Sorting cancer karyotypes by elementary operations

Michal Ozery-Flato and Ron Shamir

School of Computer Science, Tel-Aviv University, Tel Aviv 69978, Israel
{ozery,rshamir}@post.tau.ac.il

Abstract. Since the discovery of the "Philadelphia chromosome" in chronic myelogenous leukemia in 1960, there is an ongoing intensive research of chromosomal aberrations in cancer. These aberrations, which result in abnormally structured genomes, became a hallmark of cancer. Many studies give evidence to the connection between chromosomal alterations and aberrant genes involved in the carcinogenesis process. An important problem in the analysis of cancer genomes, is inferring the history of events leading to the observed aberrations. Cancer genomes are usually described in form of *karyotypes*, which present the global changes in the genomes' structure. In this study, we propose a mathematical framework for analyzing chromosomal aberrations in cancer karyotypes. We introduce the problem of sorting karyotypes by elementary operations, which seeks for a shortest sequence of elementary chromosomal events transforming a normal karyotype into a given (abnormal) cancerous karyotype. Under certain assumptions, we prove a lower bound for the elementary distance, and present a polynomial-time 3-approximation algorithm. We applied our algorithm to karyotypes from the Mitelman database, which records cancer karyotypes reported in the scientific literature. Approximately 94% of the karyotypes in the database, totalling 57,252 karyotypes, supported our assumptions, and each of them was subjected to our algorithm. Remarkably, even though the algorithm is only guaranteed to generate a 3-approximation, it produced a sequence whose length matches the lower bound (and hence optimal) in 99.9% of the tested karyotypes.

Introduction

Cancer is a genetic disease, caused by genomic mutations leading to the aberrant function of genes. Those mutations ultimately give cancer cells their proliferative nature. Hence, inferring the evolution of these mutations is an important problem in the research of cancer. Chromosomal mutations that shuffle/delete/duplicate large genomic fragments are common in cancer. Many methods for detection of chromosomal mutations use chromosome painting techniques, such as G-banding, to achieve a visualization of cancer cell genomes. The description of the observed genome organization is called a *karyotype* (see Fig. 1). In a karyotype, each chromosome is partitioned into continuous genomic regions called *bands*, and the total number of bands is the *banding resolution*. Over the last decades, a large amount of data has been accumulated on cancer karyotypes. One of the largest depositories of cancer karyotypes is the Mitelman database of chromosomal aberrations in cancer [9], which records cancer karyotypes reported in the scientific literature. These karyotypes are described using the ISCN nomenclature [8], and thus can be parsed automatically.

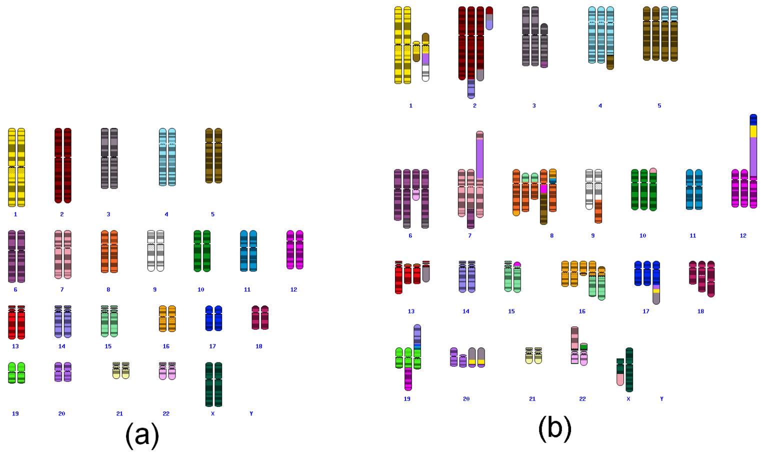


Fig. 1. A schematic view of two real karyotypes: a normal female karyotype (a), and the karyotype of MCF-7 breast cancer cell-line (b) [1]. In the normal karyotype, all chromosomes, except X and Y, appear in two identical copies, and each chromosome has a distinct single color. In the cancer karyotype presented here, only chromosomes 11,14, and 21 show no chromosomal aberrations.

Cancer karyotypes exhibit a wide range of chromosomal aberrations. The common classification of these aberrations categorizes them into a variety of specific types, such as translocations, iso-chromosomes, etc. Inferring the evolution of cancer karyotypes using this wide vocabulary of complex alteration patterns is a difficult task. Nevertheless, the entire spectrum of chromosomal alterations can essentially be spanned by four elementary operations: breakage, fusion, duplication, and deletion (Fig. 2). A *breakage*, formally known as a “double strand break”, cuts a chromosomal fragment into two. A *fusion* ligates two chromosomal fragments into one. Genomic breakages, which occur quite frequently in our body cells, are normally repaired by the corresponding inverse fusion. Mis-repair of genomic breakages is believed to be a major cause of chromosomal aberrations in cancer [4]. Other prevalent chromosomal alterations in cancer genomes are *duplications* and *deletions* of chromosomal fragments. These four elementary events play a significant role in carcinogenesis: fusions and duplications can activate oncogenes, while breakages and deletions can eliminate tumor suppressor genes.

