Preference-Based Genetic Algorithm for Solving the Bio-Inspired NK Landscape Benchmark

Christof Ferreira Torres, Sune S. Nielsen, Grégoire Danoy, and Pascal Bouvry

FSTC, University of Luxembourg
6, rue Richard Coudenhove-Kalergi,
Kirchberg, Luxembourg
christof.ferreira.001@student.uni.lu
{sune.nielsen,gregoire.danoy,pascal.bouvry}@uni.lu

Abstract. In molecular biology, the subject of protein structure prediction is of continued interest, not only to chart the molecular map of living cells, but also to design proteins with new functions. In this work a Preference-Based Genetic Algorithm (PBGA) is proposed aiming to optimise NK Landscape based benchmarks designed and shown to mimic properties of the Inverse Folding Problem (IFP) of proteins. The proposed algorithm incorporates a weighted sum model in order to combine fitness and diversity into a single objective function scoring a set of individuals as a whole. By adjusting the sum weights, direct control of the preferred emphasis on fitness vs. diversity in the algorithm population is achieved by means of a selection scheme iteratively removing the least contributing individuals. The proposed algorithm is compared to other algorithms where better results are achieved both in terms of fitness and diversity.

1 Introduction

Protein engineering in general aims at designing molecules with desired properties. A method that would allow to successfully design such molecules would find applications in a number of areas such as designing improved enzymes for biotechnology applications or new antibodies more specific towards already known targets. However evaluating and therefore optimising real biological instances is very computationally demanding. Nielsen et al. [6] recently proposed a novel NK Landscape benchmark suite that mimics the properties of the Inverse Folding Problem (IFP). The latter originally consists, given a protein sequence of N amino acids, in finding other sequences that will result in the same 3D structure. The resulting optimisation problem is highly *multi-modal* and the algorithm proposed in this work addresses this aspect by adding a novel diversity controlling mechanism. The preference-based approach employs a Weighted Sum Model (WSM) in order to control the desired bias between fitness and diversity. The resulting WSM score allows to iteratively determine and remove the individual in the combined parent and offspring population with the lowest overall fitness contribution with respect to the set preferences. The remainder of this

article is organised as follows. First the current state-of-the-art is situated in related literature in Section 2, then a detailed description of the problem and the biological background is introduced in Section 3. In Section 4 the contribution of this work in terms of achieving an adjustable level of fitness and diversity as a Preference-Based Genetic Algorithm (PBGA) is presented. Section 5 describes the experiments conducted and the results obtained for the NK benchmark suite. Finally the contribution, results and perspectives are summarised in Section 6.

2 State-of-the-art

In meta-heuristics, the subject of exploration vs. exploitation characteristics has been thoroughly studied. In this aim, a number of works have sought to maintain and control diversity in population-based meta-heuristics, e.g. crowding methods by DeJong [2], fitness sharing by Goldberg and Richardson [3], cellular algorithms by Alba and Dorronsoro [1], diversity preserving selection strategies based on hamming distance Shimodaira [7] and on altruism by Laredo et al. [4].

Preference-based algorithms have been discussed in literature [5] [8] and refer to algorithms where user preference is incorporated in the choice of regions in the solution or objective space. Preference can be incorporated in a number of ways, e.g., by modifying fitness evaluation or selection schemes. The Indicator Based Evolutionary Algorithm (IBEA) [9] is an example where an indicator that characterises the population in whole is used to guide the algorithm by eliminating least desired individuals of the parent and offspring population union. The proposed PBGA in this paper borrows the same principle of iterative elimination, determining the overall most preferable subset directly rather than achieving it as a indirect effect of designed mechanisms.

3 Bio-Inspired NK Landscape Benchmark Problem

In the NK benchmark problem as well as in the Inverse Folding Problem (IFP), a single solution is represented as a sequence $A = \{aa_i\}$ and consists of N residue positions, where $1 \leq i \leq N$ and $aa_i \in \{1, ..., 20\}$ corresponds to the set of 20 possible amino acids. The overall size and the number of local "hills and valleys" of the NK landscape model can be adjusted via changes to its two parameters, N and K. In this paper we make use of two novel NK benchmark model instances[1] proposed by Nielsen et al. [6], which are the combination of two NK models, $F^A(x)$ and $F^B(x)$, by a simple multiplication with different K and different neighbourhood definitions as defined in the table below.

