An Iterated Local Search Algorithm for the Clonal Deconvolution Problem

Maitena Tellaetxe-Abete
Molecular Oncology group
Biodonostia
Donostia-San Sebastian, Spain
maitena.tellaeche@biodonostia.org

Borja Calvo
Intelligent Systems group
University of the Basque Country UPV/EHU
Donostia-San Sebastian, Spain
borja.calvo@ehu.es

Charles Lawrie
Molecular Oncology group
Biodonostia
Donostia-San Sebastian, Spain
charles.lawrie@biodonostia.org

***Abstract*—Cancer is a disease characterized by the continuous acquisition of random mutations by cells, which are subsequently subjected to selection forces that favour the survival of some cells over others. The result of this evolutionary process called clonal evolution is a genetically heterogeneous mass known as a tumor, and identifying its composition is crucial not only for gaining further understanding of the disease, but also for designing effective therapies tailored to the particularities of the whole tumor. Thus, the clonal deconvolution problem tries to identify the different cell subpopulations that form the tumor and the phylogenetic tree that describes the evolutionary process that led to it from a series of biopsies that are an admixture of those subpopulations. This problem has been tackled from different perspectives, but as far as we know, metaheuristics have not been explored. In this article, we propose an Iterated Local Search (ILS) approach as a first metaheuristic approximation to solve this problem. Preliminary results on simulated data show that our method outperforms two well-established heuristic algorithms when running time is constrained. Moreover, the algorithm has the advantage that it is a flexible approach in which assumptions on the tumor development mode are not directly implemented, and it can therefore be easily adapted to accommodate new discoveries made on the evolution mechanisms in cancer.**

***Index Terms*—Optimization, Metaheuristics, Heuristic methods, Biology and genetics, Local search**

I. Introduction

Tumors grow through the accumulation of somatic mutations that provide them with fitness advantage. This accumulation can be described by the clonal evolution theory, according to which selection forces such as oxygen availability, physical forces of the microenvironment or therapy act by favouring the growth or survival of certain cell populations or clones over others [1]. As a result, tumors present themselves as genetically heterogeneous masses composed of clones that harbour different mutational profiles that shape their capacities, including growth rate, immunogenicity, response to treatment and ability to metastatize. This characteristic of the tumors is known as intra-tumor heterogeneity (ITH) and its study is essential not only for better understanding cancer development but also for the clinical practice, by aiding in the design of therapies adapted to the particularities of the tumor clones.

The fact that clones are not independent from each other, but rather are evolutionarily related, allows us to model the tumor history through a phylogenetic tree. Such a tree is composed of vertices or nodes that represent tumor clones and edges that represent ancestral relationships, and is usually rooted in the cell where the first mutation of the tumor arose, i.e., the most recent common ancestor of all the tumor clones. All these nodes represent clones that have arisen at some point in the evolution of the tumor. Some of those may be extinct, i.e., not present anymore in the tumor. The rest are potentially observable, but not all of them are necessarily observed in every biopsy.

Inferring such a tree, however, is rather different from the classic phylogenetic tree reconstruction problem [2]. In the classical problem, we infer a phylogeny from a collection of organisms which we observe individually; here, instead, we do not directly observe the clones, but mixtures of them, as detailed below.

For getting the input data for our problem, the procedure is to biopsy a series of samples from the tumor under study and to sequence them to obtain the set of existing mutations together with their estimated frequency or variant allele frequency (VAF) in each sample – i.e., which percentage of the sample contains each mutation. Most of the time, however, each of these samples does not contain a single clone, but is rather an admixture of clones [3]. Hence, if we aim to identify the clones present in a tumor, we must take into account the fact that these VAFs we observe are not direct estimates of the clone frequencies, but the result of a pool of them. What is more, additional bias exists in these measurements, including the ones introduced by sampling and the sequencing process itself.

This whole fact is indeed the origin of the problem dealt with in this work, as the evolutionary reconstruction is bounded to the deconvolution of the mixtures of clones. In essence, the problem we need to tackle is the identification of the clonal structure of the tumor, including the number of clones, their mutational composition and the phylogenetic tree leading to that clone mosaic, having as an input the estimated frequency of a number of mutations in a number of samples of the tumor. While several names have been given to this problem in the literature, we will refer to it as clonal deconvolution problem (CDP) (Fig. 1).

