Evolutionary Multi Objective Optimisation with Diversity as Objective 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
{sune.nielsen,gregoire.danoy,juan.jimenez,pascal.bouvry}@uni.lu
[2] Luxembourg Centre for Systems Biology,
University of Luxembourg
{wiktor.jurkowski,reinhard.schneider}@uni.lu
[3] INRIA-Lille Nord Europe, France
el-ghazali.talbi@inria.fr

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 [1] to push the algorithm farther into wide-spread 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. 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.

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. Andrea Toffolo and Ernesto Benini. Genetic diversity as an objective in multi-objective evolutionary algorithms. *Evol. Comput.*, 11(2):151–167, May 2003.