On Maximal Instances for the Original Syntenic Distance

Cedric Chauve, Guillaume Fertin

▶ To cite this version:

Cedric Chauve, Guillaume Fertin. On Maximal Instances for the Original Syntenic Distance. Theoretical Computer Science, 2004, 326, pp.29-43. hal-00417741

HAL Id: hal-00417741
https://hal.science/hal-00417741

Submitted on 16 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INSTITUT DE RECHERCHE EN INFORMATIQUE DE NANTES

On maximal instances for the original syntenic distance

Cedric Chauve [1] and Guillaume Fertin [2]

[1] LaCIM et Département d'Informatique,
Université du Québec à Montréal
Case Postale 8888, Succursale Centre-Ville
H3C 3P8, Montréal (QC), Canada

[2] IRIN UPRES-EA 2157, Université de Nantes
2 rue de la Houssinière
BP 92208 - F44322 Nantes Cedex 3, France

— Combinatoire et Bio-Informatique —

RESEARCH REPORT

N^{o} 03.03

March 2003

IRIN, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3

Cedric Chauve and Guillaume Fertin

On maximal instances for the original syntenic distance

22 p.

Les rapports de recherche de l'Institut de Recherche en Informatique de Nantes sont disponibles aux formats PostScript® et PDF® à l'URL :

http://www.sciences.univ-nantes.fr/irin/Vie/RR/

Research reports from the Institut de Recherche en Informatique de Nantes are available in PostScript® and PDF® formats at the URL:

http://www.sciences.univ-nantes.fr/irin/Vie/RR/indexGB.html

© March 2003 by Cedric Chauve and Guillaume Fertin

On maximal instances for the original syntenic distance

Cedric Chauve and Guillaume Fertin

chauve@lacim.uqam.ca, fertin@irin.univ-nantes.fr

Abstract

The syntenic distance between two multichromosomal genomes has been introduced by Ferretti, Nadeau and Sankoff as an approximation of the evolutionary distance between genomes for which the gene order is not known. This distance is the minimum number of fusions, fissions and translocations required to transform a genome into the other. Kleinberg and Liben-Nowell proved that for n-chromosomes genomes the diameter for this distance is $2n-4$ and asked for a characterization of maximal instances for the syntenic distance (pairs of n-chromosomes genomes at a distance of $2n-4$). Pisanti and Sagot generalized the result of Liben-Nowell and Kleinberg and showed that the maximal distance between a genome with m chromosomes and a genome with n chromosomes is $n+m-4$. In this paper, we give a characterization of pairs of genomes with respectively n and m chromosomes that are at maximal distance.

1 Introduction

In the past few years, the interest in determining evolutionary distance between species, in the framework of genome rearrangement, has been continuously growing. In this context, the evolution models are at a genomic level, where mutations take place between large pieces of DNA, thus affecting the order of one or several genes within chromosomes. If one considers two genomes $\mathcal{G}_1$ and $\mathcal{G}_2$, the distance between $\mathcal{G}_1$ and $\mathcal{G}_2$ is usually defined as the minimum number of mutations needed to transform one genome into the other. Depending on the mutations that are considered (and thus allowed in the model), this gives rise to several distinct problems. One can cite for instance the *reversal distance* [1, 3, 6], where mutations are described in terms of reversals of chromosomes segments, or the *transposition distance* [2], where a transposition is a mutation that takes a segment out of a chromosome and inserts it at another location in the genome. Several other models or variants also exist, that either take or do not take into account the order of the genes within chromosomes.

In this paper, we are interested in the *syntenic distance*, where the order of genes within chromosomes is not considered ; hence, every chromosome is seen as an unordered set of genes. This notion has been introduced by Ferretti, Nadeau and Sankoff [5]. In this model, the mutations that are allowed are threefold: (i) *fusion*: two chromosomes are joined to form one, (ii) *fission*: a chromosome is splitted into two, and (iii) *translocation*: two chromosomes exchange arbitrary subsets of their genes. Among others, it has been shown that computing the syntenic distance between two arbitrary genomes is NP-hard [4] ; moreover, an approximation algorithm with approximation ratio 2 is known [4], and some other variants of such an approximation algorithm have been given [5, 10]. We also mention that some specific subclasses of instances have been considered, such as *linear synteny*, *exact synteny* or *nested synteny* [4, 9, 12]. In some of these cases, the computation of the distance becomes polynomial.

However, the algorithmic issue of approximating precisely the original syntenic distance between two genomes still asks for a better solution, and thus for a better understanding of the combinatorial nature of this problem. Among the natural combinatorial notions related to distances over a set of objects is the *diameter*, that is the maximum distance between two of these objects (here two genomes over the same set of genes). For the syntenic distance, the diameter for pairs of n-chromosomes genomes has been shown to be equal to $2n - 4$ [8, 12], and one particular instance reaching this value has been given. As a natural extension of these results, Kleinberg and Liben-Nowell asked for a characterization of those instances that are maximal (that is, pairs of n-chromosomes genomes at a syntenic distance of $2n - 4$) [8]. The computation of the syntenic diameter was later generalized by Pisanti and Sagot [12] who proved that the maximal distance between an n-chromosomes genome and an m-chromosomes genome, called the *bidimensional syntenic diameter*, is $n + m - 4$. The main result of the present article is an answer to the question of Kleinberg and Liben-Nowell about maximal instances, that we generalize to the bidimensional case. Our characterization will moreover allow us to decide in polynomial time whether a pair of genomes is at maximal distance.

In Section 2, we formally state the problem and its model, we recall some known properties and we introduce some notations and definitions. In Section 3, we give necessary conditions for square instances (pairs of n-chromosomes genomes) to be maximal, while in Section 4 we prove that these conditions are sufficient. Finally, in Section 5, we fully characterize those instances that are maximal, and we extend this characterization to bidimensional instances.

