HHS Public Access

Author manuscript

IEEE/ACM Trans Comput Biol Bioinform. Author manuscript; available in PMC 2017 July 01.

Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2016 ; 13(4): 643–655. doi:10.1109/TCBB.2015.2476808.

Classifying the Progression of Ductal Carcinoma from Single-Cell Sampled Data via Integer Linear Programming: A Case Study

Daniele Catanzaro,

Louvain School of Management and the Center for Operations Research and Econometrics (CORE) of the Université Catholique de Louvain (UCL), Chausseé de Binche, 151, 7000 Mons, Belgium.

Stanley E. Shackney,

Departments of Human Oncology and Human Genetics of the Drexel University School of Medicine, Pittsburgh, PA, USA.

Alejandro A. Schäffer, and

Computational Biology Branch of NCBI, NIH, Bethesda, MD, USA.

Russell Schwartz

Department of Biological Sciences and the Computational Biology Department of the Carnegie Mellon University, Pittsburgh, PA, USA.

Abstract

Ductal Carcinoma In Situ (DCIS) is a precursor lesion of Invasive Ductal Carcinoma (IDC) of the breast. Investigating its temporal progression could provide fundamental new insights for the development of better diagnostic tools to predict which cases of DCIS will progress to IDC. We investigate the problem of reconstructing a plausible progression from single-cell sampled data of an individual with Synchronous DCIS and IDC. Specifically, by using a number of assumptions derived from the observation of cellular atypia occurring in IDC, we design a possible predictive model using integer linear programming (ILP). Computational experiments carried out on a preexisting data set of 13 patients with simultaneous DCIS and IDC show that the corresponding predicted progression models are classifiable into categories having specific evolutionary characteristics. The approach provides new insights into mechanisms of clonal progression in breast cancers and helps illustrate the power of the ILP approach for similar problems in reconstructing tumor evolution scenarios under complex sets of constraints.

Correspondence should be addressed to: daniele.catanzaro@uclouvain.be.

Digital Object Identifier no. 10.1109/TCBB.2015.2476808

Keywords

Tumor profiling; single-cell sequencing; ductal carcinoma of the breast; phylogeny estimation; parsimony criterion; computational biology; distance methods; network design; combinatorial optimization; mixed integer linear programming

1 Introduction

Ductal Carcinoma In Situ (DCIS) is considered a precursor lesion for invasive breast cancer and is found synchronously in approximatively 45% of patients affected by Invasive Ductal Carcinoma (IDC) [1]. Specifically, DCIS is the last step in a continuum of non-invasive stages of increased cellular atypia, which are believed to develop from flat epithelial atypia and atypical ductal hyperplasia [2]. Incidences of DCIS and IDC were estimated at 35 and 155 per 100,000 women in the United States, respectively some years ago [3], [4]. The incidences of DCIS and early-stage IDC are expected to increase due to improvements in accuracy of mammography and its increased usage [5]. Investigating the temporal progression of solid tumors could provide fundamental new insights for the development of more effective diagnoses and treatments. Hence, increasing research efforts have been devoted to this topic in recent years [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. The topic has been reviewed in [17], [18].

Several studies have been performed on the dynamics of genomic alterations during the evolution of breast cancer using comparative genomic hybridization (CGH) data [19], mutation data [20], [21], [22], or fluorescence in situ hybridization (FISH) data [7], [11], [23], [24]. In this case study, we focus on reanalyzing the data collected on DCIS and IDC in [23]. Specifically, the authors carried out a single-cell FISH analysis [7], [20], [21], [25] on 13 patients with synchronous DCIS and IDC of the breast. Heselmeyer-Haddad and colleagues observed both an enormous intercellular heterogeneity in DCIS and IDC (although lower in DCIS with respect to IDC) and signal patterns consistent with a non-random distribution of genomic imbalances. The presence of recurrent patterns of genomic imbalances in the evolution from DCIS to IDC led the authors to suspect that similar sequences of genetic events might underlie progression across the patient cohort, enabling a classification. However, the partial classification proposed in [23] is based on static analysis of the single-cell data, not a model of progressing genetic changes.

Recent studies in various tumor types have shown extensive intra-tumor heterogeneity [26] either when the molecular data are point mutations [27], [28] or when the data are gene copy numbers measured by FISH [29], [30], as we use in this study. Therefore, modeling tumor progression based on single-cell data, as we do here, should yield more accurate models than analyses based on tumor-wide data on copy number or mutations [31]. Subsequent phylogenetic analyses [15], [16] of the data of [23] provided a proof of principle that one can reconstruct models of progression for DCIS and IDC data capable of identifying distinguishing features of amplification and loss of specific driver genes. Such work is, however, limited by the computational difficulty of accurately fitting phylogenetic trees to tumor data, particularly with regard to extending algorithmic theory of phylogenetics to cover realistic models of genomic copy number evolution. Further, none of the existing

methods for single-tumor phylogenetics, to our knowledge, is able to explicitly consider models of progression, such as DCIS/IDC data or explicit time-series samples, that would have a clinically defined sequence of progression among samples.

Using the data of [23], in this article, we address the problem of modeling and classifying progression of ductal carcinoma by single-cell phylogenetic analysis of tumors from a population of affected individuals. By making use of assumptions derived from the observation of cellular atypia occurring in ductal carcinoma, we first design a possible parsimony-based predictive model able to reconstruct a plausible progression of the carcinoma from single-cell samples of a patient. The modeling uses integer linear programming (ILP) to fit the single-cell data to a set of constraints that are intended to capture a model of our prior biological knowledge of plausible pathways of progression. The values of some variables in an optimal solution of the IP are interpreted as a structured model of progression. The use of mathematical programming to model tumor progression from molecular data was suggested by Farahani and Lagergren [32]. Mathematical programming and optimization have been used in a previous study on the relationship between DCIS and IDC [33], but that study was based on clinical characteristics of the patients without molecular data. Interestingly, a follow-up study on the same data suggested that mixture models allowing for two different types of clinical progression fit the data much better than any single model [34]. Mixture modeling has also been used effectively to infer tumor progression pathways from CGH data [35], [36]. In that same spirit, our computational experiments using FISH data show that the corresponding predicted progressions are non-random and classifiable into categories having specific evolutionary characteristics.

2 Profiling the progression of ductal carcinoma from single-cell sampled data

In this section, we briefly describe the data collected in [23] and introduce a number of biological and evolutionary assumptions that will prove useful to approximate progression of the tumor pathology considered here.

2.1 Sample data

The literature on tumorigenesis shows that progression of a tumor proceeds over time by increasing atypia from a normal cell [10], [11]. In ductal carcinoma, such atypia may affect, among other things, the number of copies of chromosomal segments and the number of genes in those segments (see [23]). Hence, the variation of the copy number of a (set of) gene(s) can be used as a measure of progression of a cancer cell with respect to the healthy cells.

The data collected in [23] have been extracted from a cohort of 13 patients affected with synchronous DCIS and IDC and include copy numbers of five oncogenes: *COX2* (cyclooxygenase 2, located on 1q31.1), *MYC* (c-MYC, located on 8q24.21), *HER2* (human epidermal growth factor receptor 2, located on17q12), *CCND1* (cyclin D1, located on 11q13.3) and *ZNF217* (zinc finger protein 217, located on 20q13.2) and three tumor

suppressor genes namely, *DBC2* (deleted in breast cancer, located on 8p21.3), *CDH1* (cadherin1, also known as epithelial cadherin, located on 16q22.1) and *TP53* (tumor protein p53, located on 17p13.1). Numerous studies have shown that the considered oncogenes are preferentially gained in tumor pathologies of the breast and that the tumor suppressor genes are preferentially lost in these pathologies [23]. In particular, losses of *TP53* mechanistically promote instability in the genome (see [37]), and hence variations of the copy number of *TP53* deserve special attention. Given a population of individuals affected by DCIS and IDC, we assume that for each individual, data on DCIS and IDC single-cell samples (or *taxa*) are available. Here, we refer to the paired samples for one patient as a *dataset* and the datasets are numbered DAT1 through DAT13. For each cell of the either sample in a dataset, the ordered list of copy numbers of the eight genes is called a *taxon*. We encode each taxon as a sequence of eight numbers each of which represents the number of copies of one gene. For example, the sequence ⟨2.2.4.2.1 – 1.2.3⟩ describes a taxon having (in the following order): 2 copies of *COX2*, 2 copies of *MYC*, 4 copies of *HER2*, 2 copies of *CCND1*, 1 copy of *ZNF217*, 1 copy of *DBC2*, 2 copies of *CDH1*, and 3 copies of *TP53*, respectively. By convention, we use the dot to separate the copies of different genes in a given taxon and the dash to separate oncogenes from tumor suppressor genes. Different cells may have the same copy number taxon; the multiplicity of each taxon is part of the data, but not used in our modeling. The number of distinct taxa, ignoring multiplicity, in a sample ranges from 35 to 126 for this dataset (see Table 1).

2.2 Biological assumptions

Human populations show an extensive polymorphism in the number of copies of some chromosomal segments and genes, called *Copy Number Variants* (CNVs), even among twins [38], but such benign CNVs have not been reported for the eight genes studied here. All eight genes are on autosomes, not the X chromosome. Therefore, it usually holds that healthy cells carry 2 copies of each gene, i.e., healthy cells can be represented by the taxon ⟨2.2.2.2.2 – 2.2.2⟩ [23]. Hence, we consider the all-2 taxon as the origin of progression of ductal carcinoma in each patient.