During the last few years, several efforts have been made to address the CDP. A large number of probabilistic (mainly Bayesian) models have been proposed as they are a natural

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

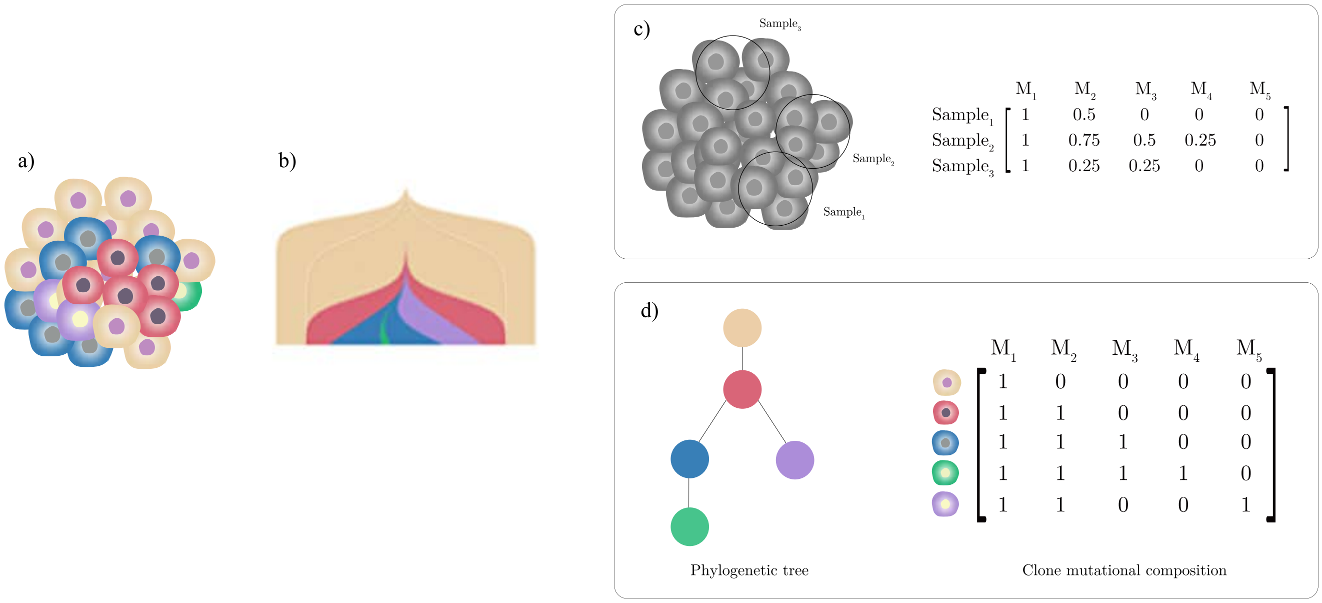


Fig. 1. Illustration of the clonal deconvolution problem. The toy tumor on a) has 5 clones, each of them represented with a different color, and has evolved as depicted in b). The input for the CDP is shown in panel c). Three samples are biopsied from the tumor, but the clones present in each sample are unknown. Instead, the information we get is the set of the identified mutations (M_1-M_5), together with their variant fractions in each sample. These are detailed in the matrix on the right. One possible solution to the CDP is shown in d), and consists of a phylogenetic tree that explains the ancestral relationships between the clones and the mutational composition of each clone. Several factors interfere with the resolution of the problem. Among them, it is worth noting that, in this specific example, the sampling does not capture any purple clone and hence, none of the possible phylogenies constructed from these data, including the one shown in d), could contain this clone.

approximation to the problem that provides an inherent way of dealing with noisy data and modelling uncertainty [4], [5], [6], [7], [8]. Still, they run into too large computing times for practical purposes, especially as the problem size increases. This has only become of special relevance recently, as the lowering of DNA sequencing costs has allowed more tumor samples to be sequenced and in much more detail, which in turn has enabled the detection of larger numbers of low frequency mutations.

At the same time, the combinatorial nature of the problem opens the door to the use of optimization techniques to solve the problem. Here, most of the proposed approaches fall into the exact [9], [10] or heuristic [9], [11] categories. Whereas exhaustive methods provide the best results, they suffer from the same limitations as probabilistic modelling and are only suitable for problems with a limited sample size and number of mutations. Regarding the heuristic methods, while they are able to produce good enough solutions quickly, they have the limitation that they are created with problem-specific knowledge. For solving the CDP, these algorithms impose restrictions based on assumptions about the mechanisms behind tumor evolution. Commonly used evolution models are, however, being questioned by recent evidence in favour of new ones [12], and there is not an easy way of introducing these latest models into heuristic algorithms without redesigning them completely. Moreover, the way most of the existing methods have to cope with large amounts of mutations is by assuming that those mutations with similar VAFs appeared for the first time in the same clone, and hence they first cluster together those mutations and treat them as a single mutation [13], [14], [15]. However, this simplification may not always be true, as two ancestrally unrelated clones can each harbour a mutation with similar allele frequencies, which is indeed especially common in low frequency variants. Thus, these mutations would incorrectly be grouped together in the same clone by these approaches.

Hence, a need exists to develop algorithms that are able to produce solutions for big problem sizes without the need to make constraining assumptions on allele frequencies, and that are flexible enough to allow for changes in their formulation as new knowledge on tumorigenesis is achieved. Under this scenario, metaheuristics are an approach worthy of study. However, these types of algorithms have largely been overlooked in this field.

Here, we base our work on an existing formulation of the CDP and introduce an Iterated Local Search (ILS) algorithm as a first metaheuristic approximation to the problem. This is a flexible approach that can equally work under different uncertainty conditions and tumor evolution models. Moreover, experiments on simulated data show competitive results when compared to other established algorithms.