2 Preliminaries

Syntenic distance. Following [5], we define a *genome* $\mathcal{G}$ over a given set of *genes* as a *partition* of this set of genes into an unordered collection of *chromosomes* (sometimes called *synteny sets*). Hence the order among chromosomes and the order of genes on a chromosome are not taken into account, but a given gene can not appear into several chromosomes. In the syntenic distance model, the mutations considered are the *fusions* of two chromosomes (they are joined to form one chromosome), the *fissions* of a chromosome (it is splitted into two chromosomes) and the *translocations* between two chromosomes (they exchange arbitrary subsets of their genes). These mutations never involve, either as an input or as a result, empty chromosomes and do not cause the duplication of a gene.

For example let $\{a, b, p, q, r, x, y\}$ be a set of genes, and $\mathcal{G}_1 = \{\{a, b, c\}, \{p, q, r\}, \{x, y\}\}$ be a genome with 3 chromosomes. The genomes $\mathcal{G}_2 = \{\{a, b\}, \{c\}, \{p, q, r\}, \{x, y\}\}$, $\mathcal{G}_3 = \{\{a, b, c, x, y\}, \{p, q, r\}\}$ and $\mathcal{G}_4 = \{\{a, p\}, \{b, c, q, r\}, \{x, y\}\}$ result respectively from a fission of the first chromosome of $\mathcal{G}_1$, a fusion of the first and third chromosomes of $\mathcal{G}_1$, and a translocation between the first and second chromosomes of $\mathcal{G}_1$.

Given two genomes $\mathcal{G}_1$ and $\mathcal{G}_2$ over the same set of genes (there is no gene that appears in only one of the two genomes), the *syntenic distance* is the minimum number of mutations (fusions, fissions and translocations) needed to transform $\mathcal{G}_1$ into $\mathcal{G}_2$. This distance will be denoted $\mathcal{D}(\mathcal{G}_1, \mathcal{G}_2)$.

Compact representation of an instance. Let $[n] = \{1, 2 \dots n\}$ and let an instance of the syntenic distance be specified by two genomes $\mathcal{G}_1 = \{\mathcal{T}_1, \dots, \mathcal{T}_m\}$ (where $\mathcal{T}_i$ is the set of genes of the i^{th} chromosome) and $\mathcal{G}_2 = \{\mathcal{U}_1, \dots, \mathcal{U}_n\}$ on the same set of genes. The *compact representation* of this instance is an unordered collection of m subsets $\{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ of $[n]$ obtained by replacing in the sets $\mathcal{T}_1, \dots, \mathcal{T}_m$ every gene g by the indices of the chromosomes of $\mathcal{G}_2$ containing g. We should immediately notice that in the compact representation, every element of the underlying set $[n]$ can appear in several of the sets $\mathcal{S}_1, \dots, \mathcal{S}_m$.

For example, let $\{a, b, c, p, q, r, x, y\}$ be a set of genes, and $\mathcal{G}_1 = \{\{p, q, x\}, \{a, b\}, \{c, r\}, \{y\}\}$ and $\mathcal{G}_2 = \{\{a, b, c\}, \{p, q, r\}, \{x, y\}\}$ be two genomes. The compact representation of this instance is $\{\{2, 3\}$, $\{1\}$, $\{1, 2\}$, $\{3\}\}$.

Remark 1 *Given a compact representation, one can clearly perform on it operations like fusions, fissions and translocations, provided that one gives a rule to deal with the presence of multiple copies of an element of $[n]$ in a set. The rule used here, and in all the papers about the syntenic distance, follows naturally from the construction of a compact instance: when a fusion or a translocation could induce two copies of an element into a set, these two copies are gathered into a single copy. For example the fusion of $\{1, 2\}$ and $\{2, 3\}$ gives $\{1, 2, 3\}$.*

For two compact representations $\mathcal{S}$ and $\mathcal{T}$, we denote by $\mathcal{D}^*(\mathcal{S}, \mathcal{T})$ the minimum number of fusions, fissions and translocations (as defined in Remark 1) to transform $\mathcal{S}$ into $\mathcal{T}$. Now let $\mathcal{G}_1$ and $\mathcal{G}_2$ be an instance of the syntenic distance, $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ be its compact representation and $\overline{\mathcal{G}}_n = \{\{1\}, \dots, \{n\}\}$. It has been proved that $\mathcal{D}(\mathcal{G}_1, \mathcal{G}_2) = \mathcal{D}^*(\{\mathcal{S}_1, \dots, \mathcal{S}_m\}, \overline{\mathcal{G}}_n)$ [4, 5]. This reduction to a compact representation allows to define an instance for genomes on a set of n chromosomes as a collection $\{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ of m subsets of $[n]$ (where m is the number of chromosomes of the genome $\mathcal{G}_1$ and n the number of chromosomes of the genome $\mathcal{G}_2$).