During progression of the carcinoma, the copy number of a gene can potentially increase indefinitely or decrease to zero. Once a gene is lost entirely in a given taxon t, it is plausible to believe that such a gene cannot be regained in a subsequent descendant of t. Hence, we exclude the possibility that a generic taxon having zero copies of a specific gene, e.g., ⟨2.2.4.2.1 – 0.2.3⟩, could be considered as the ancestor of any other taxon in the same dataset having a strictly positive number of copies of that gene, e.g., ⟨2.2.4.2.1 – k.2.3⟩, with $k > 0$. We refer to this assumption as the *ex nihilo nihil assumption*.

Similarly to [11], we also assume that: (i) invasive tumors still contain cancer cells from earlier progression steps even if they have not been sampled; (ii) the rate of cell proliferation and death is not significantly different among taxa; and (iii) the cells sampled in each dataset are a reasonable representation of the whole tumor. Moreover, we also assume that: (iv) the copy number of a gene can freely increase or decrease (provided that it remains strictly positive) along a given pathway; and (v) the temporal progression should be respected, i.e.,

that DCIS taxa should temporally precede IDC taxa (although this assumption is relaxed later).

We further consider some additional assumptions to specifically reflect a model of the effects of *TP53* loss. In particular, we assume that (vi) an increment of the number of copies of *TP53* in a given taxon may potentially cause a *Doubling-Loss Event* (DLE) in its immediate descendant, i.e., a doubling of the number of copies of (all or part of) the genes in the taxon followed by a possible loss of copies of one or more genes (see [23]). Moreover, we assume that in absence of an alteration of the copy number of *TP53* it is unlikely: (vii) to have more than three genes whose corresponding copy numbers double in a generation or (viii) to gain more than 3 copies of a gene in a generation. Finally, we assume that (ix) in case of a decrement of the copy number of *TP53* it is unlikely that the immediate descendant gains more than three copies of a gene. It is worth noting that these assumptions may not be strictly conserved on real data because *TP53* dysfunctions, or functionally equivalent abnormalities, can occur by means of a variety of mechanisms beyond *TP53* copy number variation and that might not be visible in FISH data. Nonetheless, we include these additional assumptions in part to illustrate the kind of complex constraints for which the ILP approach is especially well suited.

2.3 The estimation criterion

The assumptions described above provide a list of the characteristics that we will assume should or could be included in any plausible prediction of progression of ductal carcinoma affecting a given patient. However, it is worth noting that these assumptions provide neither a criterion to predict progression itself nor a criterion to select a prediction from among plausible alternatives. Hence, in order to predict the progression of ductal carcinoma in a patient, two problems to be solved include identifying both a construction criterion and a selection criterion. As ductal carcinoma is a somatic evolutionary process, a possible approach to identify both criteria consists of using classical evolutionary theory [39]. Specifically, provided a measure of the dissimilarity among taxa, the theory assumes that one taxon evolves from another by means of “small progressive changes,” mainly because the selective forces acting on that taxon may not be constant throughout its evolution [40], [41]. Over time, a collection of small changes will not generally provide the smallest accumulated change. However, if the changes are sufficiently small and the time scales over which taxa would be expected to have evolved are sufficiently short, the process of approximating small changes with smallest change can properly fit the corresponding evolutionary process [39]. This criterion, known in the literature as the *parsimony criterion* [39], suggests both a method to predict progression itself (e.g., by joining similar taxa at each step of the progression from the healthy taxon) and a criterion to select a prediction from among plausible alternatives (e.g., by choosing the one that globally minimize the dissimilarity among all taxa in the dataset, [42]). The large number of taxa per patient in the available FISH datasets and relatively short time scales over which they would be expected to have evolved (months to years) both support the use of parsimony over more sophisticated likelihood models for this kind of data. The small number of markers typically available and the frequent presence of ancestral cell states within the observed data also make this problem poorly suited for fast distance-based phylogeny methods, such as neighbor-joining or

UPGMA [43], and similar hierarchical clustering algorithms. However, the use of the parsimony criterion involves solving a particular network design problem whose optimal solution has to satisfy the assumptions described in Subsection 2.2. In the next section we shall formalize this problem and develop a possible mathematical programming approach to solve it exactly.

3 Modeling the parsimony criterion

The use of the parsimony criterion involves, as a first task, the identification of a measure of dissimilarity between taxa. As a generic taxon in a given dataset can be seen as a point in an eight-dimensional space, a possible measure of the dissimilarity between a generic pair of taxa in the dataset can be obtained by identifying a specific norm function able to reflect appropriately some or all of the assumptions listed Subsection 2.2. Before investigating this issue, we introduce some notation that will prove useful throughout the article.

3.1 Notation

We define D to be the set of sample data from a specific patient, $\mathscr{G}$ to be the set of the eight genes sorted according to the order described in Section 2.1, and g to be a generic gene in $\mathscr{G}$. Given a taxon $t_i \in D$, we define t_i^g to be the copy number of the g-th gene in t_i. For example, if $t_i = \langle 4.3.4.5.1 - 4.4.3 \rangle$ then $t_i^{HER2} = 4$. We define $\mathscr{O}$ and $\mathscr{S}$ to be the sets of the oncogenes and the tumor suppressor genes in $\mathscr{G}$, respectively. Finally, we define $\mathscr{DCIS}$ and $\mathscr{IDC}$ to be the subsets of DCIS and IDC taxa in D, respectively, and $\mathscr{ISI} = \mathscr{DCIS} \cap \mathscr{IDC}$. We assume only distinct taxa in the DCIS the IDC individually, but the same taxon may occur in both. We assume that D always contains the healthy taxon, defined by having two (diploid) copy number for all genes. If this healthy taxon is missing from the input data, we add it in D. Moreover, we sort taxa in D in such a way that (i) the healthy taxon is the first taxon in D; (ii) taxa in $\mathscr{DCIS}$ and $\mathscr{IDC}$ follow a tree-structured partial order; (iii) taxa in $\mathscr{DCIS}$ are located after the healthy taxon; and (iv) a taxon that occurs only in $\mathscr{IDC}$ cannot precede a taxon that occurs only in $\mathscr{DCIS}$. Given a pair of distinct taxa $t_i, t_j \in D$, $i < j$, we define $\omega_{ij}^g = \left| t_i^g - t_j^g \right|$, for all $g \in \mathscr{G}$. We define $\eta_{\mathscr{O}}$ as the number of oncogenes whose counts have doubled during the transition from taxon t_i to taxon t_j. Moreover, for a fixed dataset D we define the following sets:

$$R_1 = \left\{ (t_i, t_j) : \exists\, g \in \mathscr{G} : t_i^g = 0 \wedge t_j^g > 0 \right\} \quad (1)$$

$$R_2 = \{ (t_i, t_j) : t_i \in \mathscr{IDC} : t_j \in \mathscr{DCIS} \} \quad (2)$$

$$R_3 = \left\{ (t_i, t_j) : t_i^{TP53} = t_j^{TP53}, \eta_{\mathscr{O}} \geq 3 \right\} \quad (3)$$

$$R_4 = \left\{ (t_i, t_j) : t_i^{TP53} = t_j^{TP53} \wedge \exists\, g \in \mathcal{G} : \omega_{ij}^g \geq 3 \right\} \quad (4)$$

$$R_5 = \left\{ (t_i, t_j) : t_i^{TP53} > t_j^{TP53} \wedge \exists\, g \in \mathcal{O} : t_j^g \geq 3t_i^g \right\}. \quad (5)$$

The set R_1 denotes the set of pairs of taxa in D violating the *ex nihilo nihil* assumption. The set R_2 denotes the set of pairs of taxa in D violating the temporal progression assumption (v) in a direct transition from ancestor to descendent. The sets R_3, R_4 and R_5 denote the sets of pairs of taxa in D violating the assumptions (vii), (viii) and (ix), respectively.

3.2 Measuring the dissimilarity among taxa

As a generic taxon in the FISH dataset D can be encoded as an eight-dimensional vector, a natural choice to measure the dissimilarity between taxa $t_i, t_j \in D$ is the L_1 distance [15], defined as:

$$d_1(t_i, t_j) = \|t_i - t_j\|_1 = \sum_{g \in \mathcal{G}} \omega_{ij}^g. \quad (6)$$

However, this measure is characterized by a major drawback: it does not take into account the evolutionary process related to the transition from t_i to t_j. As a result, the same distance can be assigned to different evolutionary processes, which in turn can be interpreted as equiprobable events, even if they are not in reality. For example, consider the following list of three taxa: $t_1 = \langle 2.2.2.2.2 - 2.2.2 \rangle$, $t_2 = \langle 1.1.1.1.2 - 2.2.2 \rangle$ and $t_3 = \langle 2.2.2.2.6 - 2.2.2 \rangle$. If we use equation (6) to compute the dissimilarity between the healthy cell and the remaining two taxa we get $d(t_1, t_2) = d(t_1, t_3) = 4$. This fact means that, if the transition from t_1 to t_2 and from t_1 to t_3 happened in just one generation, the process of losing one copy in genes *COX2*, *MYC*, *HER2* and *CCND1* during evolution from t_1 to t_2 would be an event as likely as gaining four copies of *ZNF217* during evolution from t_1 to t_3. As our classification attempt is based on the parsimony criterion, we assume that the transition from t_1 to t_3 is more likely than the transition from t_1 to t_2, because it involves a change in a smaller number of genes. One approach to account for this fact consists of modifying equation (6) as follows:

$$d_2(t_i, t_j) = d_1(t_i, t_j) + c_1(|\mathcal{G}| - \mu_{ij}) \quad (7)$$

where μ_{ij} is the number of equal entries (copy numbers) in taxa $t_i, t_j \in D$ and c_1 is a positive constant used to weight added μ_{ij}. For example, if we set $c_1 = 1/8$ then we get $\mu_{12} = 5$, $\mu_{13} = 7$, $d(t_1, t_2) = 4.375$ and $d(t_1, t_3) = 4.125$. Hence, the transition from t_1 to t_3 becomes more likely than the transition from t_1 to t_2. It is worth noting that the addend $c_1(|\mathcal{G}| - \mu_{ij})$ represents the Hamming distance between vectors t_i and t_j, weighted by the positive constant c_1. We denote this addend as $H(t_i, t_j)$.