The remainder of the paper is organized as follows: the formulation of the CDP is introduced in Section II. Section III describes the algorithm we have implemented. The experimental setup is presented in Section IV and the results are reported and discussed in Section V. Finally, the conclusions and future research lines are presented in Section VI.

II. Problem formulation

In order to solve the CDP, hypotheses about the mutation acquisition process that leads to the tumor development are to be made. Although one of the main advantages of our approach is the flexibility, in this work we will pay attention to the classical evolution models that are used in most of the previous works. Thus, we assume tumors have monoclonal origin, i.e., they arise from a single cell or mutation, and attach to the perfect phylogeny model or infinite sites assumption (ISA), which states that a given mutation arises at most once in the same tumor, and that a mutation can not be lost. This model, although fairly restrictive, has widely been adopted by several authors [16], [14], [13] and hence stands as a reasonable starting point. Its main implications are two. First, if two cells share a given mutation, then they have to be part of the same clone or they are part of clones ancestrally related. Secondly, if a cell in a clone acquires a given mutation, then that mutation has to be also present in all its descendants, as mutations cannot disappear.

Our algorithm tries to solve the problem based on the Variant Allele Frequency Factorization Problem (VAFFP) formulation introduced in [14]. The intuition behind this formulation is that the mutation frequencies we observe in a series of tumor samples are the result of the combination of the tumor clonal structure and the clone proportions captured in each sample. This formulation can be expressed by means of a matrix decomposition procedure.

Let us suppose that we have m samples from a given tumor and we sequence them. Now, let n be the set of the mutations identified in at least one sample of the tumor. We can now construct an $m \times n$ VAF matrix F, where f_{ij} is the frequency of mutation j in sample i.

The tree that describes the development of the tumor can be defined by a rooted, directed clone tree T with n vertices. In this tree, each vertex corresponds to a clone and is identified by exactly one mutation, meaning that the clone is the first to contain that mutation. This implies that, under this formulation, we can interchangeably talk about clones or mutations as there is a one-to-one correspondence between them. In this tree, the edges connect clones with a direct ancestral relationship. For the tree to be in line with the ISA model, for each two directly connected vertices $v_j \to v_k$, v_k contains all the mutations of its parent v_j (no mutation loss) and it has an additional mutation only present in itself and its descendants (mutations arise only once).

Alternatively, we can represent this tree by an $n \times n$ binary clone genotype B matrix [17], in which each $b_{i.}$ row represents the mutations present in vertex (clone) v_i. ISA restrictions on the tree translate into the B matrix as follows: for any mutation j , let $\mathcal{I}_j$ be the indices of the clones containing that mutation. That is, $\mathcal{I}_j = \{i/b_{ij} = 1\}$. Then, for every pair of columns j, k, either $\mathcal{I}_j$ and $\mathcal{I}_k$ are disjoint or one contains the other. Note that, assuming a monoclonal origin for the tumor, the matrix will have a column with all ones corresponding to this founding mutation, as that initial mutation is in all the clones. We should note that, given a valid B matrix, any matrix obtained by a permutation of the rows (clones) and/or columns (mutations) leads to an equivalent matrix that represents the same evolutionary process. This property implies that any relabelling of the mutations and/or clones does not alter the tree structure, and implies that, for a tree T, we have $n! \cdot n!$ equivalent matrices.

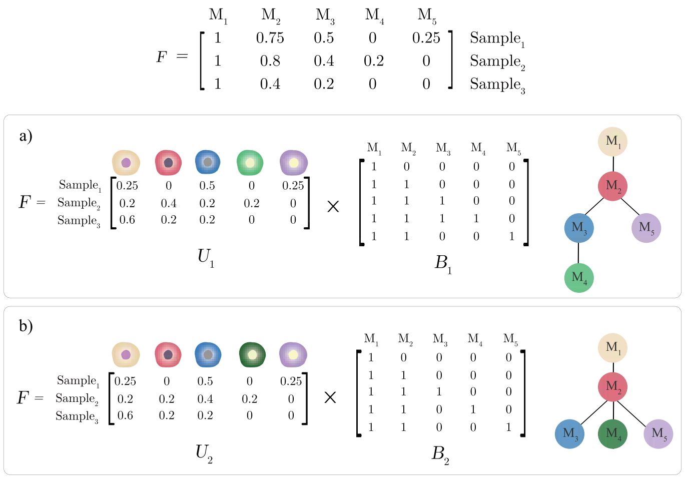


Fig. 2. The VAFFP problem formulation. This problem looks for the decomposition of a F matrix into a clone frequency matrix U and a clone genotype matrix B. Two possible solutions for the F matrix above are shown in panels a) and b). Both solutions have four clones in common (M_1, M_2, M_3, M_5) and differ in a single one (M_4): while the solution in a) places it as a direct descendant of M_3, in the solution in b) its parent node is M_2.

