Evolutionary Multi-Objective Optimisation with Quantile Constraint for the Protein Structure Similarity Problem

Sune S. Nielsen[1], Wiktor Jurkowski[2], Grégoire Danoy[1], Juan Luis Jiménez Laredo[1], Reinhard Schneider[2], El-Ghazali Talbi[3], and Pascal Bouvry[1]

[1] Faculty of Sciences, Technology and Communication,
University of Luxembourg,
6 rue R. Coudenhove-Kalergi, L-1359, Luxembourg
{sune.nielsen, gregoire.danoy, juan.jimenez, pascal.bouvry}@uni.lu
[2] Luxembourg Centre for Systems Biology,
University of Luxembourg,
7, avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg
{wiktor.jurkowski, reinhard.schneider}@uni.lu
[3] INRIA-Lille Nord Europe
40, avenue Halley Bât. A, Park Plaza 59650 Villeneuve d'Ascq France
el-ghazali.talbi@inria.fr

1 Introduction

In biology, the subject of protein structure prediction is of continued interest, not only to keep charting the molecular map of the living cell, but also to design proteins with new functions. Given a reference protein and its corresponding tertiary (3D) structure, this work is concerned with finding 1) the most *diverse* nucleotide sequences which 2) produce very *similar* 3D structures. This task is different from conventional 3D prediction seeking to predict the structure of one given sequence. In order to efficiently evaluate the protein structure *similarity objective* we introduce a fast evaluation method based on an approximate prediction of its secondary structure. This permits to use a Genetic Algorithm (GA) to efficiently probe the enormous search space of possible sequences. Since we are additionally interested in finding as many different sequences as possible, we use the *diversity-as-objective* (DAO) approach [2] to push the algorithm farther into widespread areas of the solution-space. The problem is consequently bi-objective and tackled with a Multi-Objective GA (MOGA). To circumvent the possible dominance of the *diversity objective* over the *similarity objective*, the Quantile Constraint (QC) is introduced in which the worse quantile of the population in terms the *similarity objective* is penalized. The efficiency of this MOGA is experimentally demonstrated using a a reference protein, i.e. *256b*, which consists of 106 amino-acids packed into 4 main helices.

2 Protein Structure Similarity Problem

Every protein is uniquely defined by its RNA/DNA code of length N which determines the amino acid sequence, $A = \{aa_i\}$ where $1 \leq i \leq N$. This sequence will result in a folding of the chain into a three-dimensional structure composed of *alpha-helices* and *beta-sheets* among others. The segmentation of these structure components along the sequence is defined as the secondary structure, hence a close relation between secondary and tertiary structures exists. With the PROFphd software [1], the secondary structure type $T_{pred}(i)$ can be predicted per amino acid aa_i in A with a reliability, $R_{pred}(i) \in \{1...10\}$ by means of posterior neural network training. With $T_{ref}(i)$ the actual type found at position i of the reference secondary structure, the estimated *similarity objective* score of the sequence A, $F_{sec}(A)$, is the sum of the (mis)matches between the secondary structure types predicted for A and the reference secondary structure:

$$F_{sec}(A) = \sum_{i=1}^{N} M(T_{pred}(i), R_{pred}(i), T_{ref}(i)) \; . \tag{1}$$

where

$$M(T_{pred}(i), R_{pred}(i), T_{ref}(i)) = \begin{cases} 0 & \text{if } T_{pred}(i), T_{ref}(i) \notin \{H, E\} \\ -R_{pred}(i) & \text{if } T_{pred}(i) = T_{ref}(i) \\ R_{pred}(i) & \text{if } T_{pred}(i) \neq T_{ref}(i) \end{cases}$$

As stated, biologists are not only looking for a few very good solutions wrt. $F_{sec}(A)$ but rather for a large diverse collection of good solutions. An effective and simple measure of the distance between two sequences is the Hamming-distance, defined as the number of permutations necessary to change one sequence into the other. With $A = \{aa_i\}$, $A' = \{aa'_i\}$ and $1 \leq i \leq N$, we define the Hamming distance between them as:

$$d_{Hamm}(A, A') = \sum_{i=1}^{N} d_i,\ d_i = \begin{cases} 0 & \text{if } aa_i = aa'_i \\ 1 & \text{otherwise} \end{cases} . \tag{2}$$

To obtain a non-negative diversity objective value for minimisation, we compute the average Hamming distance to all other $M-1$ individuals in the current population, minus the sequence length N. For the diversity objective we have:

$$F_{div}(A) = N - \frac{1}{M-1} \sum_{i=1}^{M-1} d_{Hamm}(A, A_i) . \tag{3}$$

3 Quantile Constraint

Initial experiments with protein *256b* have shown that the dual-objective approach delivers a constantly very high population-diversity of about 90%, but in terms of $F_{sec}(A)$, a conventional single objective GA was able to outperform it though with diversity dropping below 30%. To remedy this issue, and focus the MOGAs search on the main *similarity objective*, $F_{sec}(A)$, we introduce the Quantile Constraint (QC). At the end of every generation, the population P_t is divided according to $F_{sec}(A)$ into a C_q% and a $100 - C_q$% sized partition, with C_q being the selected quantile size. All individual sequences in the former, less fit, partition are assigned a constraint penalty. This penalty effectively prevents the less fit partition from mating, hence the population is cleaned from individuals far spread in the solution space, but with a poor $F_{sec}(A)$ score. Experiments have been conducted using $C_q \in \{5\%, 10\%, 25\%\}$ constraint thresholds.

4 Conclusion

In this paper we have presented a new approach to studying the relation between protein sequences and their resulting 3D structure as a first step in a possible way of conducting future protein design. By defining the task of finding highly diverse sequences with most similar structures we have been able to model it as an two-objective optimisation problem. We show that we are able to find many highly varying protein sequences which score better than the reference protein in terms of the secondary structure prediction. This applies to almost two thirds of individual sequences in the final population of the MOGA, which make them interesting for further studies. By adding the Quantile Constraint (QC) approach we are able to shift the focus arbitrarily between the *diversity-* and *similarity-objectives* ($F_{div}(A)$ and $F_{sec}(A)$), and to obtain better results than a standard single objective GA on $F_{sec}(A)$, while keeping a much higher diversity. In addition, the convergence of $F_{sec}(A)$ was observed as being steeper than the standard GA which promises very good solutions given a high evaluation budget.

Acknowledgments. Work funded by the National Research Fund of Luxembourg (FNR) as part of the EVOPERF project at the University of Luxembourg with the AFR contract no. 1356145. Experiments were carried out using the HPC facility of the University of Luxembourg

References

1. B Rost and C Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19(1):55–72, May 1994.
2. Andrea Toffolo and Ernesto Benini. Genetic diversity as an objective in multi-objective evolutionary algorithms. *Evol. Comput.*, 11(2):151–167, May 2003.