# BeamGA Median: A Hybrid Heuristic Search Framework

Ghada Badr* badrghada@hotmail.com
Manar Hosny mifawzi@ksu.edu.sa
Nuha Bintayyash nstayash@ksu.edu.sa
Eman Albilali ealbilali@ksu.edu.sa
S. Larabi Marie-Sainte smariesainte@ksu.edu.sa

King Saud University
College of Computer and Information Sciences
Riyadh, Saudi Arabia

## ABSTRACT

BeamGA is a general hybrid heuristic framework that can be used to solve the median problem in comparative genomics, where any distance function can be used. It starts with a heuristic search approach (local beam search) in order to generate a number of solutions. Then a Genetic Algorithm (GA) is applied to refine the solutions. It considers true biological evolution scenarios by applying the concept of common intervals during the GA optimization process.

## CCS Concepts

•**Theory of computation** → **Evolutionary algorithms;** *Stochastic approximation;* •**Computing methodologies** → *Heuristic function construction;*

## Keywords

Median Problem; Bioinformatics; Genetic Alorithm; Beam Search

## 1. INTRODUCTION

The median problem is a well known problem in comparative genomics that can be applied to derive the most reasonable rearrangement phylogenetic tree for many species. The Median problem has two important properties that can help in finding good solutions with reasonable computational effort [3]: the non-uniqueness of the problem solution, and the probability to find the median on or near the $N$ gene orders rather than the center. Genomes with equal number of genes but different order can be represented as permutations. Specifically, we are concerned with finding a permutation $M$ that minimizes the sum of distances between $M$ and a set of $N$ permutations $\pi$. The distance between two genomes $d_{\pi_1,\pi_2}$ is defined as one of the genomic distance measures. Computing the distances between genomes is quite difficult and is suspected to be NP-hard [3]. The median problem can be defined as follows:

*Input*: given three signed permutations $\pi_1$, $\pi_2$, and $\pi_3$ that represent three different taxa, where different signs mean that genes are taken from opposite DNS strands.

*Output*: finding the fourth permutation (median) $\phi$, with the smallest possible distance score $S(\phi)$ from the three original permutations, such that:

$S(\phi) = d_{\pi_1,\phi} + d_{\pi_2,\phi} + d_{\pi_3,\phi}$, $S(\phi) \geq M_{min}$, $S(\phi) \leq M_{max}$

$M_{min} = [\frac{d_{\pi_1,\pi_2}+d_{\pi_1,\pi_3}+d_{\pi_2,\pi_3}}{2}]$

$M_{max} = min\{(d_{\pi_1,\pi_2} + d_{\pi_2,\pi_3}), (d_{\pi_1,\pi_2} + d_{\pi_1,\pi_3}), (d_{\pi_2,\pi_3} + d_{\pi_1,\pi_3})\}$

Different techniques have been developed in order to solve this problem. In [3] a branch-and-bound exact method based on reversal distance was applied. However, this approach cannot provide efficient results when the distance between the species is reasonably large. Other methods have been proposed based on common intervals and/or other evolutionary distances. A common interval [1] is a subset of genes that can be rearranged but still appear joined together in two or more genomes (permutations). For example, for two permutations: $P_1 = \{1, 2, 6, 7, 4, 3, -5, 8\}$ and $P_2 = \{7, 1, -6, 4, -2, -5, -3, 8\}$, $\{1, 2, 4, 6, 7\}$ is a common interval, while $\{2, 6, 7\}$ is not. In [1] an exact algorithm is proposed to solve the reversal median problem using common intervals. However, the median problem is known to be NP-hard [3], and exact algorithms become prohibitive for large problem sizes. Thus, heuristic approaches are considered as an efficient alternative.

In this paper, we propose a new hybrid framework combining a greedy search and a heuristic approach for solving the median problem aiming to overcome the limitations of other approaches with respect to speed and number of genes. In addition, any efficient evolutionary distance can be applied in the proposed general framework. We start with a greedy search initialization step, using beam search [2], followed by a Genetic Algorithm (GA), which is applied only on common intervals. To our knowledge, this combined approach is not applied before for thiscproblem. The computational results show that the algorithm is able to handle large number of genes in good time and accuracy performance.

*Corresponding Author



*GECCO'16 Companion July 20-24, 2016, Denver, CO, USA*



ACM ISBN 978-1-4503-4323-7/16/07.

DOI: http://dx.doi.org/10.1145/2908961.2931632

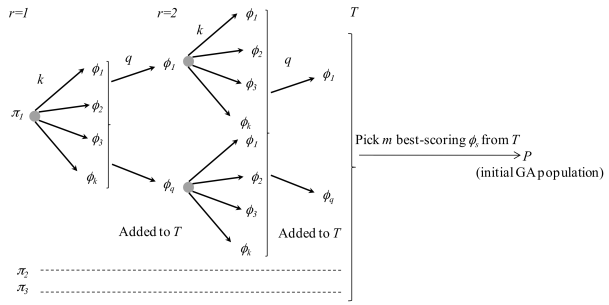


**Figure 1: Initialization Phase**

## 2. METHOD: BEAMGA

BeamGA is based on two phases, an initialization phase, where a beam search is applied in order to generate the initial population, and an optimization phase using a GA, operating on common intervals.