In order to derive the mutation frequency matrix F from a matrix B, we need information about the proportion of each clone in each sample. This information is captured in the clone frequency matrix U, which is an $m \times n$ matrix where u_{ij} is the fraction of clone j in sample i[1]. Subsequently, as shown by [14], we have that:

$$F = U \cdot B \tag{1}$$

As u_{ij} is a proportion, it follows that, first, its values have to be non-negative and secondly, rows, which represent clone proportions in each sample, must sum up to one. These conditions form what is known as the sum rule [14].

Thus, we can state the CDP as follows: given an $m \times n$ VAF F matrix, find a pair of matrices B and U that produce the observed F matrix by (1) (Fig. 2). Note that, in an error-free VAF scenario, i.e., when the F matrix does not contain errors, the exact F matrix can be found; however, when errors exist, the problem is relaxed to find the F' matrix that minimizes some distance to the observed F matrix. This problem has been shown to be NP-complete [14].

[1] Note that, regarding the relabelling issue, from all the permutations of the columns (mutations) we can restrict to that in the F matrix. Note also that the relabelling of the rows (clones) has to be taken into account in the U matrix in order to get the original F matrix.

III. Algorithm

In this work we propose an ILS algorithm to solve the CDP. The algorithm performs the search in the space of trees that fulfill the ISA. Loosely speaking, the search starts from a random valid tree. Then, neighbouring solutions (trees) are evaluated and the search continues with a solution that improves the current solution. If there is no option to move to an improving solution, a perturbation is performed and the search continues from there. The process goes on until any of the two stop criteria is met. As the evaluation function is lower-bounded at 0, the first criterion is to reach to an optimal solution. The second criterion is based on the budget of the algorithm in terms of a maximum number of evaluations.

Given a tree T and the F matrix, we first get one of the B matrices that represent T and calculate the necessarily unique U matrix using (1). This is straightforward as B is always an invertible matrix [14]. We next obtain the U' matrix from U by enforcing it to meet the aforementioned requirements: all negative values are coerced to 0 and rows are normalized to sum up to 1. It is shown that, for the U matrix to be valid, it is sufficient that it adheres to that criteria [14]. We then calculate the F' matrix with the U' matrix and the B matrix using (1) again. Finally, we calculate the objective function value as the mean absolute error between F and F':

$$e_{F'} = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \mid f_{ij} - f'_{ij} \mid \tag{2}$$

Algorithm 1 and Algorithm 2 respectively describe how the evaluation of a solution and the obtention of a B matrix for a given tree are computed. It is worth mentioning that the pseudocode in Algorithm 2 describes the calculation of a B matrix in which clones have been labelled so that clone i is the first containing mutation i, and that it is ordered so that row i represents clone i and column i represents mutation i.

The neighbourhood of a solution is defined as the collection of all trees reachable within one subtree prune and regraft (SPR) operation on that tree [18], which consists of cutting (or pruning) one edge in the tree and attaching (or regrafting) the resulting subtree to another node in the remaining tree. Note that this simple movement in the tree implies updating a number of rows and columns in the corresponding B matrix (Fig. 3). The pseudocode for the obtention of the neighbourhood is shown in Algorithm 3.

As for the selection of neighbour solutions, we employ two strategies to conduct the local search: to explore the entire neighbourhood and move to the best improving solution (greedy approach) or to continue with the first improving solution (first-improvement approach).

When getting trapped in local optima, the search continues with a solution obtained through the perturbation of that local optimum. Specifically, we perturb the solution by making a number of random SPR changes in the solution tree.

IV. Experiments

As a first approach to assess our proposal against well established algorithms, we will work on error-free simulated data. In this section we will first describe how the data are generated and, then, the experimental design will be presented.

A. Simulated data

Each problem instance consists of a matrix F of mutation frequencies in a set of samples. This matrix is built using a pair of matrices B and U representing an evolution tree that fulfils the ISA and the relative frequencies of the clones in the samples, respectively. These matrices are generated as follows.

Given a number n of mutations, we create a tree by first assigning a random mutation to the root node. Then, for each of the remaining mutations, we create a new node, assign the mutation to it and randomly set the node as a child of an existing node in the tree. In order to meet the ISA model, each of the newly added nodes inherits all the mutations present in its parent node. For the U matrix, we create each row (proportions of each clone in a sample) by randomly sampling an exponential probability density distribution with the rate parameter set to 3. As typically not all clones are present in all samples, 20% of the elements in each row are randomly selected and forced to 0. Finally, clone proportions of each row are normalized so as they sum up to one. Finally, the F matrix is calculated using (1).