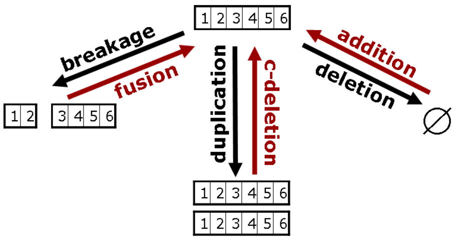


Fig. 2. Illustrations of elementary operations: breakage, fusion, duplication, and deletion. The inverse elementary operations are: fusion, breakage, c-deletion, and addition.

Based on the four elementary operations presented above, we introduce a new model for analyzing chromosomal aberrations in cancer. We study the problem of finding a shortest sequence of operations that transforms a normal karyotype into a given cancer karyotype. This is the problem of *karyotype sorting by elementary operations* (KS), and the length of a shortest sequence is the *elementary distance* between the normal and cancer karyotypes. The elementary distance indicates how far, in terms of number of operations, a cancer karyotype is from the normal one, and is *not* a metric in the mathematical sense. The elementary distance corresponds to the complexity of the cancer karyotype, which may give an indication of the tumor phase [6]. The reconstructed elementary operations can be used to detect common events for a set of cancer karyotypes, and thus point out to genomic regions suspect of containing genes associated with carcinogenesis.

Under certain assumptions, which are supported by most cancer karyotypes, the KS problem can be reduced in linear time to a simpler problem, called RKS. For the latter problem we prove a lower bound for the elementary distance, and present a polynomial-time 3-approximation algorithm. We show that approximately 94% of the karyotypes in the Mitelman database (57,252) support our assumptions, and each of these was subjected to our algorithm. Remarkably, even though the algorithm is only guaranteed to generate a 3-approximation, it produced a sequence whose length matches the lower bound (and hence optimal) in 99.9% of the tested karyotypes. Manual inspection of the remaining cases reveals that the computed sequence for each of these cases is also optimal.

The paper is organized as follows. In Section 1 we give the combinatorial formulation of the KS problem and its reduced variant RKS. In the rest of the paper we focus on the RKS problem. In Section 2 we prove a lower bound for the elementary distance for RKS. Section 3 describes our 3-approximation algorithm for RKS. Finally, in Section 4 we present the results of the application of our algorithm to the karyotypes in the Mitelman database. Due to space limits, most proofs are omitted.

1 Problem formulation

1.1 The KS problem

The KS problem receives two karyotypes as an input: K_{normal}, and the cancer karyotype, K_{cancer}. We represent each of the two karyotypes by a multi-set of chromosomes. Every chromosome in K_{normal} is presented as an interval of B integers, where each integer represents a *band*. For simplicity we assume that all the chromosomes in K_{normal} share the same B, which corresponds to the banding resolution. Every two chromosomes in the normal karyotype are either identical, i.e. are represented by the same interval, or disjoint. More precisely, we represent every chromosome in K_{normal} by the interval $[(k-1)B+1, kB]$, where k is an integer that identifies the chromosome. The normal karyotype usually contains exactly two copies of each chromosomes, with the possible exception of the sex chromosomes. Every chromosome in K_{cancer} is either a fragment or a concatenation of several fragments, where a *fragment* is a maximal sub-interval, with two bands or more, of a chromosome in the normal karyotype. More formally, a fragment is a maximal interval of the karyotype of the form $[i,j] \equiv [i, i+1, \ldots, j]$, or $[j,i] \equiv [j, j-1, \ldots, i]$, where $i < j$, $i, j \in \{(k-1)B+1, \ldots, kB\}$, and $[(k-1)B+1, kB] \in K_{\text{normal}}$. Note that in particular, a chromosome in K_{cancer} can be identical to a chromosome in K_{normal}. We use the symbol "::" to denote a concatenation between two fragments, e.g., $[i,j]::[i',j']$. Every chromosome, in both K_{normal} and K_{cancer}, is orientation-less, i.e. reversing the order of the fragments, along with their orientation, results in an equivalent chromosome. For example, $X = [i,j]::[i',j'] \equiv [j',i']::[j,i] = \overline{X}$.

We refer to the concatenation point of two intervals as an *adjacency* if the union of their intervals is equivalent to a larger interval in K_{normal}. In other words, two concatenated intervals that form an adjacency can be replaced by one equivalent interval. For example, the concatenation point in $[5,3]::[3,1] \equiv [5,1]$ is an adjacency. Typically, a breakage occurs within a band, and each of the resulting fragments contains a visible piece of this broken band. For example, if $[5,1]$ is broken within band 3, then the resulting fragments are generally denoted the by $[5,3]$ and $[3,1]$. For this reason, we do *not* consider the concatenation $[5,3]::[2,1]$ as an adjacency. A concatenation point that is *not* an adjacency, is called a *breakpoint*[1]. Further examples of concatenation points that are breakpoints are: $[1,3]::[5,6]$ and $[2,4]::[4,3]$.

We assume that the cancer karyotype, K_{cancer}, has evolved from the normal karyotype, K_{normal}, by the following four *elementary operations*:

I. **Fusion**: a concatenation of two chromosomes, X_1 and X_2, into one chromosome $X_1::X_2$.

II. **Breakage**: a split of a chromosome into two chromosomes. A split can occur within a fragment, or between two previously concatenated fragments,

[1] Formally, since the broken ends of a chromosome are not considered breakpoints here, the term "fusion-point" may seem more appropriate. However, we kept the name "breakpoint" due to its prior use and brevity.

i.e. in a breakpoint. In the former case, where the break is in a fragment $[i, j]$, the fragment is split into two complementing fragments: $[i, k]$ and $[k, j]$, where $k \in \{i+1, i+2, \ldots, j-1\}$.

