IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Optimization of a genomic data compression tool using metameric genetic algorithms

RITA FERROLHO[1, 2], ARMANDO J. PINHO[1, 2, 3] and DIOGO PRATAS[1, 2, 3, 4]
[1]Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Portugal
[2]Department of Electronics, Telecommunications and Informatics, University of Aveiro, Portugal
[3]Intelligent Systems Associate Laboratory, University of Aveiro, Portugal
[4]Department of Virology, University of Helsinki, Finland

Corresponding author: Rita Ferrolho (e-mail: ritaferrolho@ua.pt), Armando J. Pinho (ap@ua.pt) and Diogo Pratas (e-mail: pratas@ua.pt).

This work was partially funded by National Funds through the FCT (Foundation for Science and Technology), in the context of the project UIDB/00127/2020. D.P. is funded by national funds through FCT, I.P., under the Scientific Employment Stimulus - Institutional Call - reference CEECINST/00026/2018.

ABSTRACT The rapid expansion of genomic data production has amplified the need for efficient lossless compression techniques, crucial for long-term storage across research, clinical, and forensic domains. This study explores the application of metaheuristic search algorithms, with a focus on Genetic Algorithms (GAs), to optimize parameter configurations within the JARVIS3 genomic data compression tool. The goal is to identify optimized parameter sets that approach local or global optima but are typically difficult to predict given the large search space. For computing this, we created a program called OptimJV3. OptimJV3 incorporates multiple GA variants and was tested to determine which configurations most effectively enhance compression ratio and computational time. We demonstrate that OptimJV3 can consistently achieve high compression ratios for different genome sequences, from small to large genome sizes. In particular, we applied OptimJV3 to compress segments of the human genome (HG), optimizing JARVIS3 parameters across different small segment sizes of the HG sample followed by the complete compression of the HG using such parameters. Surprisingly, small segments were enough to achieve a substantial compression ratio. When we extended the GA to 200 and 500 generations, the compression was further improved. For example, using 10 MB as sample and 500 generations, we achieved 1.417 bits per base in the HG bechmark (one of the highest compression records). OptimJV3 is open-source and freely available at https://github.com/cobilab/OptimJV3.

INDEX TERMS data compression, optimization, genetic algorithms, genomic sequences.

I. INTRODUCTION

The exponential growth of genomic data production has created an urgent demand for efficient, lossless compression methods to facilitate long-term storage and analysis in fields such as genomics, clinical diagnostics, and forensic investigation [1]. Traditional compression algorithms struggle with the large, repetitive nature of genomic data, making specialized approaches essential. Genetic Algorithms (GAs), a subset of metaheuristic optimization, have shown promise in optimizing complex, multidimensional problems, yet their application to genomic data compression remains poorly explored.

In this paper, we investigate the application of GAs for parameter optimization in the JARVIS3 compression tool [3] (an evolution of JARVIS2 [2]), targeting configurations that yield improved compression ratios and processing times. We developed OptimJV3, an open-source tool that leverages multiple GA variants to explore and enhance JARVIS3's performance on diverse genomic sequences.

In the following sections we describe the algorithm, then, we present the results and close with the conclusions.

II. METHOD

The proposed method (OptimJV3) in this paper is based on a genetic algorithm (GA) framework that optimizes command sequences for data compression. The optimization process begins with an initial population of candidate solutions, which can be generated heuristically or hybridly, depending on the user's choice. Each candidate solution, or individual, consists of a set of commands that are evaluated based on their fitness, which is determined by multiple criteria, including validity, the size of the compressed sequence, and compression time. The fitness function can operate under either a single-

objective (SOGA) or a multi-objective (MOGA) framework, with the latter allowing for weighted metrics to balance the importance of different objectives during optimization, namely compression ratio and computational time.

The GA iterates through several key operations: selection, crossover, and mutation. Selection involves identifying the fittest individuals from the population, with options for various selection methods such as elitism, roulette wheel selection, and tournament selection. After selection, individuals are paired for crossover, which combines their genetic information to produce new offspring based on specified crossover rates. Two crossover operators can be used: the metameric canonical crossover (MCC) and metameric random crossover (MRC), tailored to the specific structures of command and model parameters. This is followed by mutation, where selected offspring undergo random changes to one or more of their parameters, ensuring diversity and preventing premature convergence of the population.