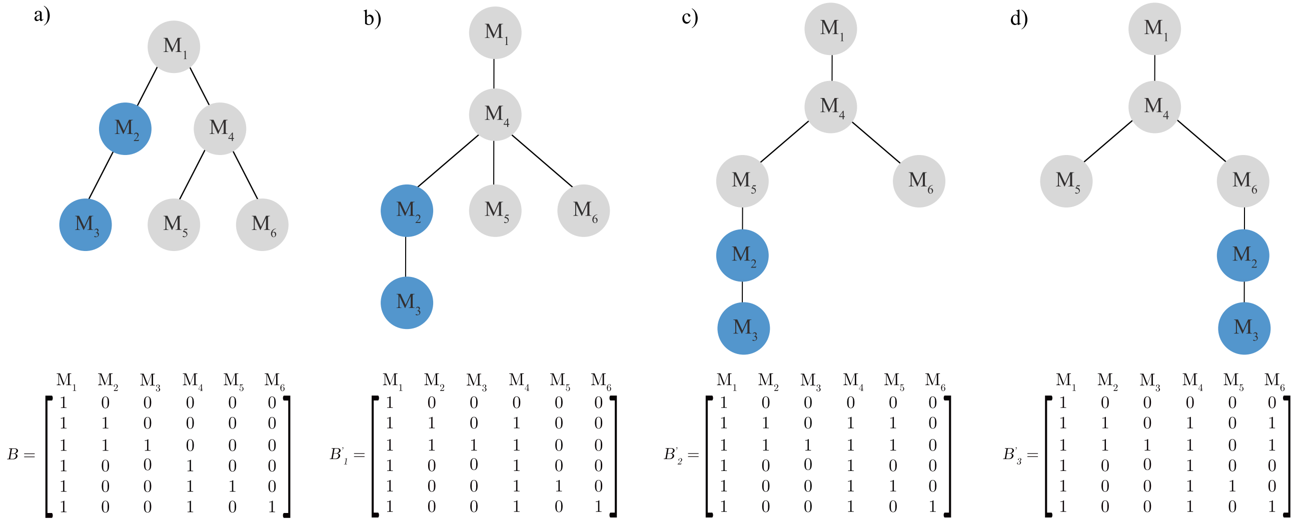
In total, we simulated 800 random instances. These comprise 4 different numbers of mutations ($n \in \{10, 25, 50, 100\}$) and 4 sample sizes ($m \in \{4, 6, 8, 10\}$), with the aim of assessing the behaviour of the algorithms with increasing sizes for the problem. For each combination, 50 instances were generated and the algorithms were run once for each instance.

B. Evaluation

We compared the performance of our algorithm to two other well established heuristic algorithms: CITUP [9] and LICHeE [13]. Evaluation was done by assessing how well the tools were able to reconstruct the original variant allele frequencies, in terms of the objective function value (2). In all cases, the running time of the algorithms was bounded to 2 hours and in the case of our algorithm, the maximum number of evaluations was limited to $10 \cdot n^2$.

The only parameter to tune in the ILS algorithm is the size of the perturbation. We performed some preliminary parameter tuning experiments, in which we found no big differences for different sizes (results not shown), so we set it to $0.2 \cdot n$.

Regarding the other two algorithms, there are several considerations to keep in mind. First of all, both can report more than one solution. For the current evaluation, the solution with minimum mean absolute error was chosen. Moreover, the two algorithms inherently cluster mutations with similar VAFs into a same clone. As our data simulation scheme does not perform any clustering but assigns a unique mutation to each clone, we tried to tune the parameters of the algorithms in order to minimize this clustering and hence make a fair evaluation. This worked for LICHeE, but posed problems with CITUP,



$$B = \begin{array}{c} \begin{array}{cccccc} M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$B'_1 = \begin{array}{c} \begin{array}{cccccc} M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$B'_2 = \begin{array}{c} \begin{array}{cccccc} M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

$$B'_3 = \begin{array}{c} \begin{array}{cccccc} M_1 & M_2 & M_3 & M_4 & M_5 & M_6 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \end{array}$$

Fig. 3. An example of the SPR operation. Panel a) shows the starting tree and an associated B matrix. Three possible SPR operations are allowed for the subtree rooted at M_2 (in blue): it can be regrafted as a child node of clone M_4, as shown in panel b); of clone M_5 as in c) or attached to clone M_6 as in d). In all cases, the elements associated to that moving subtree are updated in the B matrix. As aforementioned, these B matrices are just one of all the possible matrices for each tree. Note that the complete neighbourhood of the tree in a) is obtained by repeating this operation on the rest of possible subtrees.

```
Algorithm 1: Evaluate
    Input $Tree, F$
    Output $error$
  $B$ := Calculate B($Tree$)
  $U := F \cdot B^{-1}$
  for each $u_{ij}$ in $U$ do
      $u'_{ij}$ := max(0, $u_{ij}$)
  end
  for each $u'_{i.}$ in $U'$ do
      $u'_{i.}$ := normalize($u'_{i.}$)
  end
  $F' := U' \cdot B$
  $error$ := mean absolute error($F$, $F'$)
```

```
Algorithm 2: Calculate B
    Input $Tree$
    Output $B$
  $n$ := size($Tree$)
  Initialize identity matrix $B_n$
  for $i$ in $1 : n$ do
      for $j$ in the indeces of the descendant nodes of
        node $i$ do
          $B_{j,i} := 1$
      end
  end
```

```
Algorithm 3: Get neighbourhood
    Input $Tree$
    Output $Neighbourhood$
  Initialize empty vector $Neighbourhood$
  for each $node$ in nodes($Tree$) do
      Set $new\_parents$ to nodes($Tree$) not in parent
        node of $node$ and not in descendant nodes of
        $node$
      for each $new\_parent$ in $new\_parents$ do
          Build $new\_tree$ by setting parent of the subtree
            rooted at $node$ to $new\_parent$ //SPR move
          Add $new\_tree$ to vector $Neighbourhood$
      end
  end
```