III. **Duplication**: a whole chromosome is duplicated, resulting in two identical copies of the original chromosome.

IV. **Deletion**: a complete chromosome is deleted from the karyotype.

Given K_{normal} and K_{cancer}, we define the KS problem as finding a shortest sequence of elementary operations that transforms K_{normal} into K_{cancer}. An equivalent formulation of the KS problem is obtained by considering the inverse direction: find a shortest sequence of *inverse* elementary operations that transforms K_{cancer} into K_{normal}. Clearly, fusion and breakage operations are inverse to each other. The inverse to a duplication is a *constrained deletion* (abbreviated *c-deletion*), where the deleted chromosome is one of two or more identical copies. In other words, a c-deletion can delete a chromosome only if there exists another identical copy of it. The inverse of a deletion is an *addition* of a chromosome. Note that in general, the added chromosome need not be a duplicate of an existing chromosome and can contain any number of fragments. For the rest of the paper, we analyze KS by sorting in reverse order, i.e. starting from K_{cancer} and going back to K_{normal}. The sorting sequences will also start from K_{cancer}.

1.2 Reducing KS to RKS

In this section we present a basic analysis of KS, which together with two additional assumptions, allows the reduction of KS to a simpler variant in which no breakpoint exists (RKS). As we shall see, our assumptions are supported by most analyzed cancer karyotypes.

We start with several definitions. A sequence of inverse elementary operations is *sorting*, if its application to K_{cancer} results in K_{normal}. We shall refer to a shortest sorting sequence as *optimal*. Since every fragment contains two or more bands, we can present any band i within it by an ordered pair of its two ends, i^0, which is the end closer to the minimal band in the fragment, and i^1, the end closer to the maximal band in the fragment. More formally, we map the fragment $[i, j]$, $i \neq j$, to $[i^1, j^0] \equiv [i^1, (i+1)^0, (i+1)^1, \ldots, j^0]$ if $i < j$, and otherwise to $[i^0, j^1] \equiv [i^0, (i-1)^1, (i-1)^0, \ldots, j^1]$. We say that two fragment-ends, a and a', are *complementing* if $\{a, a'\} = \{i^0, i^1\}$. The notion of viewing bands as ordered pairs is conceptually similar to considering genes / synteny blocks as oriented, as is standard in the computational studies of genome rearrangements in evolution [3]. In this study we consider bands as ordered pairs to well identify breakpoints: as mentioned previously, a breakage usually occurs within a band, say i, and the two ends of i, i^0 and i^1, are separated between the two new resulting fragments. Thus, a fusion of two fragment-ends forms an adjacency iff these ends are complementing. We identify a breakpoint, and a concatenation point in general, by the two corresponding fragment-ends that are fused together. More formally, the concatenation point in $[a, b]$::$[a', b']$ is identified by the (unordered) pair $\{b, a'\}$. For example, the breakpoint in $[1, 2]$::$[4, 3] \equiv [1^1, 2^0]$::$[4^0, 3^1]$ is identified by $\{2^0, 4^0\}$. Note that the two other fragment-ends, 1 and 3, do not matter for that breakpoint's identity. Having defined breakpoint identities, we refer to a breakpoint as *unique* if no other breakpoint shares its identity, and otherwise we call it *recurrent*. In particular, a breakpoint in a non-unique chromosome (i.e., a chromosome with another identical copy) is recurrent. Last, we say that a chromosome X is *complex* if it contains at least one breakpoint, and *simple* otherwise. In other words, chromosome X is simple iff it consists of one fragment. Analogously, an addition is *complex* if the chromosome added is complex, and *simple* otherwise.

Observation 1 *Let S be an optimal sorting sequence. Suppose K_{cancer} contains a breakpoint, p, that is not involved in a c-deletion in S. Then there exists an optimal sorting sequence S', in which the first operation is a breakage of p.*

Proof. Since K_{normal} does not contain any breakpoint, p must be eventually eliminated by S. A breakpoint can be eliminated either by a breakage or by a c-deletion. Since p is not involved in a c-deletion, p is necessarily eliminated by a breakage. Moreover, this breakage can be moved to the beginning of S since no other operation preceding it involves p. □

Corollary 1 *Let S be an optimal sorting sequence. Suppose S contains an addition of chromosome $X = f_1{::}f_2{::}\dots{::}f_k$, where $f_1, f_2, \dots, f_k$ are fragments, and none of the $k-1$ breakpoints in X is involved in any subsequent c-deletion in S. Then the sequence S', obtained from S by replacing the addition of X with the additions of $f_1, f_2, \dots, f_k$ (overall k additions), is an optimal sorting sequence.*

Proof. By Observation 1, the breakpoints in X can be immediately broken after its addition. Thus replacing the addition of X, and the $k-1$ breakages following it, by k additions of $f_1, f_2, \dots, f_k$, yields an optimal sorting sequence. □