Finally, the evaluation of offspring includes removing duplicates and ensuring each generation maintains a diverse set of candidates. The offspring are then tested, and their fitness values are calculated to determine the next generation's population. Throughout this process, the algorithm allows for user-defined parameters, such as population size and mutation rates, providing flexibility and adaptability for different scenarios.

Additionally, OptimJV3 allows the production of evolution plots and histograms, for each sequence and for each group of sequences. OptimJV3 is open-source and publicly available at https://github.com/cobilab/OptimJV3.

III. RESULTS

The use of OptimJV3 for optimizing JARVIS3 commands yielded significant improvements in both compression size and execution time for the data sequences tested. The algorithm demonstrated its effectiveness in navigating a diverse search space by employing a steady-state approach that facilitated the continuous evolution of populations.

As depicted in Fig. 1, after several generations, the best-performing individuals achieved substantial reductions in compressed sequence sizes, often exceeding initial benchmarks established during the heuristic initialization phase. The multi-objective genetic algorithm (MOGA) further enhanced these outcomes by allowing for the optimization of multiple criteria simultaneously, leading to a balanced trade-off between compression ratio and computational resources.

We applied OptimJV3 to compress segments of the human genome (HG), optimizing JARVIS3 parameters across various small segment sizes of the HG sample, which were subsequently used for parameterize the compression of the complete HG. Surprisingly, the optimization over small segments yielded a substantial improved compression ratio, demonstrating the efficacy of the approach. Extending the genetic algorithm (GA) to 200 and 500 generations resulted in further improvements in compression performance.

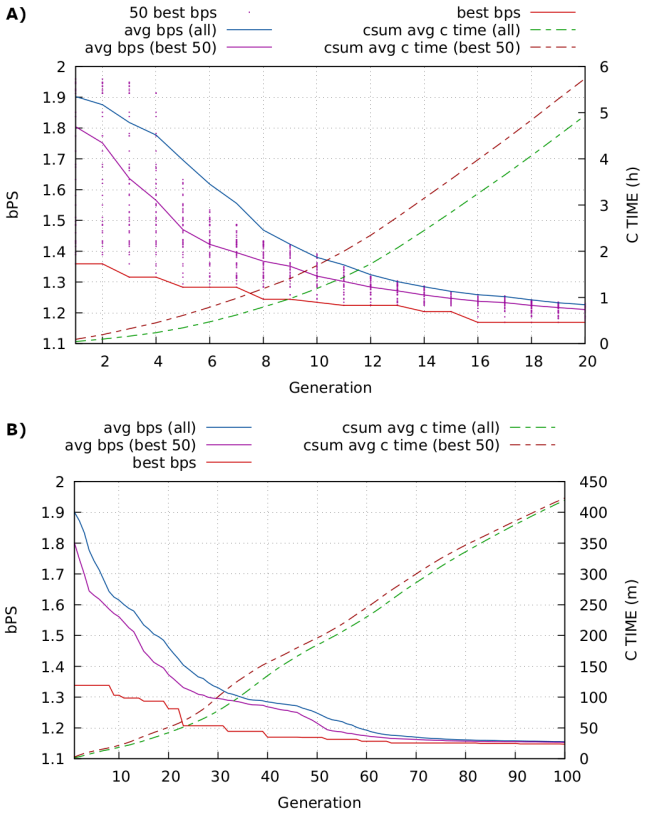

FIGURE 1. Performance examples (Bits per symbol and time) of OptimJV3 in A) Cassava genome and B) Human Y-chromosome.

For example, using a 10 MB sample and running the GA for 500 generations, we achieved an impressive compression ratio of 1.417 bits per base in the HG benchmark, marking one of the highest compression records to date.

IV. CONCLUSIONS

We created OptimJV3, an open-source tool that utilizes Genetic Algorithms (GAs) for optimizing the JARVIS3 genomic data compression model. By efficiently navigating the extensive parameter space, OptimJV3 consistently achieved high compression ratios across various genome sequences, including significant enhancements in compressing segments of the human genome. The successful application of multi-objective optimization demonstrates its potential for improving both compression efficiency and computational time, addressing the growing demand for advanced lossless compression methods in genomic data management.

REFERENCES

[1] M. Hosseini. A survey on data compression methods for biological sequences, *Information*, 2016.
[2] D. Pratas and A. J. Pinho, JARVIS2: a data compressor for large genome sequences, *Data Compression Conference*, 2023.
[3] D. Pratas, et al., JARVIS3 code repository: https://github.com/cobilab/jarvis3 , 2024.