where only default parameters worked in our instances. One last particularity about LICHeE should be mentioned. When this algorithm does not find a valid solution, it drops mutations and outputs a solution with fewer mutations than those in the F matrix. For assessing such cases, we decided to compute the mean absolute error in the subset of mutations contained in the output solution. Note that, although with this way of proceeding we alter the terms in the computation of the metric, as the error is a mean value the scale is kept and the results are comparable. Similarly, results obtained for instances of different sizes are also comparable.

V. Results and discussion

Fig. 4 and Table I summarize the performance of the four algorithms in the 800 instances of the problem described

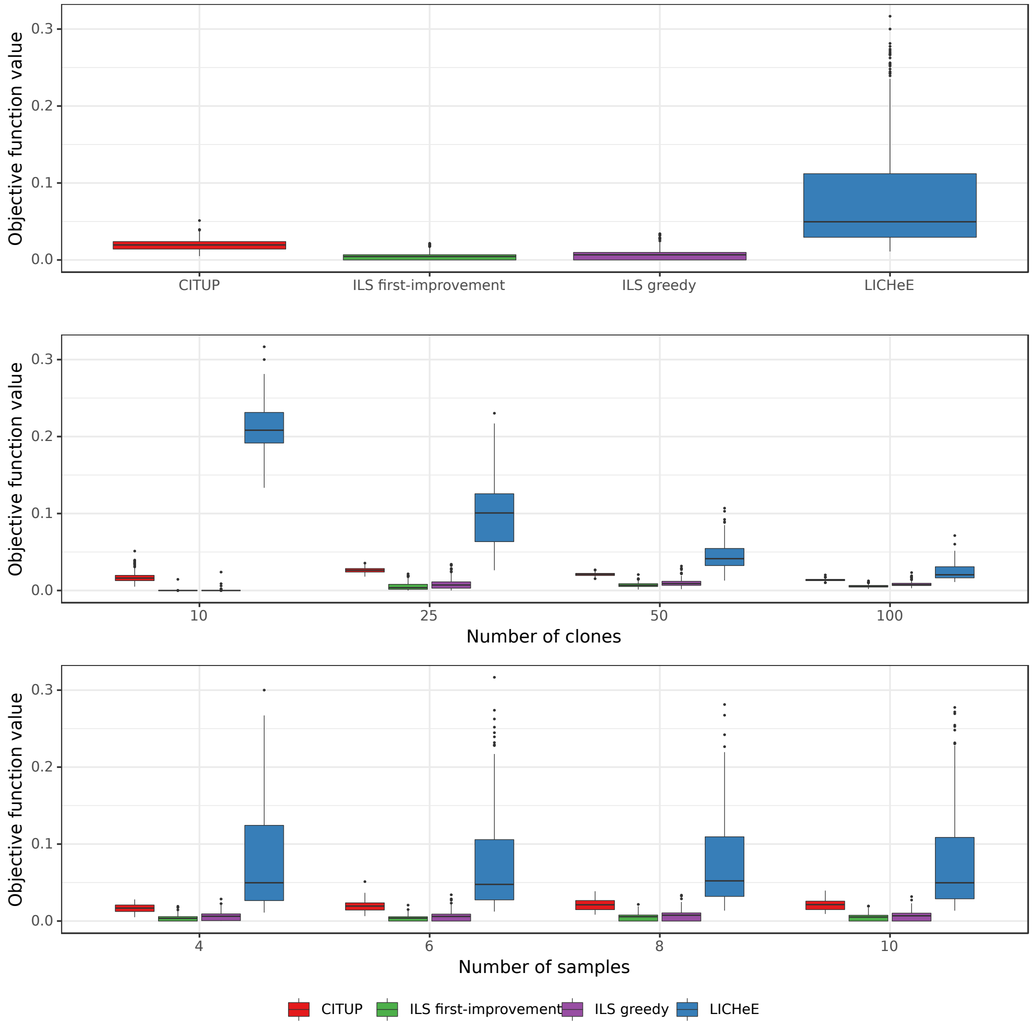


Fig. 4. Results for CITUP, LICHeE, ILS greedy and ILS first-improvement. On top, aggregated error summaries per tool are shown. Below we can see these results broken down by number of clones and by number of samples. ILS outperforms the other two algorithms in all the configurations. Results do not vary much for different numbers of samples; the number of mutations, in contrast, has a different effect on the performance depending on the approach. In the specific case of ILS, while it reaches optimal solutions for $n = 10$, it does not for larger mutation numbers.