It appears that complex additions, as opposed to simple additions, makes KS very difficult to analyze. Moreover, based on Corollary 1, complex additions can be truly beneficial only in complex scenarios in which c-deletions involve recurrent breakpoints that were formerly created by complex additions. An analysis of a large collection of cancer karyotypes reveals that only 6% of the karyotypes contain recurrent breakpoints (see Section 4). Therefore, for the rest of this paper, we make the following assumption:

Assumption 1 *Every addition is simple, i.e., every added chromosome consists of one fragment.*

Using the assumption above, the following observation holds:

Observation 2 *Let p be a unique breakpoint in K_{cancer}. Then there exists an optimal sorting sequence in which the first operation is a breakage of p.*

Proof. If p is not involved in a c-deletion, then by Observation 1, p can be broken immediately. Suppose there are k c-deletions involving p or other breakpoints identical to it. Note that following Assumption 1, from the four inverse elementary operations, only fusion can create a new breakpoint. Now, any c-deletion involving p requires another fusion that creates a breakpoint p', identical to p. Thus we can obtain an optimal sorting sequence, S', from S, by: *(i)* first breaking p, *(ii)* not creating any breakpoint point p' identical to p, *(iii)* replacing any c-deletion involving p, or one of its copies, with two c-deletions of the corresponding 4 unfused chromosomes, and *(iii)* not having to break the last instance of p (since it was already broken). In summary, we moved the breakage of p to the beginning of the sorting sequence and replaced k fusions and k c-deletions (i.e. 2k operations) with $2k$ c-deletions. □

Observation 3 *In an optimal sequence, every fusion creates either an adjacency, or a recurrent breakpoint.*

Proof. Let S be an optimal sorting sequence. Suppose S contains a fusion that creates a new unique breakpoint p. Then, following Observation 2, p can be immediately broken after it was formed, a contradiction to the optimality of S. □

In this work, we choose to focus on karyotypes that do not contain recurrent breakpoints. According to our analysis of the Mitelman database, 94% of the karyotypes satisfy this condition. Thus we make the following additional assumption:

Assumption 2 *The cancer karyotype, K_{cancer}, does not contain any recurrent breakpoint.*

Assumption 2 implies that *(i)* we can immediately break all the breakpoints in K_{cancer} (due to Observation 2), and *(ii)* consider fusions only if they create an adjacency (due to Observation 3). Hence, for each unique normal chromosome, its fragments can be used separated from all the other fragments and used to solve a simpler variant of KS: In this variant, *(i)* $K_{\text{normal}} = \{[1, B] \times N\}$, *(ii)* there are no breakpoints in K_{cancer}, and *(iii)* neither fusions, nor additions, form breakpoints. Usually, $N = 2$. Exceptions are $N = 1$ for the sex chromosomes, and $N > 2$ for cases of global changes in the ploidy. We refer to this reduced problem as RKS (restricted KS). For the rest of the paper, we shall limit our analysis to RKS only.

2 A lower bound for the elementary distance

In this section we analyze RKS and define several combinatorial parameters that affect the elementary distance between K_{normal} and K_{cancer}, denoted by $d \equiv d(K_{\text{normal}}, K_{\text{cancer}})$. Based on these parameters, we prove a lower bound on the elementary distance. Though theoretically our lower bound is not tight, we shall demonstrate in Section 4 that in practice, for the vast majority (99.9%) of the real cancer karyotypes analyzed, the elementary distance to the appropriate normal karyotype achieves this bound.

2.1 Extending the karyotypes

For the simplicity of later analysis, we extend both K_{normal} and K_{cancer} by adding each $2N$ "tail" intervals:

$$\widehat{K}_{\text{normal}} = K_{\text{normal}} \cup \{[0,1] \times N, [B, B+1] \times N\}$$
$$\widehat{K}_{\text{cancer}} = K_{\text{cancer}} \cup \{[0,1] \times N, [B, B+1] \times N\}$$

These new "tail" intervals do not take part in elementary operations: breakage and fusion are still limited to $\{2, 3, \ldots, B-1\}$, and intervals added/c-deleted are contained in $[1, B]$. Hence $d(K_{\text{normal}}, K_{\text{cancer}}) \equiv d(\widehat{K}_{\text{cancer}}, \widehat{K}_{\text{cancer}})$. Their only role is to simplify the definitions of parameters given below.

2.2 The histogram

We define the *histogram* of $\widehat{K}_{\text{cancer}}$, $H \equiv H(\widehat{K}_{\text{cancer}}) : \{[i-1, i] \mid i = 1, 2, \ldots, B+1\} \to \mathbb{N} \cup \{0\}$, as follows. Let $H([i-1, i])$ be the number of fragments in $\widehat{K}_{\text{cancer}}$ that contain the interval $[i-1, i]$. See Fig. 3(b) for an example. From the definition of $\widehat{K}_{\text{cancer}}$, it follows that $H([0,1]) = H([B, B+1]) = N$. For simplicity we refer to $H([i-1, i])$ as $H(i)$. The histogram H has a *wall* at $i \in \{1, \ldots, B\}$ if $H(i) \neq H(i+1)$. If $H(i+1) > H(i)$ (respectively, $< H(i)$) then the wall at i is called a *positive* wall (respectively, a *negative* wall). Intuitively, a wall is a vertical boundary of H. We define w to be the total size of walls in H. More formally,

$$w = \sum_{i=1}^{B} |H(i+1) - H(i)|$$

Since $H(1) = H(B+1) = N$, the total size of positive walls is equal to the total size of negative walls, and hence w is even. Note that if $\widehat{K}_{\text{cancer}} = \widehat{K}_{\text{normal}}$ then $w = 0$. The pair $(i, h) \equiv (i, [h-1, h])$, $h \in \mathbb{N}$, is a *brick* in the wall at i if $H(i) + 1 \leq h \leq H(i+1)$ or $H(i+1) + 1 \leq h \leq H(i)$. A brick (i, h) is *positive* (respectively, *negative*) if the wall at i is positive (respectively, negative). Note that the number of bricks in a wall is equal to its total size. Hence w corresponds to the total number of bricks in H.

