An Efficient Probabilistic Population-Based Descent for the Median Genome Problem

Adrien Goëffon  
INRIA Bordeaux Sud-Ouest  
351 cours de la Libération  
33400 Talence, France  
goeffon@labri.fr

Macha Nikolski  
CNRS / LaBRI  
Université Bordeaux I  
351 cours de la Libération  
33400 Talence, France  
macha@labri.fr

David J. Sherman  
INRIA Bordeaux Sud-Ouest  
351 cours de la Libération  
33400 Talence, France  
david@labri.fr

ABSTRACT
We present a novel population-based local search algorithm for the median genome problem. The primary result of this article is that this probabilistic approach significantly improves the performance of ancestral genome reconstruction compared to existing methods, making it possible to tackle problems where the contemporary genomes may contain many hundreds of markers. Moreover, our method is not limited to triples of genomes, and thus solves the median genome problem in its generality. We show that in real application cases the computational results are highly robust, suggesting that we can interpret the computed median genomes as candidates carrying the semantics of ancestral architectures.

Categories and Subject Descriptors
I.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search—Heuristic methods; J.3 [Life and Medical Sciences]: Biology and genetics

General Terms: Algorithms.

Keywords
Median Genome Problem, probabilistic neighborhood, local search.

1. INTRODUCTION
The increasing availability of fully sequenced genomes has fuelled efforts in understanding the history and function of genomes through comparison of related species and analyses in computational biology. Identifying ancestral genome architectures is one important question that can now be addressed thanks to recent advances in combinatorial methods and to recent acquisition of large-scale genome datasets. Constructing plausible hypotheses about these ancestral architectures is a computational task whose results may provide deep insight both into the past histories of particular genomes and the general mechanisms of their formation. This task is frustrated by the computational complexity of the problem as well as by the difficulty of integrating biological constrains in the algorithms without altering completely their mathematical foundations.

Mathematically, genome architectures are encoded as signed permutations of common markers. The goal of this encoding is not to align one genome against the other, but rather to compare the order of gene markers. Two main approaches to compare marker orders exist: counting the differences between two genomes in terms of breakpoints [12, 15], and counting the minimal number of edit operations that transform one genome into another [8]. Both of these approaches define a distance function on the space of signed permutations. In this paper we will follow the Hannenhalli and Pevzner approach [8] where the allowed edit operations are fusion, fission, reciprocal translocation and reversal. It was originally established that this rearrangement distance can be computed in polynomial time [8, 10]. Later work has improved these results establishing a linear-time algorithm for the rearrangement distance computation [11]. Minimizing this distance is a fundamental step in computing plausible ancestors.

Two computational approaches for construction of ancestral genome architectures were proposed. They were formulated as the Median Genome Problem (MGP) and the Multiple Genome Rearrangement Problem (MGRP). Given a set of genomes \( \{\Pi_i\} \), the former consists in computing a permutation \( \Pi \) minimizing the sum of distances to \( \{\Pi_i\} \), while the latter aims at computing the Steiner tree thus minimizing the sum of distances along its edges. The Median Genome Problem has been shown to be NP-hard for both of distance functions even in the case of only 3 genomes (see [4, 14] for breakpoint distance and [5, 6] for rearrangement distance). Nevertheless, exact resolution of MGP have been attempted, yielding optimal solutions for very small instances [17].

Approximate algorithms for MGP have been proposed for both distances. In the breakpoint case, Sankoff and Blanchette formulated the solution through a reduction to the Travelling Salesman Problem [16]. These authors proposed an algorithm that guarantees a reasonable lower bound on the sum of distances.

In the case of rearrangement distance, MGP and MGRP have been tightly linked since for real-sized cases the proposed solutions for the latter rely on successive triangulations, and thus on the solving of the former in the 3-genome case. Two existing software packages MGR [3] and rEvoluzer...
implement partial solutions of MGP. Indeed, MGR solves the problem for triples of multichromosomal genomes, while rEvolutzer can treat more than 3 genomes, but only in the unichromosomal case. Both of the proposed solutions are heuristics based on the detection of “good” reversals, operations that are guaranteed to improve the solution. The genomes \( \{ \Pi_i \} \) are re-written step by step by applying reversals until two (or three) of them become equal. There are two differences in the proposed solutions. First, the definition of what constitutes a good reversal is not exactly the same. Second, when no good reversals remain, MGR performs a k-depth search to find a best reversal, while rEvolutzer allows for backtracking.