TABLE I

OBJECTIVE FUNCTION MEDIAN VALUES OF THE SOLUTIONS OF EACH ALGORITHM, FOR A SERIES OF n MUTATIONS AND m SAMPLES

		ILS greedy	ILS first-improvement	CITUP	LICHeE
$n = 10$	$m = 4$	1.47e-17	1.40e-17	0.0123	0.201
	$m = 6$	1.58e-17	1.58e-17	0.0163	0.209
	$m = 8$	1.95e-17	1.89e-17	0.0185	0.206
	$m = 10$	2.05e-17	2.05e-17	0.0184	0.232
$n = 25$	$m = 4$	0.00680	0.00297	0.0222	0.0721
	$m = 6$	0.00467	0.00224	0.0252	0.0826
	$m = 8$	0.00954	0.00714	0.0277	0.103
	$m = 10$	0.00716	0.00508	0.0290	0.116
$n = 50$	$m = 4$	0.00782	0.00532	0.0191	0.0378
	$m = 6$	0.00753	0.00565	0.0207	0.0392
	$m = 8$	0.0104	0.00771	0.0215	0.0443
	$m = 10$	0.00991	0.00757	0.0220	0.0441
$n = 100$	$m = 4$	0.00741	0.00471	0.0124	0.0167
	$m = 6$	0.00812	0.00498	0.0138	0.0202
	$m = 8$	0.00758	0.00572	0.0141	0.0254
	$m = 10$	0.00723	0.00563	0.0137	0.0233

above. In the figure there are three boxplots. The top one shows the overview of the results obtained by the four algorithms (our proposal with first-improvement and greedy selections, CITUP and LICHeE) in all the instances. The other two boxplots summarize the results obtained grouped by number of clones and number of samples. Similarly, the table contains the median result obtained in the 50 instances generated for each combination of number of clones (n) and samples (m).

The first thing we can see in the results is that there are no 0 error solutions. As we are working with error-free simulated data, the optimal solution should have no error, but given that the computation of the error implies getting the inverse of the B matrix, this leads to numerical errors that are reflected in the evaluation function. Particularly, for the $n = 10$ instances the errors are below 10^{-16}, suggesting that indeed those solutions are optimal. Paying attention to the results obtained by the four algorithms, we can see that our proposal systematically outperforms the other two algorithms used in the comparison, with CITUP being the one that gets closer to the performance of our ILS approach. Most likely, the differences between our proposal and the other two algorithms are not so much due to the optimization algorithm itself but due to the required simplification of clustering mutations with similar frequencies. This is an important point, as one of the goals of using metaheuristics is to have the flexibility and scalability needed to avoid such simplifications.

With respect to the effect of the number of samples, we can see in the results that it is negligible. However, we have to bear in mind that we generate all the samples completely at random, that is, assuming that the distribution of the clones in the tumor is homogeneous, but in real-life data this may not be true. It is under such circumstances where analyzing the effect of having a more thorough sampling would be of interest.

Finally, we can also see that there is not a clear effect in the obtained results for the increasing number of mutations, except for our approach and between $n = 10$ and the rest. There are different factors that can shed light on this observation. First of all, having more mutations increases (at least in our case) the search space, but also provides more flexibility to find good solutions. In other words, it is not evident what effect the number of mutations has on the shape of the landscape. As for the other two algorithms, we should remember that they cope with an increasing number of mutations using clustering, so the increment in the size of the problem is to some extent masked with that strategy. Also, related to the clustering, we should not forget that the experimentation is preliminary and the way the instances have been generated (in particular, the frequencies associated to the mutations) can be related with the penalization (or lack of) due to the clustering, similarly to the implications of having a homogeneous tumor or not. In any case, in the future a more extensive and realistic evaluation of the algorithm will be needed in order to properly characterize its behaviour under different circumstances.

VI. Conclusions and future work

Solving the CDP is a challenging task that has been tackled from several methodological perspectives. In this work, we have to the best of our knowledge, explored it from a metaheuristic perspective for the first time. Preliminary results on error-free simulated data show that, although the proposed algorithm is very simple, the approach already outperforms other existing heuristic tools when running time is bounded, highlighting the potential of the method. Moreover, the good performance of our algorithms with respect to the other approaches is most likely due to the simplification (clustering mutations) required by them in order to cope with an increasing number of mutations.

Beyond this direct observation, we believe that the use of metaheuristics is promising due to several reasons. First, there is increasing evidence that the ISA model is frequently violated. Recent studies have shown that mutation loss over time and parallel evolution, i.e., the independent acquisition of a same mutation by two clones, are common in certain cancer types [12]. Hence, models allowing such scenarios and, thus, better reflecting the underlying tumor evolutionary mechanisms need to be developed [15], [19]. Whereas specialized or heuristic algorithms fail to easily adapt to such scenarios, the work we have presented is a rather general framework that can work under these new assumptions with minimal changes. This is because the way the evolutionary mode is introduced into our algorithm is through the restrictions on the B matrix and the neighbourhood structure, and not into the algorithm itself. Thus, modifying these two elements of the approach may be enough to accommodate these new models.