[1] The NK Landscape Protein IFP Benchmark Suite - `http://nk-ifp-bench.gforge.uni.lu/index.html`

Model	Setting
NK-IFP-1	$F^A(x)$: a $K = 4$ semi-adjacent circular neighbourhood is designed as follows: $\{x_{i2}, x_{i1}, x_{i+1}, x_{i+2}\}$, omitting the central position x_i. $F^B(x)$: a $K = 3$ neighbourhood of uniform random distribution.
NK-IFP-2	$F^A(x)$: a $K = 4$ semi-adjacent circular neighbourhood as *NK-IFP-1*. $F^B(x)$: a $K = 5$ neighbourhood of uniform random + 20 positions wide triangular distribution.

4 A Novel Preference-Based Approach

The main idea of the preference-based approach is to use a Weighted Sum Model (WSM) in order to constantly maintain a current population best fulfilling the defined preferences. In an iterative manner, the weakest individuals from the combination of parent and offspring populations are determined and removed until the desired population size is achieved.

Algorithm 1 Preference-Based Genetic Algorithm

```
1: Initialise(P_0)
2: t ← 0
3: while t < t_max do
4:     Q_t ← makeNewOffspringPop(P_t)
5:     R_t ← P_t + Q_t
6:     while |R_t| > |P_t| do
7:         I ← getWeakestIndividual(R_t)
8:         R_t ← R_t − I
9:     end while
10:    P_t ← R_t
11:    t ← t + 1
12: end while
```

The procedure *getWeakestIndividual* of determining the weakest individual in Algorithm 1 is defined as follows:

1. Systematically remove one individual
2. Compute the weighted sum score according to Equation 1
3. Add the individual back to the population
4. Repeat from step 1. until all individuals have been tried once and the worst individual can be determined.

The weighted sum score of a given population P is calculated as follows:

$$WSM_{score}(P) = -W_{fit} \cdot F_{fit}(P) + W_{div} \cdot F_{div}(P) \tag{1}$$

Note the negation of W_{fit} in Equation 1 as we want to maximise diversity but also minimise fitness at the same time.

The population fitness F_{fit} is computed by simply taking the average of the fitness of all M individuals of the current population P:

$$F_{fit}(P) = \frac{1}{M} \sum_{i=1}^{M} F(x) \tag{2}$$

An effective and simple measure of distance between two sequences is the Hamming-distance. For two sequences $A = \{aa_i\}$ and $A' = \{aa'_i\}$ where $1 \leq i \leq N$, the normalised Hamming distance between them is defined as:

$$d_{Hamm}(A, A') = \frac{1}{N} \sum_{i=1}^{N} d_i \quad where \quad d_i = \begin{cases} 0 & \text{if } aa_i = aa'_i \\ 1 & \text{if otherwise} \end{cases} \tag{3}$$

The population diversity F_{div} is computed by taking the average Hamming distance of each M individuals to the remaining $M-1$ individuals of the population P:

$$F_{div}(P) = \frac{1}{M \cdot (M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} d_{Hamm}(A_i, A_j), \quad \forall i \neq j \tag{4}$$

5 Experimental Results

To study the performance of the proposed algorithm with respect to fitness and diversity convergence, a number of experiments have been conducted to compare against a number of standard Genetic Algorithms such as the generational (gGA), the synchronous cellular (scGA) and finally the steady-state (ssGA). The algorithm was tested with these weight ratio settings:

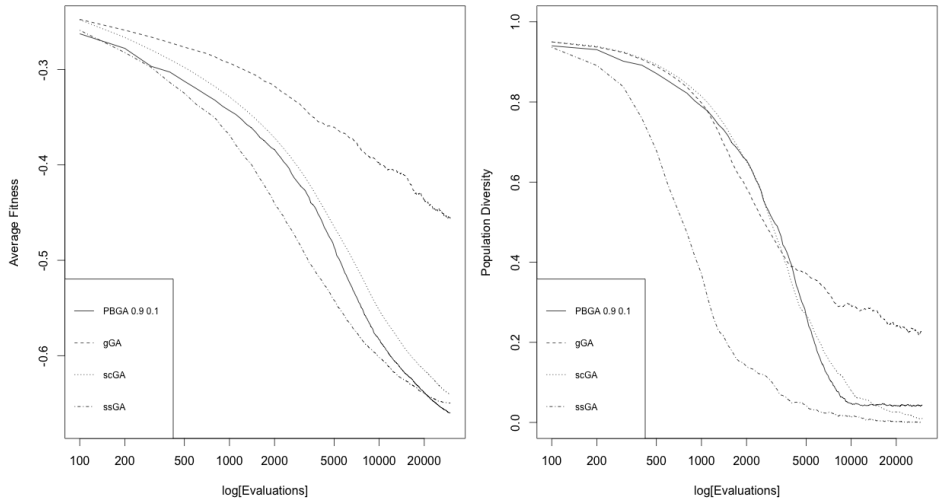
$$W_{(fit,div)} = \{(1.0, 0.0), (0.9, 0.1), (0.8, 0.2), (0.7, 0.3), (0.5, 0.5), (0.3, 0.7)\}.$$

Table 1 summarises the settings and parameters used to conduct the experiments.