The following proposition holds:

Proposition 1—*The function $d_2(\cdot, \cdot)$ induces a metric space in D.*

Proof: If $d_2(\cdot, \cdot)$ is a metric then it has to satisfy the following properties:

1) $d_2(t_i, t_j) \geq 0$ (non-negativity);

2) $d_2(t_i, t_j) = 0$ iff $t_i = t_j$ (coincidence axiom);

3) $d_2(t_i, t_j) = d_2(t_j, t_i)$ (symmetry);

4) $d_2(t_i, t_z) \leq d_2(t_i, t_j) + d_2(t_j, t_z)$ (triangle inequality).

It is easy to see that $d_2(\cdot, \cdot)$ satisfies the first three properties. To show that $d_2(\cdot, \cdot)$ also satisfies the triangle inequality, observe that the following inequality holds:

$$\begin{aligned} d_2(t_i, t_z) &= \|t_i - t_z\|_1 + H(t_i, t_z) \\ &\leq \|t_i - t_j\|_1 + H(t_i, t_j) \\ &+ \|t_j - t_z\|_1 + H(t_j, t_z). \end{aligned} \quad (8)$$

Inequality (8) can be seen as the convex combination (with multipliers equal to 1) of the following two inequalities:

$$\|t_i - t_z\|_1 \leq \|t_i - t_j\|_1 + \|t_j - t_z\|_1 \quad (9)$$

$$H(t_i, t_z) \leq H(t_i, t_j) + H(t_j, t_z). \quad (10)$$

Both (9) and (10) satisfy the triangle inequality [44], hence the statement holds.

Although $d_2(\cdot, \cdot)$ takes into account the number of genes whose copy numbers have been subjected to a change during the transition from taxon t_i to taxon t_j, it does not take into account the value of such a change. For example, consider the following list of three taxa: $t_1 = \langle 2.2.2.2.2 - 2.2.2 \rangle$, $t_2 = \langle 2.2.2.4.4 - 2.2.2 \rangle$ and $t_3 = \langle 2.2.2.2.6 - 2.2.2 \rangle$. If we set $c_1 = 1/8$ and use equation (7) to compute $d(t_1, t_2)$ and $d(t_1, t_3)$ we get $d(t_1, t_2) = 4.25$ and $d(t_1, t_3) = 4.125$. Hence, the transition from t_1 to t_3 is more likely than the transition from t_1 to t_2. However, as mentioned in the assumptions (vii)-(ix), we consider that the process of gaining three or more copies of any gene in just one generation is an unlikely event. One approach to impose these assumptions consists of penalizing the transitions involving copy number changes greater than one unit. This task can be performed by modifying equation (7) as follows:

$$d_3(t_i,t_j) = d_2(t_i,t_j) + \sum_{\substack{g\in\mathcal{G}:\\ \omega_{ij}^g=2}} c_2\omega_{ij}^g + \sum_{\substack{g\in\mathcal{G}:\\ \omega_{ij}^g=3}} c_3\omega_{ij}^g + \sum_{\substack{g\in\mathcal{G}:\\ \omega_{ij}^g\geq 4}} c_4\omega_{ij}^g \quad (11)$$

where c_k satisfying $c_k \leq c_{k+1}$, $k = 2, \ldots, 4$, are positive constants used to weight the new addends. For example, if $c_2 = 1/16$, $c_3 = 1/8$ and $c_4 = 1/5$ then $d(t_1, t_2) = 4.5$ and $d(t_1, t_3) = 5.125$. Thus, the transition from t_1 to t_2 becomes more likely than the transition from t_1 to t_3.

Proposition 2—*The function $d_3(\cdot, \cdot)$ induces a semimetric space in D.*

Proof: It is easy to see that the non-negativity, the symmetry properties hold for $d_3(\cdot, \cdot)$. To see that also the coincidence axiom holds, denote

$$Q_{ij} = \sum_{\substack{g\in\mathcal{G}:\\ \omega_{ij}^g=2}} c_2\omega_{ij}^g + \sum_{\substack{g\in\mathcal{G}:\\ \omega_{ij}^g=3}} c_3\omega_{ij}^g + \sum_{\substack{g\in\mathcal{G}:\\ \omega_{ij}^g\geq 4}} c_4\omega_{ij}^g$$

and observe that, for some $t_i, t_j \in D$, $d_3(t_i, t_j)$ can be written as

$$d_3(t_i,t_j) = d_2(t_i,t_j) + Q_{ij}. \quad (12)$$

If $t_i = t_j$, then $d_2(t_i, t_j) = 0$ and $Q_{ij} = 0$. If $t_i \neq t_j$, then $d_2(t_i, t_j) > 0$ and either $Q_{ij} = 0$ or $Q_{ij} \geq 0$. As at least one of the two addends is different from 0, the coincidence axiom holds.

Now observe that $d_3(\cdot, \cdot)$ does not satisfy the triangle inequality. This fact can be seen by showing a counterexample. Specifically, first note that if the triangle inequality held for $d_3(\cdot, \cdot)$, then for some $t_i, t_j, t_z \in D$, $d_3(t_i, t_j)$ we should have

$$\begin{aligned} d_3(t_i,t_z) &= \|t_i - t_z\|_1 + H(t_i,t_z) + Q(i,z) \\ &\leq \|t_i - t_j\|_1 + H(t_i,t_j) + Q(i,j) \\ &+ \|t_j - t_z\|_1 + H(t_j,t_z) + Q(j,z). \end{aligned} \quad (13)$$

Inequality (13) can be seen as the convex combination (with multipliers equal to 1) of the following three inequalities:

$$\|t_i - t_z\|_1 \leq \|t_i - t_j\|_1 + \|t_j - t_z\|_1 \quad (14)$$

$$H(t_i, t_z) \leq H(t_i, t_j) + H(t_j, t_z) \quad (15)$$

$$Q(i, z) \leq Q(i, j) + Q(j, z). \quad (16)$$

The L_1 distance and the Hamming distance satisfy the triangle inequality. However, inequality (16) does not. In fact, if we consider the following three taxa: $t_1 = \langle 1.1.1.1.1 - 1.1.1 \rangle$, $t_2 = \langle 2.2.2.2.2 - 2.2.2 \rangle$ and $t_3 = \langle 3.3.3.3.3 - 3.3.3 \rangle$ we have that $Q(i, z) = 16c_2 > Q(i, j) + Q(j, z) = 0$. Thus, the statement follows.

Semimetrics are particularly interesting to our purposes, as they suggest that transitions from, e.g., $t_1 = \langle 1.1.1.1.1 - 1.1.1 \rangle$ to $t_2 = \langle 2.2.2.2.2 - 2.2.2 \rangle$ and from t_2 to $t_3 = \langle 3.3.3.3.3 - 3.3.3 \rangle$, are more likely than the single-step transition from t_1 to t_3.

It is worth noting that there may exist particular types of transitions that (i) happen in one generation, (ii) involve a large part or all of the genes in $\mathcal{G}$, and (iii) are characterized by big copy number changes. For example, when a DLE occurs then it is possible to evolve from $t_1 = \langle 2.2.2.2.2 - 2.2.2 \rangle$ to $t_2 = \langle 4.4.4.4.4 - 4.4.3 \rangle$ in just one generation. We assume that DLE is relatively rare compared to more localized variations and may be not easy to detect. For example, it is not easy to determine whether the pair of taxa $t_1 = \langle 2.2.2.1.2 - 1.1.1 \rangle$ and $t_2 = \langle 2.4.4.2.4 - 2.2.2 \rangle$ represents a DLE or arise through two distinct evolutionary processes from the healthy all-2 cell. However, if a DLE was the transition that occurred, equation (11) would provide $d(t_1, t_2) = 11.25$ under the assumption that $c_1 = 1/8$, $c_2 = 1/16$, $c_3 = 1/8$ and $c_4 = 1/4$. This value is far larger than the dissimilarity between, e.g., t_1 and $t_3 = \langle 3.3.3.2.3 - 2.1.2 \rangle$ under the same cost value (equal to $d(t_1, t_3) = 7.125$). Hence, the transition from t_1 to t_2 would not be considered as likely. To enable identification of possible DLEs in the considered datasets, we modify the dissimilarity measure (11) by adding a new term in (11). Specifically, we denote η as the number of genes whose copy numbers have doubled during the transition from taxon t_i to taxon t_j and we set $\nu = |\mathcal{G}| - \eta$. Then, we consider the following measure of dissimilarity between taxa $t_i, t_j \in D$, $i < j$:

$$\begin{aligned} d_4(t_i, t_j) &= d_3(t_i, t_j) + c_5 \min\{\eta, \nu\} \\ &= d_3(t_i, t_j) + c_5 \min\{\eta, |\mathcal{G}| - \eta\}, \end{aligned} \quad (17)$$

where c_5 is a sufficiently large positive constant such that $c_5 >> c_k$, $k = 1, \ldots, 4$. The addend $c_5 \min\{\eta, |\mathcal{G}| - \eta\}$ constitutes a weighted heuristic approximation of Lee distance [44]. The following proposition holds:

Proposition 3—*The function $d_4(\cdot, \cdot)$ induces a premetric space in D.*

Proof: It is easy to see that the non-negativity and the coincidence axioms hold for $d_4(\cdot, \cdot)$. Since the triangle inequality does not hold for $d_3(\cdot, \cdot)$, then $d_4(\cdot, \cdot)$ does not induce a metric space either. To see that also the symmetry property does not hold for $d_4(\cdot, \cdot)$ it is sufficient

to consider taxa. $t_1 = \langle 2.2.2.2.2 - 1.1.1 \rangle$ and $t_2 = \langle 1.1.4.4.4 - 2.1.1 \rangle$ and to observe that $d_4(t_1, t_2) = d_3(t_1, t_2) + 4c_5$ and $d_4(t_2, t_1) = d_3(t_2, t_1) + 2c_5 = d_3(t_1, t_2) + 2c_5$.