Remark 2 *It should be noticed that in several papers about the syntenic distance between genomes, authors consider a genome as a* collection *of subsets of the set of genes, and not a* partition *of the set of genes. Following this definition, they allow the repetition of genes on a same genome, and in this framework the equivalence with compact instances does not hold anymore. Indeed if we allow a gene to appear in several chromosomes of a n-chromosomes genome $\mathcal{G}$, the compact representation of this genome is not equal to $\{\{1\}, \{2\}, \dots, \{n\}\}$ and the distance from $\mathcal{G}$ to itself, computed through this compact instance, is strictly greater than 0, which is clearly false. An extreme example would be to consider a genome $\mathcal{G}$ where every chromosome contains a copy of every gene and the compact representation $\mathcal{G}'$ of the instance $(\mathcal{G}, \mathcal{G})$. Hence $\mathcal{D}(\mathcal{G}, \mathcal{G}) = 0$, when $\mathcal{D}^*(\mathcal{G}', \overline{\mathcal{G}}_n) = 2n - 4$. The only published proof of the equivalence between general instances and compact instances in [4] does not make this point clear. However, it is not difficult to see that minor modifications make this proof hold in the case where no gene is repeated.*

From now on, we consider only the compact representation of instances of synteny, that is collections[1] $\{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ of subsets of $[n]$ and the fusion, fissions and translocations for compact instances as defined in Remark 1. For such an instance, we call an *optimal mutation sequence* any minimal (in terms of number of mutations) sequence of mutations (fusions, fissions and translocations) that transform $\{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ into $\overline{\mathcal{G}}_n$. One says that such a sequence *solves* the instance $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ and one denotes by $\mathcal{D}(\mathcal{S})$

[1] It should always be kept in mind that a genome is an unordered collection of chromosomes, which implies that two collections of subsets of $[n]$ that differ only by a permutation of the subsets they contain represent the same instance.

the length of such an optimal sequence. An instance is said to be an n*-square instance* if $n = m$. If $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ is the compact representation of an instance $(\mathcal{G}_1, \mathcal{G}_2)$, we denote by $\mathrm{Dual}(\mathcal{S})$ the compact representation of the instance $(\mathcal{G}_2, \mathcal{G}_1)$.

Structural properties of the syntenic distance. We now recall some properties that will be used in the proofs of the characterization of maximal instances for the syntenic distance. We refer the reader to [4, 10] for the proofs of these properties.

Proposition 1 (Duality). *For every instance of synteny $\mathcal{S}$, $\mathcal{D}(\mathcal{S}) = \mathcal{D}(\mathrm{Dual}(\mathcal{S}))$.*

Proposition 2 (Canonicity). *For every instance of synteny, if there is an optimal mutation sequence solving it with k_1 fusions, k_2 translocations and k_3 fissions, then there is also an optimal mutation sequence solving it and* starting *by k_1 fusions,* followed *by k_2 translocations and* finishing *by k_3 fissions.*

Let $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ and $\mathcal{S}' = \{\mathcal{S}'_1, \ldots, \mathcal{S}'_m\}$ be two instances of synteny. One says that $\mathcal{S}$ *dominates* $\mathcal{S}'$ if for all $i \in [m]$ we have $\mathcal{S}'_i \subseteq \mathcal{S}_i$.

Proposition 3 (Monotonicity). *Let $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ and $\mathcal{S}' = \{\mathcal{S}'_1, \ldots, \mathcal{S}'_m\}$ be two instances of synteny. If $\mathcal{S}$ dominates $\mathcal{S}'$, then $\mathcal{D}(\mathcal{S}) \geq \mathcal{D}(\mathcal{S}')$.*

The syntenic diameter. The diameter of the syntenic distance was studied by Kleinberg and Liben-Nowell [8] and Pisanti and Sagot [12]. Let us denote by $\mathcal{SD}(m, n)$ the maximal syntenic distance over all the instances composed of m subsets of $[n]$. In [8], the diameter for n-square instances ($m = n$) was computed and shown to be equal to $2n - 4$. The following result, proved in [12], generalizes the result of Kleinberg and Liben-Nowell to the bidimensional case.

Theorem 1 (Bidimensional syntenic diameter). *For all $n, m \geq 4$, $\mathcal{SD}(n, m) = n + m - 4$.*

From now on, an instance $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ of the syntenic distance on $[n]$ will be called a (m, n)-*maximal instance* if it can not be solved in less than $n + m - 4$ mutations. If $n = m$ maximal instances will be called *maximal n-square instances*.

Gossiping and synteny. The main tool that we will use to characterize maximal instances is a relationship between the transformation of a genome into another by translocations and a problem of gossiping, introduced by Kleinberg and Liben-Nowell in [8] and developed by Liben-Nowell in [11].

More precisely, the *gossiping* problem is used to model information dissemination in communication networks. It has been introduced in the 50s, and has received considerable attention since, leading to a wide literature. For more information on the subject, we refer to the survey [7]. The gossiping problem is defined as follows: we start with a set of n people, denoted by integers $1, \ldots, n$, each knowing a single piece of information, denoted by $\{i\}$ for the person labeled by i. Those people communicate between them, using the *telephone model*, that is a communication takes place between two people only, and both exchange all the information they have. One of the first questions that arose from this model is to determine the minimum number of calls to be made in order that everyone knows everything, that is $[n]$.

Kleinberg and Liben-Nowell pointed out an equivalence between translocations and a variant of gossiping, that they called *incomplete* gossiping. In this problem, people do not necessarily want to know all the pieces of information, but each person wants to know a specific subset of the total information (that is a subset of $[n]$). Hence, in particular, during a call a person is allowed to give to the other caller only an arbitrary subset of his/her information. Let us denote by $\mathcal{S}_i$ the information that the person i wants to know. In this model, for a given information configuration $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ (a totally ordered set of n subsets of $[n]$), one says that it can be *disseminated in k calls* if there is a sequence of k calls that leads to the situation where, for $i = 1, \ldots, n$, the person i knows at least the information $\mathcal{S}_i$ (if every person i knows exactly $\mathcal{S}_i$, one says that $\mathcal{S}$ is *exactly disseminated*).

The equivalence with translocations in the syntenic distance problem is immediate: if I_j and I_k are the sets of information that two people j and k know before calling each other, after this call they respectively know I'_j and I'_k where $I_j \cup I_k = I'_j \cup I'_k$, which leads to the remark that one could evolve from sets I'_j and I'_k to the sets I_j and I_k by a translocation. Hence, a call between two people corresponds to the reverse of a translocation, which implies immediately the following property (see [11, Section 4]).