Table 1: Experimental settings.

Setting	**Value**
Standard GAs	gGA, scGA and ssGA
Population size	100
Termination condition	30000 function evaluations
Number of independent runs	30
Selection	Binary tournament (BT)
Neighbourhood	C9 in scGA
Crossover operator	SPX, $p_c = 0.9$
Mutation operator	Uniform, $p_m = \frac{1}{N}$
Elitism	2 individuals (for gGA)

Figure 1a illustrates the convergence of fitness for the best performing PBGA setting in comparison with the gGA, scGA and the ssGA. The gGA performs the worst and the PBGA with a weight setting of (0.9, 0.1) surpasses the ssGA and achieves better final fitness results than all of the other GAs. Figure 1b illustrates the diversity convergence for the same algorithms. It is noted that the PBGA achieves a higher diversity than the scGA and ssGA while at the same time having better fitness results. Similar graphs are obtained for the NK-IFP-2 model and are hence not shown here.



(a) Fitness convergence PBGA vs. GAs (b) Diversity convergence PBGA vs. GAs

Fig. 1: NK benchmark model NK-IFP-1 average fitness and diversity convergence.

Table 2 summarises average fitness and diversity for all the algorithms tested highlighting best and worst algorithm results in light and dark grey respectively. With a weight setting of (0.9, 0.1) the PBGA achieves the best fitness for both benchmark models with -0.662 for the best value and -0.660 on average for model 1 and with -0.632 for the best value and -0.631 on average for model 2. It is interesting to note that the PBGA with a weight setting of (0.5, 0.5) achieves better results than the gGA in terms of fitness as well as diversity for both models with -0.574 vs. -0.559 for the best fitness value and -0.511 vs. -0.456 on average for model 1 and with -0.550 vs. -0.545 for the best fitness value and -0.485 vs. -0.429 on average for model 2.

In order to provide statistical confidence, the Wilcoxon test indicator was applied with a 5% significance level. With a weight setting of (0.9, 0.1), the PBGA clearly outperforms the gGA and the scGA with statistical confidence for the average fitness with values -0.662 vs. -0.559 and -0.662 vs. -0.644 respectively

for model 1 and with values -0.632 vs. -0.545 and -0.632 vs. -0.621 respectively for model 2, whereas in comparison with the ssGA the PBGA does not achieve as quick good results as the ssGA, but surpasses the ssGA in the end and achieves better average fitness values of -0.662 vs. -0.650 respectively for model 1 and with values -0.632 vs. -0.628 respectively for model 2. However, as seen in Figure 1a for model 1, the final slope is steeper than the ssGA, indicating better performance. The steeper final slope can be explained by the constantly high diversity as seen in Figure 1b for model 1, which allows for continued exploration while the standard GAs suffer from premature convergence.

Table 2: Final values in terms of fitness and diversity averaged over 30 independent runs for the two NK benchmark models.

