FROM DARWIN TO DATA: POPULARIZATION OF GENETIC ALGORITHMS

Anna Krzywiecka[1], Olga Wieromiejczyk[1], Guillem Ylla[1]
[1]Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University

Koło Naukowe Studentów Bioinformatyki IN SILICO

INTRODUCTION

Genetic algorithms (GAs) are heuristic search algorithms used for finding optimal or near-optimal solutions by mimicking biological evolution. We offer an explanation and in-depth exploration of GA functionality, particularly in optimizing complex biological processes such as primer design for polymerase chain reaction (PCR). Here we reimplement and refine the GA designed by Wu et al. (2004).

PRIMERS

PCR has various purposes, leading to many different variants. We specifically focus on optimizing primers for:


SNP genotyping


mutation detection


comparitive genotyping

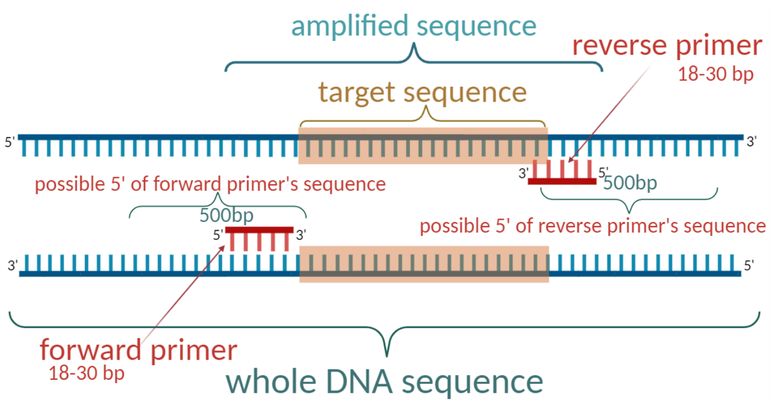
These experiments are useful in:


forensic science


diagnostics


environmental studies

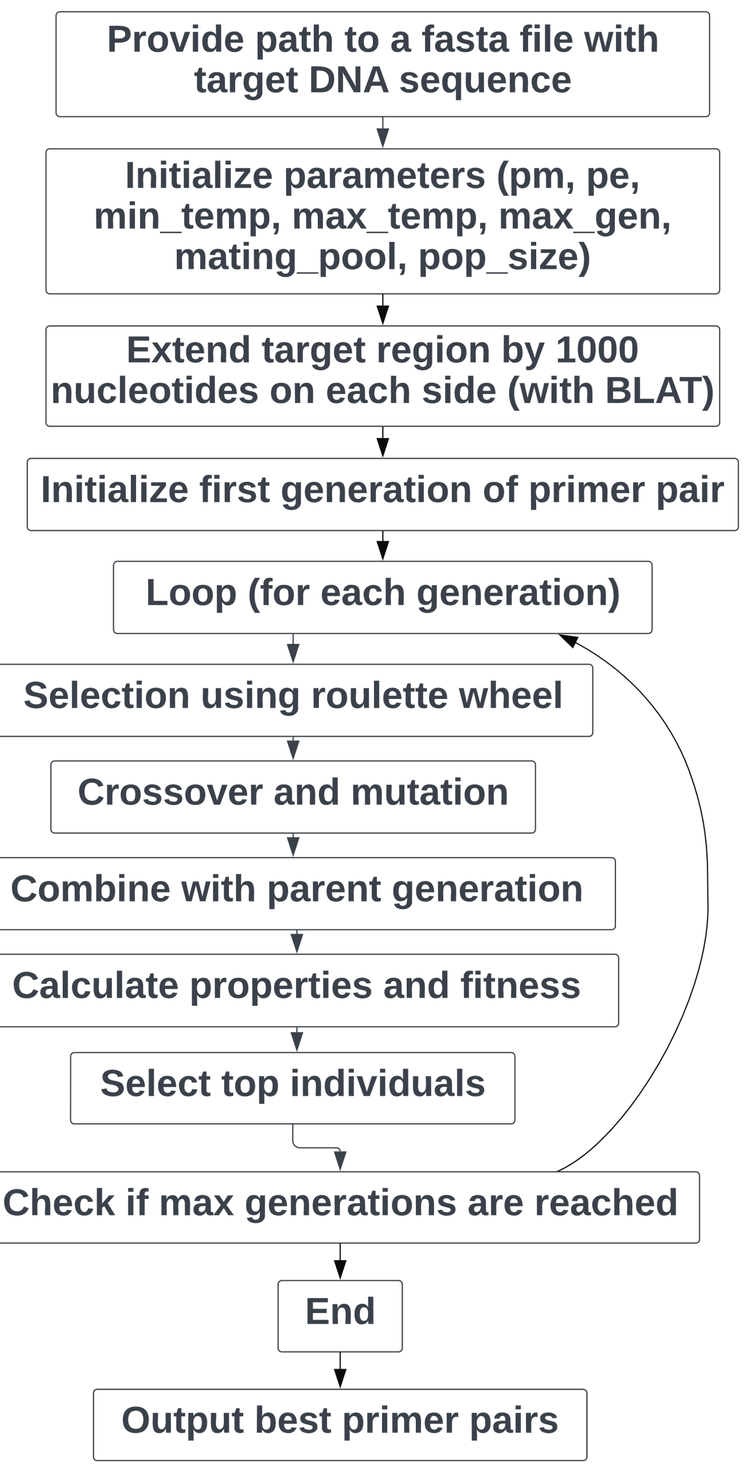
PCR makes copies of the amplified sequence. To make the reaction successful some conditions must be met, for which we consider following parameters:



Feature	Description
gc	GC content between 40% and 60%
tmd	Melting temperature difference between primers less than 5°C
uni	Specificity of primers should occur (we check it using BLAST+)
lengd	Length difference between primers not more than 5bp, the less the better
leng	Length of primer's sequence should be between 18 and 30bp
pc	Pair-complementarity shouldn't occur
term	Termination: the 3' end should be a G or a C or two of them, not more
sc	Self-complementarity shouldn't occur
fitness	$(3gc + 3tmd + 50uni + 3lengd + Leng + 10pc + 3term + 10sc + 10sc)^{-1}$

GAs

Our algorithm repeatedly modifies a population of individual primer pairs, using mutation and crossover, to create new primers. Over successive iterations, the population of primers evolves towards primer pairs with highest fitness scores.