In this paper, we present FAUCILS, a new approximate algorithm for MGP in the general case, that is, for an unrestricted number of multichromosomal genomes, while improving performances of existing approaches on restricted instances. The main originality of our approach is the definition of a probabilistic neighborhood which evolve within a population-based local search according to observations made on the population. This mechanism allows us to greatly accelerate the search and ensures more convergence, especially for real or structured instances.

2. THE MEDIAN GENOME PROBLEM

A chromosome \( \pi = (\pi_1, \ldots, \pi_m) \) is represented by a sequence of signed gene markers whose sign indicates their relative direction on the chromosome. A size-\( n \) multichromosomal genome \( \Pi \) is defined as a set of chromosomes \( \{ \pi^1, \ldots, \pi^N \} \) such that \( \sum_i |\pi^i| = n \). Markers take their value from the set of ordinals \( 1, \ldots, n \); no given marker appears in more than one chromosome.

For example, \( \{(5, -8), (1, 2, -10, 6, 4), (9), (-3, 7)\} \) is a genome of size 10. In \( \Pi \), concatenation of all chromosomes is represented as a signed permutation.

Given a set of size-\( n \) genomes \( \{ \Pi_i \} \) and a genome distance function \( d \), an instance of the combinatorial minimization problem \( MGP \) is defined by two elements \( (\tau_n, \phi) \):

1. a search space, \( \tau_n \), composed of the set of all possible size-\( n \) genomes, and
2. an objective function \( \phi : \tau_n \rightarrow \mathbb{N} \) (score) defined by \( \phi(\Pi) = \sum_i d(\Pi, \Pi_i) \).

A median genome for a given set of genomes \( \{ \Pi_i \} \) is a genome \( \Pi \) that minimizes \( \phi(\Pi) \). Every optimal solution to \( MGP \) is a median genome.

3. AN ORIGINAL POPULATION-BASED LOCAL SEARCH FOR MGP

For addressing NP-complete problems like MGP in the general case and reaching acceptable solutions in reasonable time, approximate algorithms provide the most practical approach. We present a population-based local search algorithm using an original and evolutive neighborhood reduction mechanism for the resolution of MGP in the case of rearrangement genome distances. It gives excellent results in terms of the quality of the solutions it obtains, the speed of the computation, its robustness, and its scalability.

Figure 1: Geometric analogy of MGP: median genomes are within the convex hull of \( \{ \Pi_i \} \) in the space of genomes. Starting the search from a \( \Pi \), on the perimeter greatly reduces the search space.

3.1 A descent algorithm for MGP

Stochastic Local Search (SLS) \( [9] \) is a well-known class of metaheuristics, used for the resolution of many difficult combinatorial optimization problems. SLS algorithms are iterative methods which start from an initial configuration (candidate solution of the search space) and improve it by successive local modifications. In this section we define a simple descent algorithm to MGP, where:

1. the initial configuration is taken from \( \{ \Pi_i \} \),
2. the evaluation function is the same as the objective function \( \phi \): the rearrangement distance \( d \),
3. the neighborhood relation we call \( R^1 \) is a 1-step rearrangement: \( R^1(\Pi) = \{ \Pi' \in \tau_n, d(\Pi, \Pi') = 1 \} \),
4. the move strategy is a first-improve selection (FI) which accepts better and equivalent configurations (side-walk mechanism, SW \( [16] \)), given a specified number of iterations \( nbit \).

The performance of a descent algorithm essentially depends on the neighborhood relation used \( \Pi \). In order to avoid slow processes and local optima difficulties, we use the FI+SW selection strategy combined with the large and straightforward neighborhood \( R^1 \). Figure 1 shows that configurations taken from \( \{ \Pi_i \} \) may be interesting initial candidates for the beginning of the search. Considering the evaluation function and the neighborhood relation \( R^1 \), the descent will explore only configurations from the schematic area delimited by the \( \Pi_i \). If the \( \Pi_i \) are close (for example in the case of real applications), then the resulting search space is significantly reduced.