	Model 1				Model 2			
	Fitness		Diversity		Fitness		Diversity	
Algorithm	Best	Average	Best	Average	Best	Average	Best	Average
$PBGA_{1.0\ 0.0}$	-0.649	-0.648 $\pm 0.37E{-3}$	0.005	0.002 $\pm 1.78E{-3}$	-0.628	-0.628 $\pm 0.27E{-3}$	0.004	0.001 $\pm 1.56E{-3}$
$PBGA_{0.9\ 0.1}$	-0.662	-0.660 $\pm 1.07E{-3}$	0.041	0.043 $\pm 0.84E{-3}$	-0.632	-0.631 $\pm 0.76E{-3}$	0.031	0.038 $\pm 3.25E{-3}$
$PBGA_{0.8\ 0.2}$	-0.652	-0.627 $\pm 12.3E{-3}$	0.250	0.337 $\pm 43.8E{-3}$	-0.621	-0.594 $\pm 13.4E{-3}$	0.310	0.406 $\pm 48.1E{-3}$
$PBGA_{0.7\ 0.3}$	-0.629	-0.582 $\pm 23.3E{-3}$	0.508	0.612 $\pm 52.1E{-3}$	-0.602	-0.557 $\pm 22.5E{-3}$	0.542	0.639 $\pm 48.7E{-3}$
$PBGA_{0.5\ 0.5}$	-0.574	-0.511 $\pm 31.4E{-3}$	0.774	0.833 $\pm 31.4E{-3}$	-0.550	-0.485 $\pm 32.3E{-3}$	0.787	0.846 $\pm 29.6E{-3}$
$PBGA_{0.3\ 0.7}$	-0.527	-0.458 $\pm 34.5E{-3}$	0.880	0.909 $\pm 14.6E{-3}$	-0.503	-0.440 $\pm 31.7E{-3}$	0.888	0.913 $\pm 12.4E{-3}$
gGA	-0.559	-0.456 $\pm 51.4E{-3}$	0.145	0.227 $\pm 40.9E{-3}$	-0.545	-0.429 $\pm 58.0E{-3}$	0.138	0.221 $\pm 41.1E{-3}$
scGA	-0.644	-0.641 $\pm 1.54E{-3}$	0.017	0.010 $\pm 3.36E{-3}$	-0.621	-0.619 $\pm 1.20E{-3}$	0.013	0.009 $\pm 2.18E{-3}$
ssGA	-0.650	-0.645 $\pm 0.18E{-3}$	0.005	0.001 $\pm 1.97E{-3}$	-0.628	-0.628 $\pm 0.14E{-3}$	0.001	0.001 $\pm 0.42E{-3}$

6 Conclusion

In this paper a novel Preference-Based Genetic Algorithm (PBGA) was presented in combination with a weighted sum model, which allows to shift focus arbitrarily between diversity and fitness with a direct effect on the population as a whole without relying on secondary effects from added mechanisms or operators. The PBGA was tested with two NK benchmark models and compared to other GAs where final results were found comparable or better than the standard GAs on average, while the diversity of found sequences remains higher at the same time. The best results were achieved using a weight setting of (0.9, 0.1) where 0.9 represents 90% of fitness and 0.1 represents 10% of diversity. In addition, the algorithm convergence was observed as being steeper than the standard GAs, which promises even better solutions, given an evaluation budget beyond the computational limitations set in this work. Future work will be the development of a more advanced preference evaluation model using Fuzzy logic while adding more preferences such as crowding, elitism, etc. and making the selection of preferences be adaptive.

References

1. E. Alba and B. Dorronsoro. The exploration/exploitation tradeoff in dynamic cellular genetic algorithms. *Evolutionary Computation, IEEE Transactions on*, 9(2):126–142, 2005.

2. K. A. De Jong. Analysis of the behavior of a class of genetic adaptive systems. 1975.
3. D. E. Goldberg and J. Richardson. Genetic algorithms with sharing for multimodal function optimization. In *Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms*, pages 41–49. Hillsdale, NJ: Lawrence Erlbaum, 1987.
4. J. L. Jimenez Laredo, S. S. Nielsen, G. Danoy, P. Bouvry, et al. Cooperative selection: Improving tournament selection via altruism. In *The 14th European Conference on Evolutionary Computation in Combinatorial Optimisation*, 2014.
5. A. López-Jaimes and C. A. C. Coello. Including preferences into a multiobjective evolutionary algorithm to deal with many-objective engineering optimization problems. *Information Sciences*, 277:1–20, 2014.
6. S. S. Nielsen, G. Danoy, P. Bouvry, and E.-G. Talbi. Nk landscape instances mimicking the protein inverse folding problem towards future benchmarks. In *Proceedings of the Companion Publication of the 2015 on Genetic and Evolutionary Computation Conference*, GECCO Companion '15, pages 915–921, New York, NY, USA, 2015. ACM.
7. H. Shimodaira. Dcga: A diversity control oriented genetic algorithm. In *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*, pages 367–374. IEEE, 1997.
8. L. Thiele, K. Miettinen, P. J. Korhonen, and J. Molina. A preference-based evolutionary algorithm for multi-objective optimization. *Evolutionary Computation*, 17(3):411–436, 2009.
9. E. Zitzler and S. Künzli. Indicator-based selection in multiobjective search. In *Parallel Problem Solving from Nature-PPSN VIII*, pages 832–842. Springer, 2004.