Premetrics are useful for our purposes as, in our classification attempt, we consider DLEs as *directional evolutionary processes*, i.e., processes that consider the transition from taxon t_i to t_j and from taxon t_j to t_i as two events having different probabilities. It is also worth noting that $d_4(\cdot, \cdot)$ models the assumption that a duplication event followed by some loss or (few) gain events is more probable than a large number of sequential gain or loss events. For example, $d_4(\cdot, \cdot)$ indicates that the transition e.g., from $t_1 = \langle 2.2.2.2.2 - 2.2.2 \rangle$ to $t_2 = \langle 4.4.4.4.4 - 4.4.3 \rangle$ is more likely to have occurred in two events (e.g., from $t_1 = \langle 2.2.2.2.2 - 2.2.2 \rangle$ to $t_k = \langle 4.4.4.4.4 - 4.4.4 \rangle$ and then from t_k to $t_2 = \langle 4.4.4.4.4 - 4.4.3 \rangle$) rather than in a sequence of transitions in which at each step a copy number of a gene increase (e.g., from $t_1 = \langle 2.2.2.2.2 - 2.2.2 \rangle$ to $t_{k_1} = \langle 3.2.2.2.2 - 2.2.2 \rangle$, from t_{k1} to $t_{k2} = \langle 3.3.2.2.2 - 2.2.2 \rangle$, and so on, until the transition from $t_{kq} = \langle 4.4.4.4.4 - 4.3.3 \rangle$ to $t_{k2} = \langle 4.4.4.4.4 - 4.4.3 \rangle$). In the remainder of the article, we shall use $d_4(\cdot, \cdot)$ to measure the dissimilarity between a pair of taxa in D and for simplicity of notation we will write $d_4(t_i, t_j)$ as d_{ij}.

3.3 Using integer programming to predict the progression of ductal carcinoma under the parsimony criterion

In this section, we formalize the parsimony criterion in terms of an optimization problem on a graph. To this end, given a set D of single-cell sample data extracted from a patient, consider a complete undirected weighted graph $G = (V, E)$, called *tumor graph*, having a vertex i for each taxon $t_i \in D$ and a weight d_{ij} for each edge $(i, j) \in E$. We assume that vertex 1 represents the healthy taxon $\langle 2.2.2.2.2 - 2.2.2 \rangle$, i.e., the origin of progression, and we define $V_1 = V \setminus \{1\}$. Then, predicting the progression of ductal carcinoma from D via the parsimony criterion is equivalent to solving the following problem:

Problem—The Parsimonious Tumor Progression Problem (PTPP)

Given a set D of single-cell sample data extracted from a patient and the corresponding tumor graph G = (V, E), find an arborescence rooted in vertex 1, covering all of the remaining vertices in V_1, satisfying the assumptions discussed in Subsection 2.2, and such that the sum of the distances among all pairs of adjacent taxa in the arborescence is minimized.

It is worth noting that the PTPP cannot be trivially solved by means of classical hierarchical clustering algorithms such as UPGMA [43]. In fact, in classical molecular phylogenetics the observed taxa can only be terminal vertices of a phylogeny; the internal vertices represent speciation events occurred along evolution of taxa are usually assumed to be non-contemporary to them. In contrast, in tumorigenesis, ancestor and descendant clones may co-exist throughout tumor progression [23]. This fact alone suggests that a phylogeny obtained via traditional clustering algorithms may not constitute a suitable representation of a tumorigenesis. Similarly, the PTPP cannot be trivially solved by using any greedy algorithm for the minimum spanning tree [45], at least because the solutions provided by these algorithms may not satisfy the assumptions (iv)-(v).

A possible approach to solve PTPP consists of using Integer Linear Programming (ILP), which provides both a powerful tool to model the assumptions in Subsection 2.2 and a certificate of optimality for the solution to the PTPP so computed. In particular, the presence of the assumption (v) may require the use of ILP formulations based on path variables and solvable via column generation approaches and branch-and-price methods similar to those described, for example, in [46]. However, we have observed that the strict temporal progression assumption (v) does not quite match the study design of [23]. In fact, the DCIS and IDC single-cell samples extracted from each patient were physically sampled at the same time. Even if the assumption that the DCIS state strictly precedes the IDC state is completely correct at the single-cell level, the transformation from DCIS to IDC must have happened at an earlier point in time when the DCIS was in a precursor state P of the sampled state. Since gene copy number changes are likely to have occurred between P and the sampled DCIS, the cells containing these late changes will not be temporal predecessors of cells in the sampled IDC. For example, in the dataset DAT10 (see Figure 10 in supplementary data downloadable at perso.uclouvain.be/daniele.catanzaro/SupportingMaterial/IDC.pdf) the IDC taxon $\langle 2.2.2.2.2 - 3.2.2 \rangle$ would be obliged to appear in progression later than DCIS taxa $\langle 2.2.2.3.2 - 2.2.2 \rangle$ and $\langle 2.2.2.2.2 - 2.3.2 \rangle$ even if it seems to be contemporaneous with them. Moreover, this fact leads us to suspect the possibility that the causes underling the progression from DCIS to IDC may not be exclusively restricted to the 8 genes considered in [23]. Hence, we decided to relax the strict temporal progression assumption (v) along a path by transforming it into an absence of alternating invasive/non-invasive states, i.e., an absence in the arborescence of three contiguous vertices $i \neq j \neq k \in V$ such that i precedes j, j precedes k, and such that the corresponding taxa satisfy the following property: $t_i \in \mathscr{I}\mathscr{D}\mathscr{C}$, $t_j \in \mathscr{I}\mathscr{S}\mathscr{I}$ and $t_k \in \mathscr{D}\mathscr{C}\mathscr{I}\mathscr{S}$. Moreover, we decided to relax the absence in the arborescence of pairs of taxa belonging to $R_3 \cup R_4 \cup R_5$ by penalizing their existence by means of a particular cost in the objective function. Specifically, we add in (17) the addend $c_6\lambda_{ij}$, where $c_6 \geq c_5$, and

$$\lambda_{ij} = \begin{cases} \sum\limits_{g \in \mathscr{G}} \omega_{ij}^{g} & (t_i, t_j) \in \bigcup_{s=3}^{5} R_{s'} \\ 0 & \text{otherwise.} \end{cases}$$

Then, a possible ILP formulation for the PTPP is the following:

Formulation

$$\min \sum_{\substack{i,j \in V: \\ i \neq j}} (d_{ij} + c_6\lambda_{ij})\, y_{ij} \tag{18a}$$

$$y_{ij} = 0 \qquad \forall\, i,j \in V : (i,j) \in R_1 \cup R_2 \tag{18b}$$

$$y_{ij}+y_{jk}<=1 \qquad \forall\, i<j<k \in V_1: \; t_i \in \mathscr{I}\mathscr{D}\mathscr{C},\; t_j \in \mathscr{I}\mathscr{S}\mathscr{I},\; t_k \in \mathscr{D}\mathscr{C}\mathscr{I}\mathscr{S} \tag{18c}$$

$$\sum_{\substack{i,j \in S: \\ i \neq j}} y_{ij} \leq |S| - 1 \qquad \forall\, S \subset V, S \neq \varnothing \tag{18d}$$

$$\sum_{\substack{i,j \in V: \\ i \neq j}} y_{ij}=|V| - 1 \tag{18e}$$

$$y_{j1}=0 \qquad \forall\, j \in V_1 \tag{18f}$$

$$y_{ij} \in \{0,1\} \qquad \forall\, i,j \in V: i \neq j. \tag{18g}$$

Constraints (18b) impose the absence of pairs of adjacent vertices in the arborescence belonging to $R_1 \cup R_2$, i.e., pairs of taxa violating either the *ex nihilo nihil* assumption or the strict temporal progression assumption (v) in a direct transition from ancestor to descendent. Constraints (18c) impose the absence of oscillating invasive/non-invasive states. Constraints (18d) impose the absence of cycles in the arborescence. Constraint (18e) imposes that the overall number of arcs in the arborescence is $|V| - 1$. Finally, constraints (18f) impose the absence of incoming arcs for vertex 1.