Observation 4 *For a breakage/fusion, $\Delta w = 0$; For a c-deletion/addition, $\Delta w = \{-2, 0, 2\}$.*

2.3 Counting complementing end pairs

Consider the case where $w = 0$. Then there are no gains and no losses of bands, and the number of fragments in $\widehat{K}_{\text{cancer}}$ is greater or equal to the number of fragments in $\widehat{K}_{\text{normal}}$. Note that each of the four elementary operations can decrease the total number of fragments by at most one. Hence when $w = 0$, an optimal sorting sequence would be to fuse pairs of complementing fragment-ends, not including the tails. Let us define $f \equiv f(\widehat{K}_{\text{cancer}})$ as the maximum number of disjoint pairs of complementing fragment-ends. Note there could be many alternative choices of complementing pairs. Nevertheless, any maximal disjoint pairing is also maximum. It follows that if $w = 0$, then $d(\widehat{K}_{\text{normal}}, \widehat{K}_{\text{cancer}}) = f - 2N$. Also, when $w \neq 0$, a c-deletion may need to be preceded by some fusions of complementing ends, to form two identical fragments. In general, the following holds:

Observation 5 *For breakage* $\Delta f = 1$*; For fusion,* $\Delta f = -1$*; For c-deletion,* $\Delta f \in \{0, -1, -2\}$*; For addition,* $\Delta f \in \{0, 1, 2\}$.

Lemma 1 *For breakage/addition,* $\Delta(w/2+f) = 1$*; For fusion/c-deletion,* $\Delta(w/2+f) = -1$.

2.4 Simple bricks

A brick (i, h) is *simple* if: *(i)* $(i, h-1)$ is not a brick, and *(ii)* $\widehat{K}_{\text{cancer}}$ does not contain a pair of complementing fragment-ends in i. Thus, in particular, a simple brick cannot be eliminated by a c-deletion. On the other hand, for a non-simple brick, (i, h), there are two fragments ending in the corresponding location (i.e. i). Nevertheless, it may still be impossible to eliminate (i, h) by a c-deletion if these two fragments are not identical. We define $s \equiv s(\widehat{K}_{\text{cancer}})$ as the number of simple bricks.

Observation 6 *For breakage,* $\Delta s \in \{0, -1\}$*; For fusion,* $\Delta s \in \{0, 1\}$*; For c-deletion,* $\Delta s = 0$*; For addition,* $|\Delta s| \leq 2$.

2.5 The weighted bipartite graph of bricks

We now define the last parameter in the lower bound formula for the elementary distance. It is based upon matching pairs of bricks, where one is positive and the other is negative. Note that in the process of sorting $\widehat{K}_{\text{cancer}}$, the histogram is flattened, i.e., all bricks are eliminated, which can be done only by using c-deletion/addition operations. Observe that if a c-deletion/addition eliminates a pair of bricks, then one of these bricks is positive and the other is negative. Thus, roughly speaking, every sorting sequence defines a matching between pairs of positive and negative bricks that are eliminated together.

Given two bricks, $v = (i, h)$ and $v' = (i', h')$, we write $v < v'$ (resp. $v = v'$) if $i < i'$ (resp. $i = i'$). Let V^+ and V^- be the sets of positive and negative bricks respectively. We say that v and v' have the same *sign*, if either $v, v' \in V^+$, or $v, v' \in V^-$. Two bricks have the same *status* if they are either both simple, or both non-simple. Let $BG = (V^+, V^-, \delta)$ be the weighted complete bipartite graph, where $\delta : V^+ \times V^- \rightarrow \{0, 1, 2\}$ is an edge-weight function defined as follows. Let $v^+ \in V^+$ and $v^- \in V^-$ then:

$$\delta(v^+, v^-) = \begin{cases} 0 & v^+ \text{ and } v^- \text{ are both simple } \textbf{and } v^- < v^+ \\ 0 & v^+ \text{ and } v^- \text{ are both non-simple } \textbf{and } v^+ < v^- \\ 1 & v^+ \text{ and } v^- \text{ have opposite status} \\ 2 & \text{otherwise} \end{cases}$$

For an illustration of BG see Fig 3(c). A *matching* is a set of vertex-disjoint edges from $V^+ \times V^-$. A matching is *perfect* if it covers all the vertices in BG (recall that $|V^+| = |V^-|$). Thus a perfect matching is in particular a maximum matching. Given a matching M, we define $\delta(M)$ as the total weight of its edges. Let $m \equiv m(\widehat{K}_{\text{cancer}})$ denote the minimum weight of a perfect matching in BG. The problem of finding a minimum-weight perfect matching can be solved in polynomial-time (see, e.g., [7, Theorem 11.1]). We note there exists a simple efficient algorithm for computing m, which relies heavily on the specific weighting scheme, δ.

