes in Computer Science No. 1993, 2001.