4 Numerical experiments

To test the predictions provided by Formulation 3.2 we reanalyzed the [23] datasets obtained from a population of 13 individuals affected by invasive ductal carcinoma of the breast. For completeness, we report in Table 1 a simple description of each dataset in terms of the number of DCIS and IDC single-cell taxa. We refer the interested reader to [23] for a more extensive and systematic description the datasets.

We implemented Formulation 3.2 in ANSI C++ by using Xpress Optimizer libraries v18.10.00. The experiments run on a Pentium 4, 3.2 GHz, equipped with 2 GByte RAM and operating system Gentoo release 7 (Linux kernel 2.6.17). When solving the instances of Formulation 3.2, we activated Xpress automatic cuts, Xpress pre-solving strategy, and used Xpress primal heuristic to generate the first upper bound for the problem. We did not limit

the computation time of the above Formulation because time is not a hard constraint for this application.

As progression of a tumor in a patient usually is an unobservable process, there is no general way to validate empirically a candidate set of costs for the dissimilarity measure proposed in (17). After a number of preliminary attempts aiming to obtain a measure able both to provide integer values and to guarantee the respect of the hierarchical assumptions described in Subsection 2.2, we set the costs in (17) as follows: $c_1 = 30$, $c_2 = 60$, $c_3 = 90$, $c_4 = 500$, $c_5 = c_6 = 1000$. However, we observe that this is not the only possible choice and that there exist alternative measures of dissimilarities that could be considered as plausible as the one proposed in this article. For example, an alternative measure of the dissimilarity between taxa $t_i, t_j \in D$, $i < j$, could be obtained by maximizing the likelihood related to the transition from a taxon t_i to t_j, under the hypothesis that the copy number of each gene changes independently of the others. This task could be performed e.g., by solving the following optimization problem:

$$\max d(t_i, t_j) = \prod_{g \in \mathcal{G}} p_u^{\alpha_{ij}^g} \cdot p_i^{\beta_{ij}^g} \cdot p_d^{\gamma_{ij}^g} \quad (19a)$$

$$s.t. \quad p_u + p_i + p_d = 1 \quad (19b)$$

$$p_s \geq 0 \qquad \forall\, s \in \{u, i, d\} \quad (19c)$$

where p_u is the probability that the copy number of gene g remains unchanged in the transition from taxon t_i to taxon t_j; p_i is the probability that the copy number of gene g increases in the transition from taxon t_i to taxon t_j; p_d is the probability that the copy number of gene g decreases in the transition from taxon t_i to taxon t_j; α_{ij}^g is a positive number equal to 1 if the copy number of gene g remains unchanged in the observed transition from taxon t_i to taxon t_j and 0 otherwise; β_{ij}^g is a positive number equal to ω_{ij}^g if the copy number of gene g increases in the observed transition from taxon t_i to taxon t_j and 0 otherwise; and γ_{ij}^g is a positive number equal to ω_{ij}^g if the copy number of gene g decreases in the observed transition from taxon t_i to taxon t_j and 0 otherwise.

The likelihood-based approach entails the quantification of the assumptions described in Subsection 2.2 and the expression of these assumptions in terms of constraints for Problem (19a)-(19c). We did not investigate the likelihood-based approach much further, as studying alternative measures of dissimilarity for the considered FISH data is not the main scope of the present article. Here, we merely observe that the likelihood-based approach and the measures of dissimilarities discussed in Subsection 3.2 are related. In fact, for fixed values

of the probabilities p_u, p_i and p_d, the logarithm base (e.g., p_i) of the objective function in the likelihood-based approach becomes

$$\sum_{g\in\mathcal{G}} \left(\alpha_{ij}^{g}\, log_{p_i}\,(p_u) + \beta_{ij}^{g} + \gamma_{ij}^{g}\, log_{p_i}\,(p_d)\right). \tag{20}$$

Whenever the logarithms are defined in (20), the addend $\alpha_{ij}^{g}\, log_{p_i}\,(p_u) + \beta_{ij}^{g} + \gamma_{ij}^{g}\, log_{p_i}\,(p_d)$ would add

$$\begin{cases} K_1 & \text{if } \left|t_i^g - t_j^g\right| = 0 \\ \left|t_i^g - t_j^g\right| & \text{if } \left|t_i^g - t_j^g\right| > 0 \\ \left|t_i^g - t_j^g\right| K_2 & if \left|t_i^g - t_j^g\right| < 0 \end{cases} \tag{21}$$

where $K_1 = \log_{p_i}(p_u)$ and $K_2 = \log_{p_i}(p_d)$ are two positive constants. This fact shows one relationship between the likelihood-based approach and, e.g., $d_1(\cdot, \cdot)$. Similar relationships exist also for $d_k(\cdot, \cdot)$, $k \in \{2, 3, 4\}$. Investigating further these relationships warrants additional analysis.

5 Results and discussion

Figures 1, 2 and 3 show examples of predicted progression trees for three paired DCIS/IDC cases, provided to illustrate some of the similarity and differences between inferred phylogenies by dataset. The remaining full trees are omitted in the main manuscript for brevity but can be found in the supplementary data downloadable at perso.uclouvain.be/daniele.catanzaro/SupportingMaterial/IDC.pdf.

A few general trends are apparent from manual examination of the trees. The progression of the IDC is characterized by an elevated tendency to lose copies of *TP53* (17.79% ± 5.11%) and other tumor suppressor genes (36% ± 5.35%). Moreover, the tumor suppressor genes have a high level of *spontaneous variation* (see Figure 4), i.e., the tendency of a single gene in a taxon to increase or decrease its copy number with respect to its immediate ancestor while the copy numbers of the remaining genes in both taxa are unchanged. *CDH1* shows the highest average rate of spontaneous variation (16.53 ± 6.59), followed by *TP53* (15.36 ± 5.18) and *DBC2* (14.23 ± 6.51). Among the oncogenes, *COX2* shows the highest level of spontaneous variation (13.19 ± 6.02), followed by *ZNF217* (10.75 ± 6.07), *CCND1* (10.13 ± 4.69), *HER2* (10.03 ± 3.33) and *MYC* (9.79 ± 3.87). Interestingly, *ZNF217* is more prone to spontaneous variations in datasets DAT07 and DAT09 and this phenomenon seems to be correlated to the variation induced by *CDH1* (see Figures 22 and 24 in supplementary data). The doubling-loss event is less frequent (7%±3.99%) than loss of tumor suppressor genes. It also appears to be more specific to particular datasets than others.

We offer a tentative classification of the datasets in Table 2, manually clustering the datasets based on apparent genetically distinct subsets of tumors. Specifically, we can distinguish between progressions showing a preponderance of doubling-loss phenomena (“Doubling-

Driven" column in Table 2) and progressions showing a low or absent presence of the doubling-loss event ("Doubling-Absent" column in Table 2). The first group is the largest and can be in turn subdivided in three main subgroups, namely: the *regular*, the *abnormal-with-*TP53*-predominance* and the *abnormal-with-*CDH1*-predominance*. The regular group is the largest subgroup and includes datasets DAT02 (Figure 2), DAT03, DAT04, DAT11, DAT12 and DAT13. As a general trend, this group shows a high spontaneous variation, usually affecting the tumor suppressor genes *CDH1* and *TP53*, which in general tend to be lost. Moreover, the copy numbers of the genes in general do not tend to increase with respect to the root node to a similar degree as is seen in the abnormal groups and usually do not exceed 8 copies. The doubling-loss event is more predominant in the regular group than in others, it usually tends to affect (almost) all genes (see, e.g., taxa ⟨2.2.2.1.2–1.1.1⟩ and ⟨4.4.4.2.4 – 2.2.2⟩, or ⟨2.3.2.1.2 – 1.1.1⟩ and ⟨4.6.4.2.4 – 1.2.2⟩ in Figure 2). It can be considered as a possible driving mechanism of progression of the carcinoma, being located in several internal vertices of corresponding predictions.

The abnormal-with-*TP53*-predominance subgroup includes datasets DAT08 and DAT06 (Figure 3). It is characterized by a very high spontaneous variation for *TP53* with frequent loss either of the tumor suppressor gene *CDH1* if *DBC2* is affected by high spontaneous variations or, vice versa, loss of the tumor suppressor gene *DBC2* if *CDH1* is affected by high spontaneous variations. The copy number of the genes tends to increase more than in the other categories, but usually does not exceed 8 copies. The instance DAT06 shows multiple situations in which some or all of the tumor suppressor genes are lost and this phenomenon usually comes together with a simultaneous loss of one or more oncogenes. A similar situation can be observed also in DAT08 (see supplementary data), although the loss of gene copies is less predominant. Interestingly, dataset DAT08 shows a very high level of variation of *TP53* with copy numbers ranging from 0 to 5. This fact seems to suggest the presence of a strong selective pressure acting on this gene. Similarly, in both datasets, *COX2* and *MYC* show a high level of variation, particularly *COX2*, although with a tendency towards gain or doubling events. The doubling-loss event is less preponderant in these tumors than in the regular subgroup and it is "abnormal" in the sense that it usually does not affect all of the genes but just some of the genes (see, e.g., taxa ⟨4.2.2.2.3–2.2.2⟩ and ⟨5.4.3.4.1 – 2.2.3⟩ in Figure 3 or taxa ⟨2.2.2.2.2 – 1.1.2⟩ and ⟨4.2.2.2.2 – 2.1.3⟩ in Figure 8 of supplementary data). Also in this case, the doubling-loss event can be considered as a possible source of progression of the carcinoma, being located in several internal vertices of corresponding predictions.