3.2 Probabilistic population-based local search

Traditionally, descent algorithms are sensitive to either stochastic factors or initial configurations and consequently may not be sufficiently robust – that is, different executions may diverge – although the SW mechanism and the use of a large neighborhood can reduce this drawback. A commonly used solution is to perform several descents from different initial configurations (different replications in a multi-start descent process). In genetic local search algorithms, local
search processes and crossovers between elements (individuals) of a set or multiset of current configurations (population) provide intensification and diversification phases.

A local search process applied to many independent replications is sometimes called a population-based local search even though there is no interaction between individuals [13]. Here we do not use any crossover operations, but simulate an alternative evolutionary process in order to accelerate the searches and to make multi-start descents more convergent.

We introduce a probabilistic population-based local search algorithm which favours, at each step of the search, the selection of most pertinent neighbors [7] with respect to the population. Structural information about each individual is used to estimate a selection probability at each step of the search. In this process, all replications are dependant, while the descents are carried out simultaneously.

In this section we present a multi-start descent for MGP. We use the descent mechanism presented in section 3.1 adding to each neighbor a selection probability.

Let \( \mathcal{P} \) be the population of our population-based descent, which initially contains individuals taken from \( \{ \Pi_1, \ldots, \Pi_l \} \). Now let us consider a probabilistic function \( p : \tau_\Pi \times \tau_\Pi \times \tau_\Pi \to [0, 1] \), such that \( p(\Pi, \Pi', \mathcal{P} \setminus \{ \Pi \}) \) gives a selection probability of \( \Pi' \in R^d(\Pi) \). Such a probabilistic function is quite similar to the one used for simulated annealing move strategy. The difference here is that only better or equivalent neighbors are accepted by the move strategy, whereas neighbors are generated by a probability distribution (probabilistic neighborhood [11]). The aim is not to escape to local optima, but to favor neighbors which share properties with other individuals in \( \mathcal{P} \).

This probabilistic function is connected to the notion of adjacencies, that we define in the way analogous to Nadeau and Taylor [12].

**Definition 1.** Two consecutive elements \( \pi_i \) and \( \pi_{i+1} \) of a chromosome \( \pi \in \Pi \) are said to be adjacent in \( \Pi \). We note this adjacency by \( (\pi_i, \pi_{i+1}) \).

We consider additional adjacencies at the extremities of each chromosome by introducing marker 0. For a chromosome \( (\pi_1, \ldots, \pi_n) \), two adjacencies are added: \((0, \pi_1)\) and \((\pi_n, 0)\). Notice that \( (\pi_i, \pi_j) = (-\pi_j, -\pi_i) \) and \((0, \pi_1) = (-\pi_0, 0)\). Finally, we note \( A(\Pi) \) the set of all adjacencies in \( \Pi \). We have \( |A(\Pi)| = n + N \) (\( n \) is the number of markers and \( N \) the number of chromosomes).

Each move (rearrangement) breaks one (fission) or two (reversal, fusion, reciprocal translocation) adjacencies. The probabilistic neighborhood encourages adjacencies which are not or are less, represented in the population to be broken.

The probabilistic neighbor selection operates as follows: let \( \Pi' \in R^d(\Pi) \); if \( \phi(\Pi') \leq \phi(\Pi) \), then \( \Pi' \) replaces \( \Pi \) in \( \mathcal{P} \) in function of the proportional representation of the broken adjacencies in \( \mathcal{P} \setminus \{ \Pi \} \):

\[
p(\Pi, \Pi', \mathcal{P} \setminus \{ \Pi \}) = 1 - \frac{|\{ \Pi'' \in \mathcal{P} \setminus \{ \Pi \} : (A(\Pi) \setminus A(\Pi')) \cap A(\Pi'') \neq \emptyset \}|}{|\mathcal{P}| - 1}
\]

Algorithm 1 provides an overview of our probabilistic population-based descent we called FAUCILS for Fast Ancestor (inference) Using Convergent and Intelligent Local Search.