Proposition 4 *Let $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ be a collection of subsets of $[n]$. If this configuration can be disseminated in k calls, then $\mathcal{S}$ can be solved by k translocations and $\mathcal{D}(\mathcal{S}) \leq k$.*

When during a call two people, say i and j, exchange totally their available information, we say that this call is *complete* and we denote it by (i, j).

Reduction of an instance. Finally, before starting the description of our characterization of maximal instances, let us introduce the the notion of reduction of an instance, that will be used intensively in our proofs. Let X be a subset of $[n]$ ($X = \{x_1, \ldots, x_k\}$, with $x_1 < \ldots < x_k$) and $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ be an instance of synteny. We call the *X-reduction of $\mathcal{S}$* the instance $\mathcal{S}^X = \{\mathcal{S}^X_1, \ldots, \mathcal{S}^X_m\}$ obtained from $\mathcal{S}$ as follows: in each $\mathcal{S}_j$, with $1 \leq j \leq n$, every element $x_i \in X$ is replaced by a single occurrence of x_1 and the resulting collection $\mathcal{S}' = \{\mathcal{S}'_1, \ldots, \mathcal{S}'_m\}$ of m subsets of $[n]$ are normalized on $[n-k+1]$ to give $\mathcal{S}^X$. Hence $\mathcal{S}^X$ is an instance on $[n-k+1]$. For example, let $\mathcal{S} = \{\{1,2,3,4,5\}, \{1,2,3,4\}, \{1,2,5\}, \{1,2\}, \{1\}\}$ and $X = \{1,2,4\}$. Then $\mathcal{S}' = \{\{1,3,5\}, \{1,3\}, \{1,5\}, \{1\}, \{1\}\}$ and $\mathcal{S}^X = \{\{1,2,3\}, \{1,2\}, \{1,3\}, \{1\}, \{1\}\}$ is an instance with $n = 3$.

3 Necessary conditions for maximal square instances

We recall that an n-square (compact) instance is an instance composed of n subsets of $[n]$. In this section, we describe some conditions that an n-square instance should satisfy in order to be a maximal n-square instance. These conditions formalize the quite intuitive idea that a maximal instance does not involve a small subset of $[n]$.

Lemma 1 *Let $n \geq 5$ and $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ be an n-square instance. If there exists $i \in [n]$ such that $|\mathcal{S}_i| \leq n-3$ then $\mathcal{S}$ is not a maximal n-square instance.*

Proof: We consider the gossiping problem for $\mathcal{S}$ and we show that $\mathcal{S}$ can be disseminated in $2n-5$ complete calls, thus proving by Proposition 4 that $\mathcal{D}(\mathcal{S}) \leq 2n-5$. We first note that $\mathcal{S}$ is dominated by the instance $\mathcal{S}' = \{\mathcal{S}'_1, \ldots, \mathcal{S}'_n\}$, defined by $\mathcal{S}'_i = [n]$ for all $i \in [n]/\{1\}$, and $\mathcal{S}'_1 = [n]/\{n-2, n-1, n\}$. It is possible to disseminate $\mathcal{S}'$ in a sequence of $2n-5$ complete calls: first, for i from 1 to $n-4$, $(i, n-3)$ (i calls $n-3$) ; then $(n-2, n-3)$, $(n-1, n)$, $(n-2, n-1)$ $(n-3, n)$ (after this n calls have been made) ; and finally, for j from 2 to $n-4$, (j, n). At the end of the process, every person i, for $i \in [n]/\{1\}$, knows $[n]$, while the person 1 knows $[n-3]$. Since this process uses $2n-5$ calls, we conclude by Propositions 3 and 4 that $\mathcal{D}(\mathcal{S}) \leq \mathcal{D}(\mathcal{S}') \leq 2n-5$, and thus $\mathcal{S}$ is not maximal. □

Hence, we can now restrict our study to instances where every set $\mathcal{S}_i$ is of size at least $n-2$. The next three lemmas give necessary conditions on these instances that will be proved to be sufficient in the following section.

Lemma 2 *Let $n \geq 4$ and $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ be an n-square instance. If there exists $i, j \in [n]$ and $x, y \in [n]$ ($i \neq j$ and $x \neq y$) such that $x \notin \mathcal{S}_i$, $y \notin \mathcal{S}_i$ and $x \notin \mathcal{S}_j$, then $\mathcal{S}$ is not a maximal n-square instance.*

Proof: Here again we consider the gossiping problem associated to this instance, and show that $\mathcal{S}$ can be disseminated in $2n-5$ calls. $\mathcal{S}$ is dominated by the square instance $\mathcal{S}'$ such that $\mathcal{S}'_i = [n]$ for all $i \in [n]/\{n-3, n-2\}$, $\mathcal{S}'_{n-3} = [n-2]$ and $\mathcal{S}'_{n-2} = [n-1]$. $\mathcal{S}'$ can be disseminated by the following sequence of $2n-5$ complete calls: for i from 1 to $n-1$, $(i, i+1)$ (after this, n and $n-1$ know $[n]$, $n-2$ knows $[n-1]$ and $n-3$ knows $[n-2]$) ; then for j from 1 to $n-4$, (j, n). Thus, $\mathcal{D}(\mathcal{S}) \leq \mathcal{D}(\mathcal{S}') \leq 2n-5$ by Propositions 3 and 4, and $\mathcal{S}$ is not a maximal n-square instance. □

Lemma 3 *Let $n \geq 5$ and $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ be an n-square instance. If there exists three distinct integers $i, j, k \in [n]$ and $x \in [n]$ such that $|\mathcal{S}_i| \leq n-2$, $x \notin \mathcal{S}_j$ and $x \notin \mathcal{S}_k$, then $\mathcal{S}$ is not a maximal n-square instance.*

Proof: Due to Lemmas 1 and 2 above, we just have to focus on instances dominated by the instance $\mathcal{S}'(n)$ containing $n-3$ copies of $[n]$, one copy of $[n-2]$ and one copy of $[n]/\{n-2\}$. We now distinguish two cases: $n = 5$ and $n \geq 6$.