The abnormal-with-*CDH1*-predominance subgroup includes datasets DAT01 (Figure 1), DAT07, and DAT09. It is characterized by a very high spontaneous variation of *CDH1* with respect to *TP53*, high spontaneous variation of *CCND1* and *ZNF217*, and a low tendency to lose tumor suppressor genes. This subgroup shows the highest absolute gene copy numbers, with some single oncogenes showing copy numbers up to 25 in single cells (see e.g., DAT07 in supplementary data). The doubling-loss event is still present, although to a lesser degree than in the previous two subgroups.

Finally, the doubling-absent subgroup includes datasets DAT05 and DAT10. It is characterized by a very high spontaneous variation of *CDH1*, high spontaneous variation of

COX2 and *MYC*, and a low tendency to lose tumor suppressor genes. This subgroup does not show a significative presence of doubling-loss phenomena, although some genes (namely, *COX2* and *MYC*) tend to acquire high copy numbers with respect to the root, especially in DAT10. Specifically, in this dataset progression of the carcinoma seems to be caused by iterated increments of *COX2* over time, which seems to be associated with variation in the copy number of *CDH1*.

The apparent subgrouping of tumors suggests variations patient to patient, not just in random accumulation of mutations, but also in the phenotype for generating mutations. More specifically, distinct subsets of tumors show preferences for generation of aneuploidy versus amplification or loss of specific driver genes as well as in the selection of driver genes. These results are consistent with both a generic driver-passenger model, in which defined subtypes of tumors arise due to recurrent selection for malformations that produce specific selective advantages to tumors [47], [48], and with the mutator phenotype model, in which distinct subsets of tumors are driven by distinct mechanisms of generating genetic diversity [49]. In evolutionary terms, our results suggest intra-cellular heterogeneity is driven by variability in both mechanisms of diversification and selection for specific driver genes.

In principle, predicting tumor progression from single-cell sample data extracted from an individual is not a time-sensitive application. However, exploring the relationships between the computation times of Formulation (18) and the size of the instances of the problem may provide a better understanding of the computational performance of Formulation (18). As single-cell sample data are currently very difficult to obtain, we decided to considered a set of artificial instances of the problem characterized by 100, 150, 200, 250 and 300 taxa, respectively. For a fixed size (number of taxa) $t \in \{100, 150, 200, 250, 300\}$ we generated 20 random arborescences, each rooted in the healthy taxon and spanning t vertices. The algorithm used to generate a generic random arborescence T consisted of repeating the following steps: (i) given a vertex i, generate k children of i, k being a pseudorandom integer in [0, 10]; (ii) select a random vertex of T and repeat step (i) until t vertices are generated. In our experiments, we used the Mersenne twister library [50] as the pseudorandom generator. For each random arborescence T, we simulated the progression of tumor cells by randomly choosing on each arc (i, j) of T the type of mutation on vertex j (namely, copy number increment, copy number decrement and pure doubling) and the genes involved in the mutation. Specifically, we first set taxon j equal to its ancestor i. Subsequently, we set the probabilities of having a copy number increment, a copy number decrement and a pure doubling phenomenon roughly similar to the average frequencies of observing those phenomena in the considered real datasets (namely 0.84, 0.15 and 0.01, respectively). Then, we generated a random number r in [0, 1] and used this value to determine the type of mutation on taxon j. In particular, we used the following three cases: if $r \leq 0.84$ an increment of gene copy number arises; if $0.84 < r \leq 0.99$ a decrement of gene copy number arises; if $r > 0.99$ a pure doubling phenomenon arises. If the selected type of mutation is an increment or a decrement of gene copy number, the number of genes and the genes themselves subjected to the corresponding mutation are randomly chosen according to the following rules: (i) the number of selected genes cannot exceed 3; (ii) the ex nihilo nihil assumption has to be satisfied in the transition from taxon i to taxon j; (iii) the increment (or decrement) of gene copy number cannot exceed 2. If the selected type of mutation is a pure doubling,

then the copy numbers of all genes (but *TP53*) of vertex j are doubled and the copy number of *TP53* is increased by 1 unit. Finally, the nature of taxon j (i.e., whether it belongs to $\mathscr{DCIS}$, $\mathscr{ISI}$ or $\mathscr{IDC}$) is randomly determined by strictly respecting the temporal progression. Figure 5 shows the average computation times taken by Formulation (18) to solve the random instances so generated. The five average computation times fit a polynomial function of the number of taxa, with exponent approximately 2 (i.e., quadratic). At present, this is not a serious issue, as extracting large size single-cell sample datasets from patients is expensive and datasets containing more than 300 taxa are rarely analyzed. However, it reasonable to believe that in the coming years the cost for extracting single-cell sample data from patients will decrease; hence, datasets of larger and larger size will become more and more common. Therefore, investigating alternative ILP formulations able to reduce the solution time necessary to analyze a large dataset deserves further research efforts.

6 Conclusion

We proposed an approach to reconstruct a plausible progression of ductal carcinoma from single-cell sampled data of an affected individual. The approach is based on combining a generic parsimony model for evolutionary tree inference with a complex set of system-specific assumptions derived from prior knowledge regarding cellular atypia occurring in ductal carcinoma. ILP tools provide a way to describe and efficiently solve for optimal tree models in the presence of such complex constraints. Given the enormous variability between tumor types and study designs, our ILP strategy may have much broader utility for interpreting complex cellular variation data with reference to complicated system-specific biological constraints. While some aspects of our specific constraint sets are likely to be applicable across broad classes of tumor types (e.g., modeling the role of TP53 in chromosome instability or imposing timing constraints on defined pre-cancerous and cancerous progression stages), though, we do expect that some problem-specific expertise is likely to be needed to develop comparable approaches for other tumor progression systems. For example, we have studied another data collection of paired samples with single-cell FISH data on cervical cancer in which the pairs are from the primary tumor and from a metastasis [15]. For such a primary/metastasis study design, progression models could give insight into what evolutionary changes allow the metastatic sample to spread. The genes selected for FISH analysis in the cervical cancer data mostly differ from the genes in the DCIS/IDC, so one would need to adapt our modeling here to the cancer characteristics of the genes evaluated by FISH in that other data collection.

Our model suggests that progressions estimated from a population of 13 affected individuals are non-random and classifiable into several categories seemingly distinguished by distinct selective pressures and distinct mechanisms for generating genetic diversity. The complex and heterogeneous evolutionary landscape they reveal may have important implications for strategies for cancer treatment. In particular suggesting that developing effective therapies may require considering both the current spectrum of driver mutations in a tumor and the mutator phenotype by which it is generating diversity. For example, Martins et al. [11] point out that whether a breast cancer patient is a good candidate for drugs that inhibit PARP is

currently thought to depend on whether the patient's cancer genome has a defect in DNA damage repair leading to genomic instability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by the Belgian National Fund for Scientific Research (FRS-FNRS) (D.C), U.S. National Institutes of Health grants 1R01CA140214 (R.S.) and 1R01AI076318 (R.S.) and the Intramural Research Program of the NIH, NLM (A.A.S.). We thank the anonymous reviewers whose insightful suggestions helped us improve the manuscript.

Biography

Daniele Catanzaro received the PhD degree in Operations Research from the Free University of Brussels, Belgium, in 2008, for his studies in combinatorial optimization, medical bioinformatics and molecular phylogenetics. After being a Postdoctoral Fellow of the Belgian National Fund for Scientific Research (FNRS) and Assistant Professor at the Department of Operations of the Rijksuniversiteit Groningen, the Netherlands, he joined in 2014 the Louvain School of Management and the Center for Operations Research and Econometrics (CORE) of the Université Catholique de Louvain, where he is currently "Chargé de Cours" (Assistant Professor) of Operations Research.

Stanley E. Shackney received his M. D. in 1964. He was Senior Research Investigator and Attending Physician in the Medical Oncology Branch at the NIH from 1966 to 1983, and Director at the Laboratory of Cancer Cell Biology and Genetics at the Allegheny-Singer Research Institute in Pittsburgh, PA from 1985 to 2008. He later founded the company Intelligent Oncotherapeutics, Inc. He co-authored nearly 100 published papers and made significant contributions to the emerging field of tumor heterogeneity and evolution. He passed away on July 13, 2014.

Alejandro A. Schäffer received his B. S. in Applied Mathematics and his M.S. in mathematics from Carnegie Mellon University in 1983. He received his PhD in Computer Science from Stanford University in 1988, focusing on theoretical computer science. In 1992, he switched his research focus to software for genetics. He is best known for leading the development of the genetic linkage analysis package FASTLINK and for doing the implementation of the PSI-BLAST module of the sequence analysis package BLAST. He has also carried out genomic data analysis as one member of large teams doing medical genetics studies, especially studies identifying genes that when mutated cause human primary immunodeficiencies. Dr. Schäffer is currently a Computer Scientist at the National Center for Biotechnology Information, National Institutes of Health.

Russell Schwartz Russell Schwartz received his BS, MEng, and PhD degrees from the Department of Electrical Engineering and Computer Science at the Massachusetts of Technology, the last in 2000. He later worked in the Informatics Research group at Celera Genomics. He joined the faculty of Carnegie Mellon University in 2002, where he has pursued a variety of research directions in the field of computational biology, with special interest in computational genetics and computational biophysics. He is currently a Professor in the Carnegie Mellon Department of Biological Sciences and Computational Biology Department.