Furthermore, the algorithm has also the advantage that it does not explicitly model the error associated with the clone frequencies. Indeed, the same objective function is valid whether we deal with error-free data or not.

In short, we have proposed a novel approach to the CDP which works in the current scenario and is flexible and scalable enough to cope with new knowledge about cancer evolution and bigger problem sizes. However, the work has to be improved in several ways. Regarding the algorithm, other more sophisticated methods should be explored, as the simple experimentation conducted in this work already shows that our approach is not able to reach optimal solutions except for very small problem sizes.

Another point for future work is the evaluation function. In this work we have focused on the error, but there are other possible features of the solution that could be interesting to explore, such as topology-related measures.

Finally, as we have already said, the experimentation conducted in this work is very limited, as this is a preliminary work. Future developments should be compared in more complex scenarios (both simulated and real) against a broader representation of the state-of-the-art and analyzed not only from the error perspective but also from other points of view.

ACKNOWLEDGMENT

M.T.A. is recipient of PREDOC predoctoral fellowship from Department of Education, University and Research of the Basque Government (PRE_2019_2_0211). C.L. is supported by the IKERBASQUE foundation for science and B.C. is funded by the Basque Government (IT-1244-19 and ELKA-RTEK programme) and the Spanish Ministry of Economy, Industry and Competitiveness (TIN2016-78365-R).

REFERENCES

[1] P. C. Nowell, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, pp. 23–28, 1976.

[2] W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, vol. 155, no. 3760, pp. 279–284, 1967.

[3] J. M. Alves, T. Prieto, and D. Posada, "Multiregional tumor trees are not phylogenies," *Trends in cancer*, vol. 3, no. 8, pp. 546–550, 2017.

[4] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris, "Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors," *Genome biology*, vol. 16, no. 1, p. 35, 2015.

[5] Y. Jiang, Y. Qiu, A. J. Minn, and N. R. Zhang, "Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing," *Proceedings of the National Academy of Sciences*, vol. 113, no. 37, pp. E5528–E5537, 2016.

[6] A. Fischer, I. Vázquez-García, C. J. Illingworth, and V. Mustonen, "High-definition reconstruction of clonal composition in cancer," *Cell reports*, vol. 7, no. 5, pp. 1740–1752, 2014.

[7] N. Donmez, S. Malikic, A. W. Wyatt, M. E. Gleave, C. C. Collins, and S. C. Sahinalp, "Clonality inference from single tumor samples using low coverage sequence data," in *International Conference on Research in Computational Molecular Biology*. Springer, 2016, pp. 83–94.

[8] O. E. Ogundijo, K. Zhu, X. Wang, and D. Anastassiou, "A sequential monte carlo algorithm for inference of subclonal structure in cancer," *PloS one*, vol. 14, no. 1, 2019.

[9] S. Malikic, A. W. McPherson, N. Donmez, and C. S. Sahinalp, "Clonality inference in multiple tumor samples using phylogeny," *Bioinformatics*, vol. 31, no. 9, pp. 1349–1356, 2015.

[10] E. Husić, X. Li, A. Hujdurović, M. Mehine, R. Rizzi, V. Mäkinen, M. Milanič, and A. I. Tomescu, "Mipup: minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ilp," *Bioinformatics*, vol. 35, no. 5, pp. 769–777, 2019.

[11] F. Strino, F. Parisi, M. Micsinai, and Y. Kluger, "Trap: a tree approach for fingerprinting subclonal tumor composition," *Nucleic acids research*, vol. 41, no. 17, pp. e165–e165, 2013.

[12] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel, "Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors," *Genome research*, vol. 27, no. 11, pp. 1885–1894, 2017.

[13] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou, "Fast and scalable inference of multi-sample cancer lineages," *Genome biology*, vol. 16, no. 1, p. 91, 2015.

[14] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael, "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data," *Bioinformatics*, vol. 31, no. 12, pp. i62–i70, 2015.

[15] M. El-Kebir, G. Satas, L. Oesper, and B. J. Raphael, "Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures," *Cell systems*, vol. 3, no. 1, pp. 43–53, 2016.

[16] I. Hajirasouliha, A. Mahmoody, and B. J. Raphael, "A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data," *Bioinformatics*, vol. 30, no. 12, pp. i78–i86, 2014.

[17] D. Gusfield, "Efficient algorithms for inferring evolutionary trees," *Networks*, vol. 21, no. 1, pp. 19–28, 1991.

[18] M. Bordewich and C. Semple, "On the computational complexity of the rooted subtree prune and regraft distance," *Annals of combinatorics*, vol. 8, no. 4, pp. 409–423, 2005.

[19] H. Zafar, A. Tzen, N. Navin, K. Chen, and L. Nakhleh, "Sifit: inferring tumor trees from single-cell sequencing data under finite-sites models," *Genome biology*, vol. 18, no. 1, p. 178, 2017.