When $n = 5$, it is possible to disseminate $\mathcal{S}'(5)$ with the following five complete calls: $(4,5)$, $(1,2)$, $(2,5)$, $(1,3)$, $(3,4)$. Thus, in 5 translocations, it is possible to solve $\mathcal{S}'(5)$ that is not a maximal 5-square instance.

Now suppose $n \geq 6$. In that case, we first perform on $\mathcal{S}'(n)$ $n-5$ fusions over $n-4$ copies of $[n]$. Let us denote by $\mathcal{S}''$ the resulting instance, that is $\mathcal{S}'' = \{[n], [n], [n-2], [n]/\{n-2\}, [n]/\{n-2\}\}$. As $\mathcal{S}'(5)$ is nothing else that the X-reduction of $\mathcal{S}''$ for $X = [n-4]$, we can deduce from the sequence of translocations described above to solve $\mathcal{S}'(5)$ a sequence of 5 translocations that can be applied to $\mathcal{S}''$. Hence, after these n mutations ($n-5$ fusions and 5 translocations) one has an instance $\{[n-4], \{n-3\}, \{n-2\}, \{n-1\}, \{n\}\}$, that can be transformed into $\overline{\mathcal{G}}_n$ by $n-5$ fissions on the subset $\{1, 2 \ldots n-4\}$ producing the $n-4$ subsets $\{1\}, \{2\} \ldots \{n-4\}$. Altogether, we used $2n-5$ mutations, and thus $\mathcal{S}'$ is not maximal, which implies, by monotonicity (Proposition 3) that $\mathcal{S}$ is not a maximal n-square instance. $\square$

Lemma 4 *Let $n \geq 6$ and $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ be an n-square instance. If there exists $i, j \in [n]$ ($i \neq j$) such that $|\mathcal{S}_i| \leq n-2$ and $|\mathcal{S}_j| \leq n-2$, then $\mathcal{S}$ is not a maximal n-square instance.*

Proof: Thanks to Lemma 2, we know that if there exists an x such that $x \notin \mathcal{S}_i$ and $x \notin \mathcal{S}_j$, then $\mathcal{S}$ is not maximal. Now let us focus on the remaining cases, that corresponds to instances dominated by $\mathcal{S}'(n)$, where $\mathcal{S}'(n)$ is composed of $n-2$ copies of $[n]$, one copy of $[n-2]$ and two copies of $[n]/\{1,2\}$. Following the same method as in proof of Lemma 3, we can notice that: the sequence of 7 complete calls $(5,6), (3,4), (1,2), (4,5), (2,3), (2,6), (1,4)$ disseminates $\mathcal{S}'(6)$; if we first perform on $\mathcal{S}'(n)$ $n-6$ fusions over $n-5$ copies of $[n]$, followed by the 6 translocations induced by the calls described above to disseminate $\mathcal{S}'(6)$ (they lead to the instance $\{[n-5], \{n-4\}, \{n-3\}, \{n-2\}, \{n-1\}, \{n\}\}$) and $n-6$ fissions to transform $[n-5]$ into $\{1\}, \{2\} \ldots \{n-5\}$, we can solve $\mathcal{S}'(n)$ with $(n-6)+7+(n-6) = 2n-5$ mutations. Hence, by monotonicity, $\mathcal{S}$ is not a maximal n-square instance. $\square$

Altogether, the four previous lemmas (Lemmas 1 to 4) lead to the following proposition.

Proposition 5 *Let $n \geq 6$, $\mathcal{A}(n) = \{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$ be the n-square instance defined by $\mathcal{A}_i = [n]/\{i\}$ for $i = 1, \ldots, n$, and $\mathcal{B}(n) = \{\mathcal{B}_1, \ldots, \mathcal{B}_n\}$ be the n-square instance defined by $\mathcal{B}_i = [n]/\{i\}$ for $i = 1, \ldots, n-2$, $\mathcal{B}_{n-1} = [n]/\{n-1, n\}$ and $\mathcal{B}_n = [n]$. Then every maximal n-square instance dominates $\mathcal{A}(n)$ or $\mathcal{B}(n)$.*

4 Sufficient conditions for maximal square instances

In the previous section we proved that every maximal instance has an equivalent instance that dominates either $\mathcal{A}(n)$ or $\mathcal{B}(n)$ (or both). Here we prove that these two instances are maximal n-square instances (Proposition 6). We first restrict our study to the case of translocations (Lemmas 5 and 6 below). In Lemma 7 we show that this restriction is sufficient.

Lemma 5 *If $n \geq 4$, every sequence of translocations solving $\mathcal{A}(n)$ has length at least $2n-4$.*

Proof: First we consider the gossiping problem. Let $C = c_1, \ldots, c_\ell$ be a sequence of complete calls disseminating exactly an instance $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$ that dominates $\mathcal{A}(n)$ and is such that there is at least one subset $\mathcal{S}_i$ of size exactly $n-1$ (say $\mathcal{S}_i = [n]/\{j\}$ where j can be equal to i). We can consider that the calls of C are complete because replacing an incomplete call by the complete call involving the same two people does not prevent the dissemination of $\mathcal{S}$. For every $i \in [n]$, we define the *final call for* i as the least call c_ℓ involving the person i and such that this person knows $\mathcal{S}_i$ after the call c_ℓ.