4. EXPERIMENTS

For experiments we use different kinds of instances: real and random ones, with different numbers of genes and chromosomes by genome.

4.1 Real instances

First we assess our algorithm FAUCILS on two sets of 10 triplets of yeast genomes. The data, provided by Généol- vures Consortium[http://www.genolevures.org](http://www.genolevures.org) (GDR CNRS 2354), consists in five sequenced yeasts from the Kluyveromyces clade: Kluyveromyces lactis (Killa), Saccharomyces kluyveri (Sakl), Zygosaccharomyces rouxii (Zyro), Ashbya gossypii (Ergo) and Kluyveromyces thermotolerans (Klth). From these data, two sets of permutations have been computed: the first one with 135 markers (K135), and the second one with 499 markers (K499). For the comparison with MGR, which calculates only 3-genomes medians \( N = 3 \), we separate in ten instances each possible triplet of genomes: Killa-Sakl-Zyro is K135-1 and K499-1, Killa-Sakl-Ergo is K135-2 and K499-2, ... These five genomes have respectively 6, 8, 7, 6 and 8 chromosomes. We add a real test instance composed by the genomes of Human, Cat and Mouse, and available on the MGR web page[http://nber.sdsce.edu/GRIMM/mgr.cgi](http://nber.sdsce.edu/GRIMM/mgr.cgi).

Table 4 shows performances of FAUCILS and MGR on these real instances. FAUCILS is a stochastic algorithm, and two executions may return different results; for each instance we perform multiple executions. Table 4 indicates the best results \( \phi_b \) of 20 executions, their frequency \( f \), the mean scores \( \phi_m \), the worst scores \( \phi_w \), the standard deviations \( \sigma \) and the mean computation times of one execution. FAUCILS was run with its default parameters: \( l = 3 \) (i.e. a population size of 9 when \( k = 3 \)), and one million LS iterations \( nbit \): MGR was first run with its default parameters, and secondly with the heuristic option H1 (MGR-H1) for speeding up the search. Each execution was performed on

\[\text{http://www.genolevures.org}\]
\[\text{http://nber.sdsce.edu/GRIMM/mgr.cgi}\]
### Table 1: Comparison between FAUCILS and MGR on real instances

<table>
<thead>
<tr>
<th>Instance</th>
<th>k</th>
<th>n</th>
<th>FAUCILS</th>
<th>MGR</th>
<th>MGR-H1</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>K135-1</td>
<td>3</td>
<td>135</td>
<td>178</td>
<td>151</td>
<td>133</td>
<td>-3</td>
</tr>
<tr>
<td>K135-2</td>
<td>3</td>
<td>135</td>
<td>150</td>
<td>100</td>
<td>133</td>
<td>-3</td>
</tr>
<tr>
<td>K135-3</td>
<td>3</td>
<td>135</td>
<td>140</td>
<td>133</td>
<td>133</td>
<td>0</td>
</tr>
<tr>
<td>K135-4</td>
<td>3</td>
<td>135</td>
<td>140</td>
<td>133</td>
<td>133</td>
<td>0</td>
</tr>
<tr>
<td>K135-5</td>
<td>3</td>
<td>135</td>
<td>140</td>
<td>133</td>
<td>133</td>
<td>0</td>
</tr>
<tr>
<td>K135-6</td>
<td>3</td>
<td>135</td>
<td>140</td>
<td>133</td>
<td>133</td>
<td>0</td>
</tr>
<tr>
<td>K135-7</td>
<td>3</td>
<td>135</td>
<td>140</td>
<td>133</td>
<td>133</td>
<td>0</td>
</tr>
<tr>
<td>K135-8</td>
<td>3</td>
<td>135</td>
<td>140</td>
<td>133</td>
<td>133</td>
<td>0</td>
</tr>
<tr>
<td>K135-9</td>
<td>3</td>
<td>135</td>
<td>140</td>
<td>133</td>
<td>133</td>
<td>0</td>
</tr>
<tr>
<td>K135-10</td>
<td>3</td>
<td>135</td>
<td>140</td>
<td>133</td>
<td>133</td>
<td>0</td>
</tr>
<tr>
<td>K499-1</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-2</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-3</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-4</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-5</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-6</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-7</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-8</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-9</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K499-10</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>K135</td>
<td>3</td>
<td>135</td>
<td>178</td>
<td>151</td>
<td>133</td>
<td>-3</td>
</tr>
<tr>
<td>K499</td>
<td>3</td>
<td>499</td>
<td>300</td>
<td>250</td>
<td>250</td>
<td>0</td>
</tr>
<tr>
<td>HCM</td>
<td>3</td>
<td>114</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Comparison between FAUCILS and MGR on real instances. \( \phi_b \) is the best score returned by FAUCILS, \( |f| \) is its frequency, \( \phi_m \) is the mean score, \( \phi_w \) is the worst score, \( \sigma \) is the standard deviation based on 20 different executions.