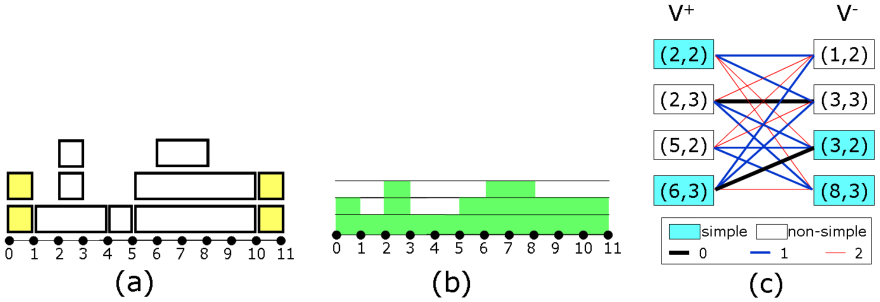


Fig. 3. An example of an (extended) cancer karyotype $\widehat{K}_{\text{cancer}}$ and its combinatorial parameters. (a) The (extended) cancer karyotype is $\widehat{K}_{\text{cancer}} = \{[0,1] \times 2, [1,4], [4,5], [5,10] \times 2, [10,11] \times 2, [2,3] \times 2, [6,8]\}$. Here $N = 2$, $B = 10$. The number of disjoint pairs of complementing fragment-ends, f, is 5. (b) The histogram $H \equiv H(\widehat{K}_{\text{cancer}})$, which has walls at 1, 2, 3, 5, 6, and 8. There are four positive bricks: $(2,2)$, $(2,3)$, $(5,2)$, and $(6,3)$, and four negative bricks: $(1,2)$, $(3,3)$, $(3,2)$, and $(8,3)$. Hence $w = 8$. Four of the eight bricks are simple: $(2,2)$, $(3,2)$, $(6,3)$, and $(8,3)$, thus $s = 4$. (c) The weighted-bipartite graph of *BG*. It is not hard to verify that $M = \{\ ((2,3),(3,3)),\ ((6,3),(3,2)),\ ((2,2),(1,2)),\ ((5,2),(8,3))\ \}$ is a minimum-weight perfect matching and hence $m = 2$.

Let K' be obtained from K by an elementary operation (a move). For a function F defined on karyotypes, define $\Delta(F) = F(K') - F(K)$.

Lemma 2 $d \geq w/2 + f - 2N + s + m \geq 0$.

3 The 3-approximation algorithm

Algorithm 1 below is a polynomial procedure for the RKS problem. We shall prove that it is a 3-approximation, and then describe a heuristic that aims to improve it.

Lemma 3 *Algorithm 1 transforms $\widehat{K}_{cancer}$ into $\widehat{K}_{normal}$ using at most $3w/2 + f - 2N + s + m$ inverse elementary operations.*

Algorithm 1 Elementary Sorting (RKS)

```
1: $M \leftarrow$ a minimum-weight perfect matching in $BG$
2: for all $(v^-, v^+) \in M$ where $v^- < v^+$ do
3:    Add the interval $[v^-, v^+]$.
4: end for /* Now $v^+ < v^-$ for every $(v^+, v^-) \in M$, where $v^+ \in V^+$, $v^- \in V^-$ */
5: for all $v \in V^+ \cup V^-$, $v$ is simple, and $v \neq 1, B$ do
6:    if $v \in V^+$ then
7:       Add the interval $[1, v]$
8:    else
9:       Add the interval $[v, B]$
10:   end if
11: end for /* Now $v^+ < v^-$ for every $(v^+, v^-) \in M$, where $v^+ \in V^+$, $v^- \in V^-$ and
    all the bricks are non-simple */
12: for all $v^- \in V^-$, $v^- < B$ do
13:    Add the interval $[v^-, B]$
14: end for /* Now all the bricks are non-simple, and $v^- = B, \forall v^- \in V^-$ */
15: while $V^+ \neq \emptyset$ do
16:    $v^+ \leftarrow \max V^+$
17:    for all $p > v^+$, $p < B$ do
18:       Fuse any pair of intervals complementing at $p$.
19:    end for
20:    C-delete an interval $[v^+, B]$
21: end while
```

Theorem 1 *Algorithm 1 is a polynomial-time 3-approximation algorithm for RKS.*

Note that the same result applies to multi-chromosomal karyotypes, by summing the bounds for the RKS problem on each chromosome. Note also that the results above imply also that $d \in [w/2 + f - 2N + s + m, 3w/2 + f - 2N + s + m]$

We now present Procedure 2, a heuristic that improves the performance of Algorithm 1, by replacing Steps 12-21. The procedure assumes that *(i)* all bricks are non-simple, and *(ii)* $v^+ < v^-$, for every $(v^+, v^-) \in M$, $v^- \in V^-$. In this case, $m = 0$, and the lower bound is reached only if no additions are made. Thus, Procedure 2 attempts to minimize the number of extra addition operations performed. For an interval I, let $L(I)$ and $R(I)$ be the left and right endpoints of I respectively.

4 Experimental results

In this section we present the results of sorting real cancer karyotypes, using Algorithm 1, combined with the improvement heuristic in Procedure 2.