The facts that $\mathcal{S}_i = [n]/\{j\}$ and that C disseminates exactly $\mathcal{S}$ imply that there is a subsequence $c_{i_1}, \ldots, c_{i_{n-2}}$ of C such that $c_{i_{n-2}}$ is the final call for i, for every $k \in [n-3]$ there is a person involved in both the calls c_{i_k} and $c_{i_{k+1}}$, and for $k \in [n-3]$ none of the two people involved in the call c_{i_k} knows the information j before the call. But by definition of $\mathcal{A}(n)$, exactly $n-1$ people should know the information j at the end of the calls C and such a dissemination of the information j needs at least $n-2$ calls such that one person involved in this call knows this information before the call. As none of these calls can belong to $c_{i_1}, \ldots, c_{i_{n-2}}$, we have $\ell \geq 2n-4$.

Altogether with Proposition 4, this implies that every instance $\mathcal{S}$ dominating $\mathcal{A}(n)$ and not composed of n copies of $[n]$, can not be solved by translocations only with less than $2n-4$ steps. Otherwise, if $\mathcal{S}$ contains n copies of $[n]$, by monotonicity (Proposition 3) and the value of the syntenic diameter (Theorem 1), one needs at least $2n-4$ mutations to solve $\mathcal{S}$, which concludes the proof. □

Lemma 6 *If $n \geq 4$, every sequence of translocations solving $\mathcal{B}(n)$ has length at least $2n-4$.*

Proof: Here again we consider a sequence of complete calls $C = c_1, \ldots, c_\ell$ that disseminates an instance $\mathcal{S}$ dominating $\mathcal{B}(n)$. It follows from the proof of Lemma 5 that if there exists i such that $|\mathcal{S}_i| = n-1$ (the person i knows exactly $n-1$ pieces of information after the calls of C), then $\ell \geq 2n-4$. Hence the only instance $\mathcal{S}$ dominating $\mathcal{B}(n)$ that could be solved in less than $2n-4$ translocations is such that $\mathcal{S}_n = [n]/\{a,b\}$ for some $a, b \in [n]$, and $\mathcal{S}_i = [n]$ for $i = 1, \ldots, n-1$.

We now consider such an instance $\mathcal{S}$ and we say that a call is an *$\{a,b\}$-call* if the information pieces a and b are known, after the call, by the two people involved in this call, while it was not the case before the call. Similarly to the proof of Lemma 5, we can notice that in order that n knows exactly $[n]/\{a,b\}$ after all the calls, C should contain a subsequence C' of at least $n-3$ calls involving people who do not know either a or b at the time of the call (these calls are all non $\{a,b\}$-calls). Moreover, as the $n-1$ people other than n should know the information pieces a and b after the calls of C (which can be done for people a and b in one complete call (a,b)), C should contain a subsequence C'' of at least $n-2$ $\{a,b\}$-calls (so $C' \cap C'' = \emptyset$). Hence, if the first call involving a or b is not (a,b) (that is do not belong either to C' or C''), $\ell \geq 2n-4$. Otherwise the first call involving people a and b is (a,b), which leads to $\ell \geq 2n-5$ and the fact that all the $\{a,b\}$-calls are such that exactly one of the people involved knows the information pieces a and b before the call. Now let us consider such a sequence of $2n-5$ calls disseminating $\mathcal{S}$. The $n-2$ $\{a,b\}$-calls of this sequence can be partitioned into three different kinds of calls:

C_1. calls that are final for none of the two involved people ;

C_2. calls such that one person knows $[n]$ before the call, but not the second person ;

C_3. calls such that none of the two people knows $[n]$ before the call, but both do after the call.

We now define a mapping from C_1 to C_3 as follows: every call of C_3, say (i,j), where i knows the information a and b before the call (but not j by definition of $\{a,b\}$-calls) is mapped to a call of C_1 involving the person i and such that i does not know the information pieces a and b before this call but does after the call. This mapping is well defined because if i knows a and b before a call, it learned it during another call that was not final for him, that is a call from C_1. It is an injective mapping due to the fact that if the call (i,j) belongs to C_3, it is the only one of C_3 involving one of these two people. Hence we have $|C_1| \geq |C_3|$, which implies that $|C_1| + |C_2| + |C_3| \geq |C_2| + 2|C_3|$. But, as exactly $n-1$ persons want to know the complete information $[n]$ at the end of the calls, we have that $n-1 = |C_2| + 2|C_3|$. This leads to a contradiction with the fact that the total number of $\{a,b\}$-calls is $n-2$, which, altogether with the equivalence between calls and translocations, concludes the proof. □

Lemma 7 *If $n \geq 3$, no optimal sequence that contains at least a fission or a fusion can solve $\mathcal{A}(n)$ or $\mathcal{B}(n)$ in less than $2n-4$ mutations.*

Proof: We reason by induction on n. First, we can verify that the property holds when $n = 3$: one can not solve $\mathcal{A}(3)$ or $\mathcal{B}(3)$ with only one fusion or one fission.

Now suppose that $n > 3$. It follows from the monotonicity property (Proposition 3) and from the fact that $\mathcal{A}(n)$ or $\mathcal{B}(n)$ are square instances, that, if there is an optimal sequence of mutations solving one of these two instances and containing at least one fusion, then there is an optimal sequence of mutations $\sigma = \sigma_1, \dots, \sigma_\ell$ solving the same instance and where σ_1 is a fusion and σ_ℓ is a fission.