Due to the large number of instances, only the first 10 are listed. The full table is available in the supplementary material.

In order to assess the performance of our population-based local search algorithm with respect to the structure and the size of the instance, we generate two types of random instances.

4.2 Random instances

In order to assess the performance of our population-based local search algorithm with respect to the structure and the size of the instance, we generate two types of random instances.

First, we use completely random instances (R) containing a specified number of markers, and a minimum and maximum number of chromosomes by genome (N). On these instances of size 50 and 100, FAUCILS obtains better results than MGR systematically (see Table 2). For larger instances, only one MGR run ended, with an uncompetitive result (\( \Delta = -22 \)). These instances seem to be difficult because of their structure: each genome is a random point of \( r \), and the MGR algorithm seems very dependent on the structure of each instance (see the divergences between all computational times on tables 1, 2, and 3).

In order to estimate the impact of the structure of the instance, we generate simulated instances (S), for which distances between genomes are bounded. An arbitrary ancestral genome is generated from which a specified number of random rearrangements are applied to give three genomes. We specify the number of genes (n) and chromosomes (N), and the number of rearrangements done during the simulation (r); this parameter is an upper bound of the optimal median genome score.

The results are given in Table 3. We can see that, with \( r = 10 \) or \( r = 50 \), instances are very easy to solve. But when the distances between genomes increase (\( r = 100 \) and \( r = 200 \)), FAUCILS is very competitive and can find in short computation time solutions considerably better than MGR. Moreover, the algorithm is robust as small values of \( \sigma \) show. For these instances (S), we have to reduce the number of local search iterations to 2000, for an equivalent efficiency.

The evolution of the ratio \( \phi_r \) gives an empirical indication of the difficulty of the structure of the search space. Indeed, for \( r = 200 \), the minimal number of rearrangements required for reconstructing an evolutionary scenario is about 25% lower than the number of rearrangements made during the simulation. Adding to the relative difficulty to find near-optimal genomes for these instances, we can presume that this ratio represents the quantity of lost information and can be a good indicator for comparing the difficulty of simulated instances.

Finally, we have executed rEvoluzzer [2] on each unichromosomal instance: S100-10-1, S100-50-1, S100-100-1, S100-200-1, R50-1, R100-1, R200-1, R-500-1. Except for the three first instances, where rEvoluzzer founds in few seconds or minutes the same scores as FAUCILS (10, 50, 95), the program did not return any solution for the five other instances, even given one week of computation.

---

[https://www.grid5000.fr](https://www.grid5000.fr)
4.3 Influence of the probabilistic neighborhood

One of the main originalities of FAUCILS is that neighbors are selected with a non-uniform probability. The foremost aim is to select more pertinent neighbors as a function of the similarities between individuals in the current population. Since the population is initialized by the given genomes (instance), the probabilistic selection will have a larger impact on structured instances, that is when genomes share adjacencies; it is notably the case of real data instances.