References

[1]. von Minckwitz G, Darb-Esfahani S, Loibl S, Huober J, Tesch H, Solbach C, Holms F, Eidtmann H, Dietrich K, Just M, Clemens M, Hanusch C, Schrader I, Henschen S, Hoffmann G, Tiemann K, Diebold K, Untch M, Denkert C. Responsiveness of adjacent ductal carcinoma in situ and changes in HER2 status after neoadjuvant chemotherapy/trastuzumab treatment in early breast cancer: Results from the GeparQuattro study (GBG 40). Breast Cancer Research and Treatment. 2012; 132:863–870. [PubMed: 21667238]

[2]. Sgroi DC. Preinvasive breast cancer. Annual Reviews of Pathology. 2010; 5:193–221.

[3]. Virnig BA, Tuttle TM, Shamliyan T, Kane RL. Ductal carcinoma in situ of the breast: A systematic review of incidence, treatment, and outcomes. Journal of the National Cancer Institute. 2010; 102(no. 3):170–178. [PubMed: 20071685]

[4]. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA A Cancer Journal for Clinicians. 2011; 61:69–90. [PubMed: 21296855]

[5]. Fenton JJ, Xing G, Elmore JG, Bang H, Chen SL, Lindfors KK, Baldwin L. Short-term outcomes of screening mammography using computer-aided detection: A population-based study of medicare enrollees. Annals of Internal Medicine. 2013; 158(no. 8):580–587. [PubMed: 23588746]

[6]. Pennington, G.; Smith, CA.; Shackney, S.; Schwartz, R. Cancer phylogenetics from single-cell assays. Department of Biological Sciences, Carnegie Mellon University; 2006. Tech. Rep.

[7]. — — . Reconstructing tumor phylogenies from heterogeneous single-cell data. Journal of Bioinformatics and Computational Biology. 2006; 5(no. 2a):407–427.

[8]. Frumkin D, Wasserstrom A, Itzkovitz S, Stern T, Harmelin A, Eilam R, Rechavi G, Shapiro R. Cell lineage analysis of a mouse tumor. Cancer Research. 2008; 68:5924–5931. [PubMed: 18632647]

[9]. Letouzé E, Allory Y, Bollet MA, Radvanyi F, Guyon F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. Genome Biology. 2010; 11:R76. [PubMed: 20649963]

[10]. Riester M, Attolini CS-O, Downey RJ, Singer S, Michor F. A differentiation-based phylogeny of cancer subtypes. PLoS Computational Biology. 2010; 6:e100077.

[11]. Martins FC, De S, Almendro V, Gönen M, Park SY, Blum JL, Herlihy W, Ethingthon G, Schnitt SJ, Tung N, Garber JE, Fetten K, Michor F, Polyak K. Evolutionary pathways in BRCA1-associated breast tumors. Cancer Discovery. 2012; 2(no. 6):503–511. [PubMed: 22628410]

[12]. Podlaha O, Riester M, De S, Michor F. Evolution of the cancer genome. Cell. 2012; 28(no. 4): 155–163.

[13]. Shlush LI, Chapal-Ilani N, Adar R, Pery N, Maruvka Y, Spiro A, Schouval R, Rowe JM, Tzukerman M, Bercovich D, Izraeli S, Marcucci G, Bloomfield CD, Zuckerman T, Skorecki K, Shapiro E. Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. Blood. 2012; 120:603–612. [PubMed: 22645183]

[14]. Kim KI, Simon R. Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinformatics. 2014; 15:27. [PubMed: 24460695]

[15]. Chowdhury SA, Shackney S, Heselmeyer-Haddad K, Ried T, Schäffer AA, Schwartz R. Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. Bioinformatics. 2013; 29(no. 13):i198.

[16]. — — . Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. PLoS Computational Biology. 2014; 10(no. 7):e1003740. [PubMed: 25078894]

[17]. Attolini CSO, Michor F. Evolutionary theory of cancer. Annals of the New York Academy of Sciences. 2009; 1168:23–51. [PubMed: 19566702]

[18]. Beerenwinkel N, Schwartz RF, Gerstung M, Markowetz F. Cancer evolution: Mathematical models and computational inference. Systematic Biology. 2015; (no. 1):e1–e25. [PubMed: 25293804]

[19]. Kainu T, Juo S, Desper R, Schäffer AA, Gillanders E, Rozenblum E, Freas-Lutz D, Weaver D, Stephan D, Bailey-Wilson J, Kallioniemi O, Tirkkonen M, Syryjäkoski K, Kuukasjärvi T, Koivisto P, Karhu R, Holli K, Arason A, Johannesdottir G, Bergthrosson JT, Johannsdottir H, Egilsson V, Barkardottir RB, Johannsson O, Haraldsson K, Sandberg T, Holmberg E, Grönberg H, Olsson H, Borg Å, Vehmanen P, Eerola H, Heikkila P, Pyrhönen S, Nevanlinna H. Somatic deletions in hereditary breast cancers implicate 13q21 as a putative novel breast cancer susceptibility locus. Proceedings of the National Academy of Science USA. 2000; 97:9603–9608.

[20]. Navin N, abd JK, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472(no. 7341):90–94. [PubMed: 21399628]

[21]. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B, Wigler M, Navin N, Hicks J. Genome-wide copy number analysis of single cells. Nature Protocols. 2012; 7:1024–1041. [PubMed: 22555242]

[22]. Zare H, Wang J, Hu A, Weber K, Smith J, Nickerson D, Song C, Witten D, Blau CA, Noble WS. Inferring clonal composition from multiple sections of a breast cancer. PLoS Computational Biology. 2014; 10:e1003703. [PubMed: 25010360]

[23]. Heselmeyer-Haddad K, Garcia LYB, Bradley A, Ortiz-Melendez C, Lee WJ, Christensen R, Prindiville SA, Calzone KA, Soballe PW, Hu Y, Chowdhury SA, Schwartz R, Schäffer AA, Ried T. Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity, yet conserved genomic imbalances and gain of MYC during progression. American Journal of Pathology. 2012; 181(no. 5):1807–1822. [PubMed: 23062488]

[24]. Almendro V, Cheng YK, Randles A, Itzkovitz S, Marusyk A, Ametller E, Gonzalez-Farre X, Munoz M, Russnes HG, Helland E, Rye IH, Borresen-Dale AL, Maruyama R, van Oudenaarden A, Dowsett M, Jones RL, Reis-Filho J, Gascon P, Gönen M, Michor F, Polyak K. Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of cellular diversity for genetic and phenotypic features. Cell Reports. 2014; 6(no. 3):514–527. [PubMed: 24462293]

[25]. Pennisi E. Single-cell sequencing tackles basic and biomedical questions. Science. 2012; 336(no. 6048):976–977. [PubMed: 22628633]

[26]. Marusyk A, Polyak K. Tumor heterogeneity: Causes and consequences. Biochim Biophys Acta (BBA)-Reviews on Cancer. 2010; 1805(no. 1):105–117. [PubMed: 19931353]

[27]. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez PP, Matthews N, Stewart A, Tarpey P, Varela I, Phillimore B, Begum S, McDonald NQ, Butler A, Jones D, Raine K, Latimer C, Santos CR, Nohadani M, Eklund AC, Spencer-Dene B, Clark G, Pickering L, Stamp G, Gore M, Szallasi Z, Downward J, Futreal PA, Swanton C. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine. 2012; 366(no. 10):883–892. [PubMed: 22397650]

[28]. Sottoriva A, Spiteri I, Piccirillo SG, Touloumis A, Collins VP, Marioni JC, Curtis C, Watts C, Tavaré S. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. Proceedings of the National Academy of Sciences USA. 2010; 110(no. 10):4009–4014.

[29]. Snuderl M, Fazlollahi L, Le LP, Nitta M, Zhelyazkova BH, Davidson CJ, Akhavanfard S, Cahill DP, Aldape KD, Betensky RA, Louis DN, Iafrate AJ. Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. Cancer Cell. 2011; 20(no. 6):810–817. [PubMed: 22137795]

[30]. Szerlip NJ, Pedraza A, Chakravarty D, Azim M, McGuire J, Fang Y, Ozawa T, Holland EC, Huse JT, Jhanwar S, Leversha MA, Mikkelsen T, Brennan CW. Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopul ations with distinct growth factor response. Proceedings of the Nationall Academy of Sciences USA. 2012; 109(no. 8):3041–3046.

[31]. Sprouffske K, Pepper JW, Maley CC. Accurate reconstruction of the temporal order of mutations in neoplastic progression. Cancer Prevention Research. 2011; 4(no. 7):1135–1144. [PubMed: 21490131]

[32]. Farahani HS, Lagergren J. Learning oncogenetic networks by reducing to MILP. PLoS ONE. 2013; 8:e65773. [PubMed: 23799047]

[33]. Sontag L, Axelrod DE. Evaluation of pathways for progression of heterogeneous breast tumors. Journal of Theoretical Biology. 2005; 232(no. 2):179–189. [PubMed: 15530488]

[34]. Lin S. Mixture modeling of progression pathways of heterogeneous breast tumors. Journal of Theoretical Biology. 2007; 249(no. 2):254–261. [PubMed: 17892886]

[35]. Beerenwinkel N, Rahnenführer J, Kaiser R, Hoffmann D, Selbig J, Lengauer T. Mtreemix: A software package for learning and using mixture models of mutagenetic trees. Bioinformatics. 2005; 21(no. 9):2106–2107. [PubMed: 15657098]

[36]. Tolliver D, Tsourakakis C, Subramanian A, Shackney S, Schwartz R. Robust unmixing of tumor states in array comparative genomic hybridization data. Bioinformatics. 2010; 26(no. 12):89–102.