Let us denote by $\mathcal{S}$ the starting instance (that is $\mathcal{A}(n)$ or $\mathcal{B}(n)$), by $\mathcal{T} = \{\mathcal{T}_1, \dots, \mathcal{T}_{n-1}\}$ the instance resulting from the the fusion σ_1 applied to $\mathcal{S}$, and by $\mathcal{U} = \{\mathcal{U}_1, \dots, \mathcal{U}_{n-1}\}$ the instance resulting from the action of the sequence of mutations $\sigma_2, \dots, \sigma_{\ell-1}$ applied to $\mathcal{T}$. The fact that the mutation σ_ℓ is a fission implies that the instance $\mathcal{U}$ is a set partition of $[n]$ into $n-2$ sets of size 1 and one set of size 2. Let $\mathcal{U}_i = \{a, b\}$ be the set of size 2 (with $a < b$).

Suppose that the instance $\mathcal{T}^{\{a,b\}}$ (an X-reduction of $\mathcal{T}$, where $X = \{a, b\}$) is an $(n-1)$-square instance that dominates $\mathcal{A}(n-1)$ or $\mathcal{B}(n-1)$. By definition of $\mathcal{T}$ and $\mathcal{U}$, the sequence of mutations $\sigma_2, \dots, \sigma_{\ell-1}$ induces a sequence of $\ell - 2$ mutations solving $\mathcal{T}^{\{a,b\}}$, and also $\mathcal{A}(n-1)$ or $\mathcal{B}(n-1)$ (by monotonicity, Proposition 3). Now, if σ_2 is a fusion, by induction we have $\ell - 2 = 2n - 6$. Otherwise, the mutations $\sigma_2, \dots, \sigma_{\ell-1}$ are translocations and Lemmas 5 and 6 imply that $\ell - 2 = 2n - 6$. Both cases lead to $\ell = 2n - 4$.

To conclude the proof, we then have to show that $\mathcal{T}^{\{a,b\}}$ is an $(n-1)$-square instance that dominates $\mathcal{A}(n-1)$ or $\mathcal{B}(n-1)$. It can be easily deduced from the fact that $\mathcal{T}$ is (up to a permutation of its subsets) one of the four instances $\mathcal{H}, \mathcal{I}, \mathcal{J}$ and $\mathcal{K}$ on $[n]$ defined respectively by:

- $\mathcal{H}_i = [n]/\{i\}$ for $i = 1, \dots, n-2$, and $\mathcal{H}_{n-1} = [n]$,
- $\mathcal{I}_i = [n]/\{i\}$ for $i = 1, \dots, n-3$, and $\mathcal{I}_{n-2} = \mathcal{I}_{n-1} = [n]$,
- $\mathcal{J}_i = [n]/\{i\}$ for $i = 1, \dots, n-3$, $\mathcal{J}_{n-2} = [n]/\{n-2, n-1\}$ and $\mathcal{J}_{n-1} = [n]$,
- $\mathcal{K}_i = [n]/\{i\}$ for $i = 1, \dots, n-4$, $\mathcal{K}_{n-3} = [n]/\{n-3, n-2\}$, and $\mathcal{K}_{n-2} = \mathcal{K}_{n-1} = [n]$,

and from a (tedious but easy) study of the different values that a and b can have, that $\mathcal{T}^{\{a,b\}}$ is an $(n-1)$-square instance that dominates at least one of the instances $\mathcal{A}(n-1)$ or $\mathcal{B}(n-1)$ (as before we consider such instances up to a permutation of their subsets). $\square$

Proposition 6 *If $n \geq 6$, then $\mathcal{A}(n)$ and $\mathcal{B}(n)$ are maximal n-square instances.*

Proof: This is an immediate consequence of Lemmas 5, 6 and 7. $\square$

5 Characterization of maximal instances

We can now state our main characterization results and their algorithmic consequences. Our first result is a characterization of maximal n-square instances for $n \geq 6$.

Theorem 2 *Let $n \geq 6$, $\mathcal{A}(n) = \{\mathcal{A}_1, \dots, \mathcal{A}_n\}$ be the n-square instance defined by $\mathcal{A}_i = [n]/\{i\}$ for $i = 1, \dots, n$, and $\mathcal{B}(n) = \{\mathcal{B}_1, \dots, \mathcal{B}_n\}$ be the n-square instance defined by $\mathcal{B}_i = [n]/\{i\}$ for $i = 1, \dots, n-2$, $\mathcal{B}_{n-1} = [n]/\{n-1, n\}$ and $\mathcal{B}_n = [n]$. An n-square instance is a maximal n-square instance if and only if it dominates at least one of the n-square instances $\mathcal{A}(n)$, $\mathcal{B}(n)$ or $\mathrm{Dual}(\mathcal{B}(n))$.*

Proof: This result follows immediately from Propositions 1, 5 and 6 and the fact that $\mathrm{Dual}(\mathcal{A}(n)) = \mathcal{A}(n)$. $\square$

We can now extend Theorem 2 to the case of general instances, that is we consider instances composed of m subsets of $[n]$, with $n \neq m$.

Remark 3 *We can limit our study to the case $n > m$, and the case $m > n$ follows by duality (Proposition 1).*

The next results gives a characterization of (m, n)-maximal instances in terms of reduced instances.