Figure 2 shows the evolution of the minimum score (in the left) and the average score of all individuals (in the right) during the search for the complete Kluyveromyces genomes instances (K135 at the top, K499 at the bottom). We compare the probabilistic descent (algorithm 1) to the same search without a probabilistic selection (in this case \( p(\Pi, \Pi', \mathcal{P} \setminus \{\Pi\}) = 1 \)) in Table 2. The key idea is to use a probabilistic neighborhood which evolves during the search according to the partial results of all descents performed simultaneously.

This study shows that this probabilistic population-based local search adds semantics to the search in order to reduce the neighborhood. It exploits the structure of the instance for a quick convergence of the population.

### Table 2: Comparison between FAUCILS and MGR on random instances

<table>
<thead>
<tr>
<th>Instance</th>
<th>n</th>
<th>N</th>
<th>FAUCILS</th>
<th>MGR</th>
<th>MGR-H1</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>R50-1</td>
<td>50</td>
<td>1</td>
<td>78</td>
<td>82</td>
<td>80.6</td>
<td>82</td>
</tr>
<tr>
<td>R50-2</td>
<td>50</td>
<td>1</td>
<td>90</td>
<td>5/20</td>
<td>80.8</td>
<td>82</td>
</tr>
<tr>
<td>R50-3</td>
<td>50</td>
<td>3</td>
<td>80</td>
<td>5/20</td>
<td>80.2</td>
<td>82</td>
</tr>
<tr>
<td>R50-4</td>
<td>50</td>
<td>5</td>
<td>79</td>
<td>5/20</td>
<td>80.2</td>
<td>82</td>
</tr>
<tr>
<td>R100-1</td>
<td>100</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R100-2</td>
<td>100</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R100-3</td>
<td>100</td>
<td>3</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R100-4</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R200-1</td>
<td>200</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R200-2</td>
<td>200</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R200-3</td>
<td>200</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R500-1</td>
<td>500</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R500-2</td>
<td>500</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R500-3</td>
<td>500</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 3: Comparison between FAUCILS and MGR on simulated instances

<table>
<thead>
<tr>
<th>Instance</th>
<th>n</th>
<th>N</th>
<th>div</th>
<th>FAUCILS</th>
<th>MGR</th>
<th>MGR-H1</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>S100-10-1</td>
<td>100</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-10-2</td>
<td>100</td>
<td>5</td>
<td>10</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-10-3</td>
<td>100</td>
<td>10</td>
<td>10</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-50-1</td>
<td>100</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-50-2</td>
<td>100</td>
<td>5</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>S100-50-3</td>
<td>100</td>
<td>10</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>S100-100-1</td>
<td>100</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-100-2</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-100-3</td>
<td>100</td>
<td>10</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-200-1</td>
<td>100</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-200-2</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>S100-200-3</td>
<td>100</td>
<td>10</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this paper we proposed a new efficient algorithm for the resolution of the Median Genome Problem (MGP) in the general case. We have notably introduced an novel way for speeding up and making more convergent multi-start descents for the resolution of MGP, especially for real structured instances. The key idea is to use a probabilistic neighborhood which evolves during the search according to the partial results of all descents performed simultaneously.

Experiments realized both on real and random instances show that our software FAUCILS is able to find largely better solutions than MGR, the current reference in the domain. Moreover, this local search approach is very fast and scalable: contrary to other existing techniques, FAUCILS can...
treat an unbounded number of multichromosomal genomes, which may contain hundreds or thousands of markers. Future work will involve finding ways to evaluate the quality of solutions in the case of big instances, and to extend this MGP algorithm for the resolution of MGRP. The Median Genome Rearrangement Problem is a very hard computational problem for the resolution of which existing algorithms calculate multiple median genomes.

6. ACKNOWLEDGMENTS

We would like to thank Guillaume Bourque and Matthias Bernt for their helpful comments and for providing us their software MGR and rEvoluzer. We thank Géraldine Jean for the construction of the real signed permutations (Kluyveromyces genomes), and the Genolevures Consortium for making these data available.
Experiments presented in this paper were carried out using the Grid’5000 experimental testbed, an initiative from the French Ministry of Research through the ACI GRID incentive action, INRIA, CNRS and RENATER and other contributing partners (see https://www.grid5000.fr).

7. REFERENCES