### *Initialization Phase: Beam Search.*

Given three signed permutations $\pi_1$, $\pi_2$, $\pi_3$, and $r$: the current number of rearrangement operations (levels) that can be applied on $\pi$, where one random operation is done per level to generate a random permutation; $k$: the number of random neighbors (permutations) that can be generated from $\pi$ at level $r$, where one random operation is done to generate one random neighbor; $q < k$: the number of best-scoring neighbors that are added to priority queue $T$ (beam size), and $m$: maximum GA population size.

First, we generate a population of size $m$ using beam search. The search starts by initializing a priority queue $T$ to be empty. For each $\pi_i$, we generate $k$ neighbors by performing one rearrangement operation on the previous permutation, and put the best generated $q$ (beam size) from $k$ neighbors in $T$ based on their median scores. The process is repeated for $r$ levels. Finally, we pick $m$ best-scoring $S(\phi)$ from $T$ and add them to $P$ (the initial GA population). the different steps of the beam search are shown in Fig. 1.

### *Optimization Phase: Genetic Algorithm.*

Given an initial population $P$, each individual in the population (a possible median $\phi$) is assigned a fitness value based on its median score $S(\phi)$ with respect to the three original permutations. Then a GA is applied. Parents are selected using tournament selection, crossover is performed by swapping a randomly chosen common interval between parents to produce two children, and mutation is done on each child by a random rearrangement operation (e.g reversal) within a randomly chosen common interval. Then, elitist replacement chooses the best individuals from the new and the old generations. The GA is repeated until the perfect median is reached or no improvement can be achieved for a number of iterations.

## 3. RESULTS

Reversals are randomly applied to the identity permutation to generate three sets of three signed permutations (taxa), simulating a shared common ancestor, where: $i$ is the number of random reversals that can be applied on the identity permutation in order to get the three original permutations $\pi$ for every taxon ($i = 2$, 6, or 10); $r \leq i/2$ is the current level, where $i/2$ is the maximum number of rearrangement operations (levels) that can be applied on $\pi$; $n$ is the number of genes $(25, 50, \ldots, 700)$; probability of crossover $Px = 0.8$, and probability of mutation $Pm = 0.1$.

**Table 1: Results for $n$ from 25 to 700, $i = 6$.**

| $n$ | Actual score | Min of min best score | Avg of min best score | Avg of avg best score | Total time (s) |
|---|---|---|---|---|---|
| 25 | 14.5000 | 16.0000 | 30.3333 | 31.8748 | 1.0291 |
| 200 | 18.0000 | 22.0000 | 46.0000 | 47.1142 | 2.0563 |
| 700 | 18.0000 | 22.0000 | 46.6667 | 48.0960 | 6.0468 |

**Table 2: Results for $i = 2$, 6 and 10, and $n = 50$ .**

| $i$ | Actual score | Min of min best score | Avg of min best score | Avg of avg best score | Total time (s) |
|---|---|---|---|---|---|
| 2 | 5.5000 | 6.0000 | 13.6667 | 14.3879 | 0.7891 |
| 6 | 17.0000 | 19.0000 | 39.0000 | 40.0976 | 1.0124 |
| 10 | 28.5000 | 36.0000 | 62.3333 | 63.6603 | 1.4708 |

Two sets of experiments are conducted, where accuracy and time performance are reported for each. One set is when varying value of $n$, while keeping $i$ fixed. The other is when varying value of $i$, while keeping $n$ fixed. Experiments are repeated 10 times for each of the three sets of taxa and the best score is calculated for each experiment. The results are summarized in Tables 1 and 2. The results show that the minimum of the minimum best score is close to the actual score. Table 1 shows that the maximum time was 6 seconds when $n = 700$, but the time taken is increasing linearly with $n$. Table 2 also illustrates a linear time performance with $i$ for $n = 50$, and a maximum of only 1.5 seconds for $i = 10$.

## 4. CONCLUSION

We proposed a new hybrid heuristic approach for solving the median problem, namely *BeamGA*. The accuracy of BeamGA for the median score is excellent when compared to the actual score. Also the approach shows a very good time performance. For more accurate assessment, the performance needs to be tested on real biological data. The technique can also to be used to solve the median problem using any distance measure, including transposition.

## 5. REFERENCES

[1] M. Bernt, D. Merkle, and M. Middendorf. Solving the Preserving Reversal Median Problem. In *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, volume 5(3), 2008.

[2] D. Furcy and S. Koenig. Limited discrepancy beam search. pages 125–131., 2005.

[3] A. Siepel. Exact algorithms for the reversal median problem. Master's thesis, University Of New Mexico, Albuquerque, New Mexico., 2001.