4.1 Data preprocessing

For our analysis, we used the Mitelman database (version of February 27, 2008), which contained 56,493 cancer karyotypes, collected from 9,088 published stud-

Procedure 2 Heuristic for eliminating non-simple bricks

```
while $V^+ \neq \emptyset$ do
    $v^+ \leftarrow \max V^+$
    for all $p > v^+, p < B, p \notin V^-$ do
        Fuse any pair of intervals complementing at $p$.
    end for
    if $\exists I_1, I_2$, where $I_1 = I_2$ and $L(I_1) = v^+$, and $R(I_1) \in V^-$ then
        Let $I_1, I_2$ be a pair of intervals with minimal length satisfying the above.
        C-delete $I_1$
    else if $\exists I_1, I_2$, where $L(I_1) = L(I_2) = v^+$ and $R(I_1) < R(I_2) \in V^-$ then
        Let $I_1, I_2$ be a pair of intervals with minimal length satisfying the above.
        Add the interval $[R(I_1), R(I_2)]$
    else
        Let $u^- = \min\{v^- \in V^- | v^- > v^+\}$
        Add the interval $[u^-, B]$
    end if
end while
```

ies. The karyotypes in the Mitelman database (henceforth, MD) are represented in the ISCN format and can be automatically parsed and analyzed by the software package CyDAS [5]. We refer to a karyotype as *valid* if it is parsed by CyDAS without any error. According to our processing, 49,622 (88%) of the records were valid karyotypes. Since some of the records contain multiple distinct karyotypes found in the same tissue, the total number of simple (valid) karyotypes that we deduced from MD is 61,137.

A karyotype may contain uncertainties, or missing data, both represented by a '?' symbol. We ignored uncertainties and deleted any chromosomal fragments that were not well defined.

4.2 Sorting the karyotypes

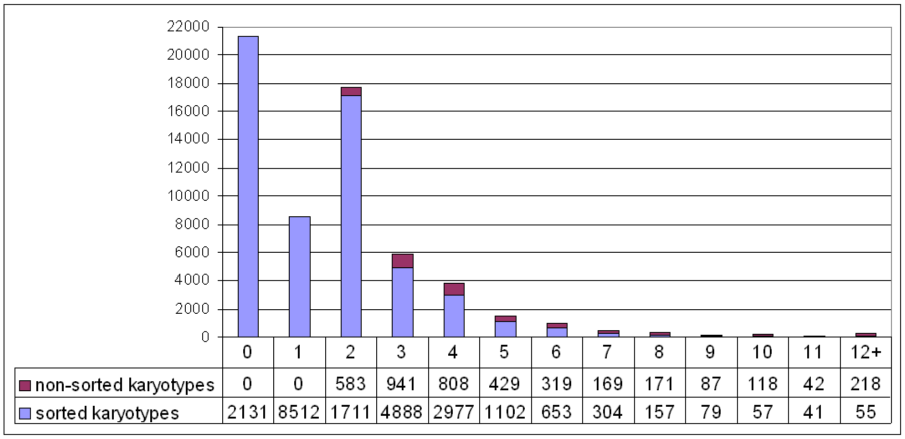
Out of the 61,137 karyotypes analyzed, only 3,885 karyotypes (6%) contained recurrent breakpoints. Our analysis focused on the remaining 57,252 karyotypes. We note that 37% (21,315) of these karyotypes do not contain any breakpoint at all. (In these karyotypes, no bands that are not adjacent in normal chromosomes are fused, but some chromosome tails as well as full chromosomes may be missing or duplicated). Following our assumptions (see Section 1.2), we broke all the breakpoints in each karyotype. We then applied Algorithm 1, combined with Procedure 2, to the fragments of each of the chromosomes in these karyotypes. We used the ploidy of each karyotype, as the normal copy-number (N) of each chromosome. In 99.9% (57,223) of the analyzed karyotypes our algorithm achieved the lower-bound, and thus the produced sequences are optimal. Each of the remaining 29 karyotypes contained a chromosome for which the computed sequence was larger in 2 than the lower-bound. Manual inspection revealed that for each of these cases the elementary distance was indeed 2 above the lower bound. Hence the computed sequences were found to be optimal in 100% of the analyzed cases.

4.3 Operations statistics

We now present statistics on the (direct) elementary operations performed by our algorithm. The 57,252 analyzed karyotypes, contained 84,601 (unique) breakpoints in total. Hence the average number of fusions (eq. breakpoints) per karyotype is approximately 1.5. The distribution of the number of breakpoints per a karyotypes, including the non-sorted karyotypes (i.e karyotypes with recurrent breakpoints), is presented in Fig. 4. The most frequent number of breakpoints after zero is 2, which may point to the prevalence of reciprocal translocations in the analyzed cancer karyotypes. Table 1 summarizes the average number of operations per sorted karyotype.

Table 1. Average number of elementary operations per (sorted) cancer karyotype

breakage	fusion	deletion	duplication	all
2.4	1.5	2.6	1.1	7.6



	0	1	2	3	4	5	6	7	8	9	10	11	12+
■ non-sorted karyotypes	0	0	583	941	808	429	319	169	171	87	118	42	218
□ sorted karyotypes	2131	8512	1711	4888	2977	1102	653	304	157	79	57	41	55

Fig. 4. The distribution of number of breakpoints (i.e. fusions of non-adjacent bands) per karyotype. “Sorted karyotypes” correspond to karyotypes with *no* recurrent breakpoints. “Non-sorted karyotypes” correspond to karyotypes with recurrent breakpoints. About 35% of all the karyotypes do not contain any breakpoint.