[37]. Efeyan A, Serrano M. P53: Guardian of the genome and policeman of the oncogenes. Cell Cycle. 2007; 6(no. 9):1006–1010. [PubMed: 17457049]

[38]. Wong KK, de Leeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL. A comprehensive analysis of common copy-number variations in the human genome. The American Journal of Human Genetics. 2007; 80(no. 1):91–104. [PubMed: 17160897]

[39]. Albert, VA. Parsimony, phylogeny, and genomics. Oxford University Press; Oxford, UK: 2005.

[40]. Beyer WA, Stein M, Smith T, Ulam S. A molecular sequence metric and evolutionary trees. Mathematical Biosciences. 1974; 19:9–25.

[41]. Waterman MS, Smith TF, Singh M, Beyer WA. Additive evolutionary trees. Journal of Theoretical Biology. 1977; 64:199–213. [PubMed: 839800]

[42]. Catanzaro, D. Estimating phylogenies from molecular data. In: Bruni, R., editor. Mathematical approaches to polymer sequence analysis and related problems. Springer; New York: 2011.

[43]. Felsenstein, J. Inferring Phylogenies. Sinauer Associates; Sunderland, MA: 2004.

[44]. Bourbaki, N. Topological Vector Spaces. Springer; New York, NY: 2002.

[45]. Ahuja, RK.; Magnanti, TL.; Orlin, JB. Network Flows: Theory, Algorithms, and Applications. Prentice Hall; United States edition: 1993.

[46]. Martin, RK. Large scale linear and integer optimization: A unified approach. Springer; New York: 1999.

[47]. Hanahan D, Weinberg R. The hallmarks of cancer. Cell. 2000; 100:57–70. [PubMed: 10647931]

[48]. — — . The hallmarks of cancer: The next generation. Cell. 2011; 144:646–674. [PubMed: 21376230]

[49]. Loeb L. Mutator phenotype may be required for multistage carcinogenesis. Cancer Research. 1991; 51:3075–3079. [PubMed: 2039987]

[50]. Matsumoto M, Nishimura T. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation. 1998; 8(no. 1):3–30.

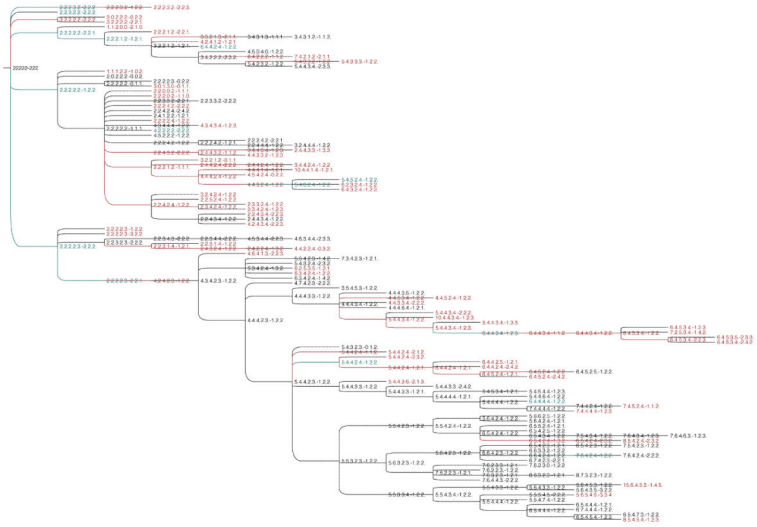

Fig. 1.

Predicted progression from DCIS to IDC for the dataset DAT01. For ease of interpretation, each sample is represented by a sequence of numbers separated by dots and a dash. Specifically, the first five numbers represent the copy numbers for the oncogenes *COX2*, *MYC*, *HER2*, *CCND1* and *ZNF217* and the last three numbers represent the copy numbers for the tumor suppressor genes *DBC2*,*CDH1* and *TP53*. The samples in black and red refer to DCIS or IDC single-cells, respectively. The samples in blue refer to single-cells found in both DCIS and IDC.

Fig. 2.
Predicted progression from DCIS to IDC for the dataset DAT02.

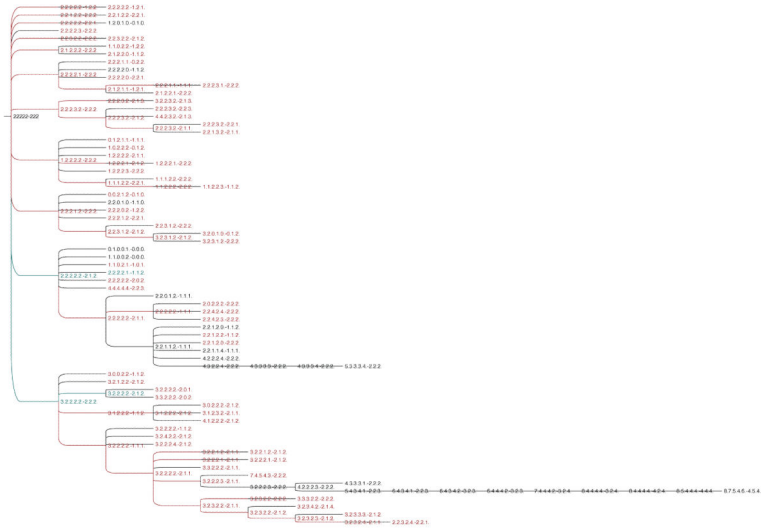

Fig. 3.
Predicted progression from DCIS to IDC for the dataset DAT06.

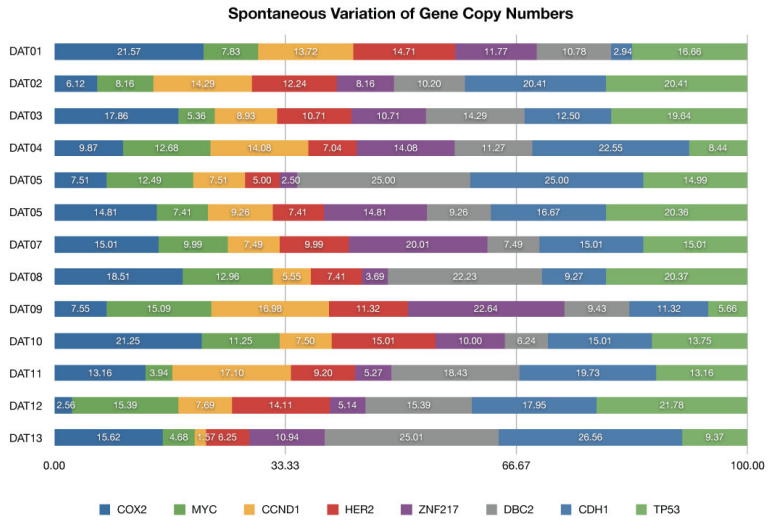

Fig. 4.

Predicted spontaneous variation of the gene copy number (expressed in percentage) in the considered datasets.

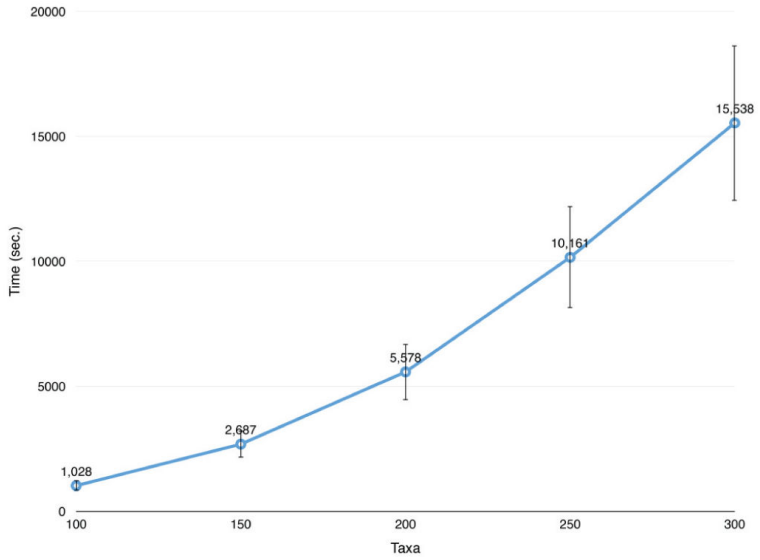

Fig. 5.

Average computation times taken by Formulation (18) to solve instances of the problem containing 100, 150, 200, 250 and 300 taxa, respectively. The vertical bars indicate the standard deviations.

TABLE 1

Characteristics of the analyzed datasets.

Dataset	Number of DCIS taxa	Number of IDC taxa	Overall number of distinct taxa
DAT01	123	119	207
DAT02	35	76	100
DAT03	79	69	115
DAT04	102	120	181
DAT05	58	57	77
DAT06	33	93	118
DAT07	84	76	128
DAT08	92	69	109
DAT09	71	69	137
DAT10	126	102	138
DAT11	77	99	146
DAT12	92	124	172
DAT13	57	97	114

TABLE 2

A possible classification of the analyzed datasets.

Doubling-Driven			**Doubling-Absent**
Regular	**Abnormal**		
	***TP53*-driven**	***CDH1*-driven**	
DAT02	DAT06	DAT01	DAT05
DAT03	DAT08	DAT07	DAT10
DAT04		DAT09	
DAT11			
DAT12			
DAT13			