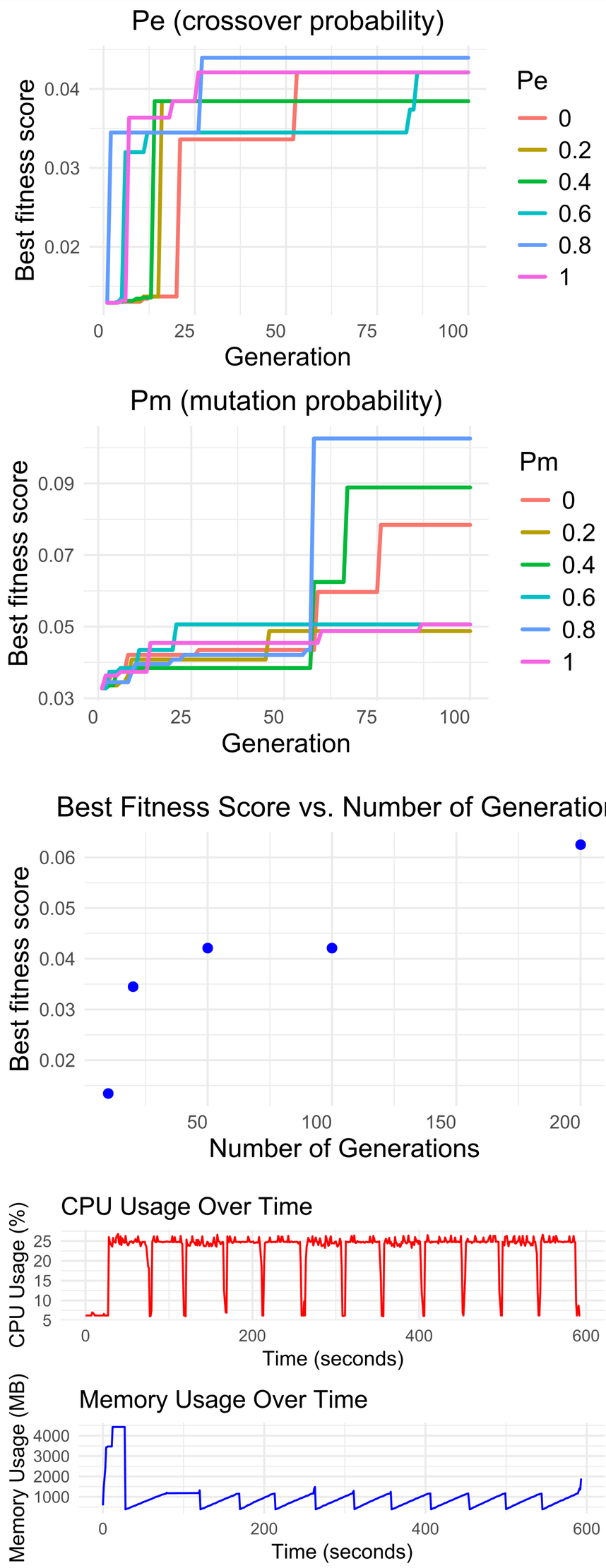


We utilize BLAT to find the target sequence in the human genome and when calculating properties we use BLAST+ to check the specificity of a primer.

In our algorithm there are several parameters, which the user specifies in our code:

Name	Description
pm	The probability of performing the mutation process
pe	The probability of performing the crossover process
min_ temp	The minimal acceptable melting temperature, Recommended: more than 50°C
max_ temp	The maximal acceptable melting temperature, Recommended: less than 72°C
max_ gen	The maximum number of generations
mating_ pool	The number of new individuals created in each generation
pop_ size	The number of individuals in a population

RESULTS



1. Higher crossing-over probability results in faster rise of the best fitness scores.
2. Higher mutation probability might result in more specific results.
3. Using a greater number of generations should result in a better fitness score.
4. The maximum CPU usage depends on the number of threads specified by the user. In this instance the specified amount was 4 and the code was run on a 16 CPU device.
5. Memory usage over time indicates that BLAT is the most memory consuming operation and afterward cyclically the memory usage increases and drops which correlates with BLAST specificity checking.

SOURCES

Figures were created with BioRender.com.
Dieffenbach, C. W., Lowe, T. M. J., & Dveksler, G. S. General Concepts for PCR Primer Design, 1993 Dec;3(3):S30-7. doi: 10.1101/gr.3.3.s30.
Katoch, S., Chauhan, S. S.,& Kumar, V. (2021). A review on genetic algorithm: past, present, and future. Multimedia Tools and Applications, 80(5), 8091–8126. https://doi.org/10.1007/s11042-020-10139-6
Wu, J. S., Lee, C., Wu, C. C., & Shiue, Y. L. (2004). Primer design using genetic algorithm. Bioinformatics, 20(11), 1710–1717. https://doi.org/10.1093/bioinformatics/bth147