5 Discussion

In this paper we propose a new mathematical model for the evolution of cancer karyotypes, using four simple operations. Our model is in some sense the result of an earlier work [10], where we showed that chromosome gain and loss are dominant events in cancer. The analysis in [10] relied on a purely heuristic algorithm that reconstructed events using a wide catalog of complex rearrangement events, such as inversions, tandem-duplication etc. Here we make our first attempt to reconstruct rearrangement events in cancer karyotypes in a more rigorous, yet simplified, manner.

The fact that we model and analyze bands and karyotypes may seem out of fashion. While modern techniques today allow *in principle* detection of chromosomal aberrations in cancer at an extremely high resolution, the clinical reality is that karyotyping is still commonly used for studying cancer genomes, and to date it is the only abundant data resource for cancer genomes structure. Moreover, our framework is not limited to banding resolution as the “bands” in our model may represent any DNA blocks.

Readers familiar with the wealth of computational works on evolutionary genome rearrangements (see [3] for a review), may wonder why we have not used traditional operations, such as inversions and translocations, as has been previously done, e.g., by Raphael et al. [12]. The reason is that while inversions and translocations are believed to dominate the evolution of species, they form less than 25% of the rearrangement events in cancer karyotypes [10], and 15% in malignant solid tumors in particular. The extant models for genome rearrangements do not cope with duplications and losses, which are frequently observed in cancer karyotypes, and thus are not suitable for cancer genomes evolution. Extending these models to allow duplications results, even for the simplest models, in computationally difficult problems (e.g. [11, Theorem 10]). On the other hand, the elementary operations in our model can easily explain the variety of chromosomal aberrations viewed in cancer (including inversions and translocations). Moreover, each elementary operation we consider is strongly supported by a known biological mechanism [2]: breakage corresponds to a double-strand-break (DSB); fusion can be viewed as a non-homologous end-joining DSB-repair; whole chromosome duplications and deletions are caused by uneven segregation of chromosomes.

Based on our new model for chromosomal aberrations, we defined a new genome sorting problem. To further simplify this problem, we made two assumptions, which are supported by the vast majority of reported cancer karyotypes. We presented a lower bound for this simplified problem, followed by a polynomial 3-approximation algorithm. The application of this algorithm to 57,252 real cancer karyotypes yielded solutions that achieve the lower bound (and hence an optimal solution) in almost all cases (99.9%). This is probably due to the relative simplicity of reported karyotypes, especially after removing ones with repeated breakpoints (cf. Fig. 4).

In the future, we would like to extend this preliminary work by weakening our assumptions in a way that will allow the analysis of the remaining non-analyzed karyotypes (6% of the data), which due to their complexity, are likely to correspond to more advanced stages of cancer. Our hope is that this study will lead to further algorithmic research on chromosomal aberrations, and thus help in gaining more insight on the ways in which cancer evolves.

Acknowledgements

We thank the referees for their careful and critical comments. This study was supported in part by the Israeli Science Foundation (grant 385/06).

References

1. NCI and NCBI's SKY/M-FISH and CGH Database, 2001. http://www.ncbi.nlm.nih.gov/sky/skyweb.cgi.
2. D.G. Albertson, C. Collins, F. McCormick, and J. W. Gray. Chromosome aberrations in solid tumors. *Nature Genetics*, 34:369–376, 2003.
3. G. Bourque and L.Zhang. Models and methods in comparative genomics. *Advances in Computers*, 68:60–105, 2006.
4. D.O. Ferguson and W.A. Frederick. DNA double strand break repair and chromosomal translocation: Lessons from animal models. *Oncogene*, 20(40):5572–5579, 2001.
5. B. Hiller, J. Bradtke, H. Balz, and H. Rieder. CyDAS: a cytogenetic data analysis system. *BioInformatics*, 21(7):1282–1283, 2005. http://www.cydas.org.
6. M. Höglund, A. Frigyesi, T. Säll, D. Gisselsson, and F. Mitelman. Statistical behavior of complex cancer karyotypes. *Genes, Chromosomes and Cancer*, 42(4):327–341, 2005.
7. B. Korte and J. Vygen. *Combinatorial optimization : theory and algorithms*. Springer, Berlin, 2002.
8. F. Mitelman, editor. *ISCN (1995): An International System for Human Cytogenetic Nomenclature*. S. Karger, Basel, 1995.
9. F. Mitelman, B. Johansson, and F. Mertens (Eds.). Mitelman Database of Chromosome Aberrations in Cancer, 2008. http://cgap.nci.nih.gov/Chromosomes/Mitelman.
10. M. Ozery-Flato and R. Shamir. On the frequency of genome rearrangement events in cancer karyotypes, 2007. Presented in the first annual RECOMB satellite workshop on computational cancer biology.
11. A. J. Radcliffe, A. D. Scott, and E. L. Wilmer. Reversals and transpositions over finite alphabets. *SIAM J. Discret. Math.*, 19(1):224–244, 2005.
12. B.J. Raphael, S. Volik, C. Collins, and P. Pevzner. Reconstructing tumor genome architectures. *Bioinformatics*, 27:162–171, 2003.