Lemma 8 *An instance of the syntenic distance $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ on $[n]$, where $n > m \geq 4$, is an (m, n)-maximal instance if and only if for every subset $\{a, b\}$ of $[n]$, $\mathcal{S}^{\{a,b\}}$ is an $(m, n-1)$-maximal instance.*

Proof: Suppose that $\mathcal{S}$ is not an (m,n)-maximal instance: there is a sequence of less than $n+m-4$ mutations solving it such that the last mutation is a fission that splits a subset $\{a,b\}$ of $[n]$ (by canonicity and the fact that $n>m$). Hence there is a subset $\{a,b\}$ of $[n]$ such that $\mathcal{S}^{\{a,b\}}$ is not an $(m,n-1)$-maximal instance. Now, suppose that there exists $\{a,b\}\subset[n]$ such that $\mathcal{S}^{\{a,b\}}$ is not an $(m,n-1)$-maximal instance. This instance can be solved by a sequence of less than $n+m-5$ mutations. If we complete this sequence by a fission of $\{a,b\}$ we solve $\mathcal{S}$ in less than $n+m-4$ mutations and $\mathcal{S}$ is not an (m,n)-maximal instance. □

We now generalize the notion of reduction of an instance for the syntenic distance. Let $\mathcal{S}$ be an instance composed of m subsets of $[n]$ $(n>m)$, $P=p_1,\ldots,p_m$ be a set partition of $[n]$ into m non-empty subsets (the $p_i's$) and $S^0,\ldots,S^m$ be the only sequence of instances defined by $S^0=\mathcal{S}$ and $S^{i+1}=(S^i)^{p_i}$ (the p_i-reduction of S^i) for $i=0,\ldots,m-1$. The P-reduction of $\mathcal{S}$, denoted by $\mathcal{S}^P$, is defined as the m-square instance $\mathcal{S}^P=S^m$.

Lemma 9 *An instance of the syntenic distance $\mathcal{S}=\{\mathcal{S}_1,\ldots,\mathcal{S}_m\}$ on $[n]$, where $n>m\geq 4$, is an (m,n)-maximal instance if and only if for every set partition $P=p_1,\ldots,p_m$ of $[n]$ into m non-empty subsets, $\mathcal{S}^P$ is an m-square maximal instance.*

Proof: First it is clear that if, for a set partition $P=p_1,\ldots,p_m$ of $[n]$, $\mathcal{S}^P$ can be solved with less than $2m-4$ mutations, then by adding to these mutations a sequence of $n-m$ fissions one can define a sequence of less than $n+m-4$ mutations solving $\mathcal{S}$.

So we just have to concentrate on the case where for every set partition P, $\mathcal{S}^P$ is an m-square maximal instance, and prove that in this case $\mathcal{S}$ is maximal. If $n-m=1$, it follows immediately from Lemma 8. Otherwise, suppose that $n-m>1$ and that $\mathcal{S}$ is not an (m,n)-maximal instance. From Lemma 8 one can say that there is $\{a,b\}\subset[n]$ (with $a<b$) such that $\mathcal{S}^{\{a,b\}}$ (denoted by $\mathcal{T}$) is not an $(m,n-1)$-maximal instance. Hence by induction, we know that there is a set partition Q of $[n-1]$ into m non-empty sets such that $\mathcal{T}^Q$ is not an m-square maximal instance. We now define a set partition P of $[n]$ as follows: every element x of Q greater than or equal to b is replaced by $x+1$ and b is added to the set containing a. It is immediate to verify that every sequence of k mutations solving $\mathcal{T}^Q$ induces a sequence of k mutations solving $\mathcal{S}^P$, which is not an m-square maximal instance. □

Thus, if we want a characterization of (m,n)-maximal instances that will be easier to translate in a decision algorithm, one just has to find which instances can not be reduced to a non m-square maximal instance. Such instances are described in the following result.

Theorem 3 *Let $n>m\geq 6$, $\mathcal{C}(m,n)=\{\mathcal{C}_1,\ldots,\mathcal{C}_m\}$ and $\mathcal{D}(m,n)=\{\mathcal{D}_1,\ldots,\mathcal{D}_m\}$ be the instances on $[n]$ defined by: $\mathcal{C}_i=[n]/\{i\}$ for $i=1,\ldots,m-2$ and $\mathcal{C}_{m-1}=\mathcal{C}_m=[n]/\{m-1\}$, and $\mathcal{D}_i=[n]/\{i\}$ for $i=1,\ldots,m-1$ and $\mathcal{D}_m=[n]/\{m,m+1\}$. An instance $\mathcal{S}=\{\mathcal{S}_1,\ldots,\mathcal{S}_m\}$ on $[n]$ is an (m,n)-maximal instance if and only if it dominates at least one of the two instances $\mathcal{C}(m,n)$ and $\mathcal{D}(m,n)$.*

Proof: Let us consider an instance $\mathcal{S}$ that does not dominate one of the two instances $\mathcal{C}(m,n)$ and $\mathcal{D}(m,n)$. It follows easily from the definition of these two instances that one subset $\{a,b,c\}$ included in $[n]$ such that for every set partition $P=p_1,\ldots,p_m$ of $[n]$, where $p_1=\{a\}$ $p_2=\{b\}$ and $p_3=\{c\}$, the instance $\mathcal{S}^P$ would violate of least one of the necessary conditions (for a square instance to be maximal) described in Lemmas 1, 2, 3 or 4 For instance, if $\mathcal{S}$ contains a subset $\mathcal{S}_i$ of size less or equal than $n-3$, one just needs to take for a,b,c three elements that do not belong to $\mathcal{S}_i$.

Hence, by Lemma 9, we just need to verify that every instance $\mathcal{S}$ taken that dominates $\mathcal{C}(m,n)$ and $\mathcal{D}(m,n)$ is an (m,n)-maximal instance. This point can be shown by induction on $n-m$. If $n-m=1$, for every subset $\{a,b\}$ of $[n]$, it is immediate to verify that $\mathcal{S}^{\{a,b\}}$ dominates one of the instances $\mathcal{A}(n)$ or $\mathcal{B}(n)$ defined in Theorem 2, which, altogether with Lemma 8 implies that $\mathcal{S}$ is an (m,n)-maximal instance. If $n-m>1$, suppose that $\mathcal{S}$ is not an (m,n)-maximal instance. By Lemma 8, there exists a subset $\{a,b\}$ of $[n]$ such that $\mathcal{S}^{\{a,b\}}$ is not an $(m,n-1)$-maximal instance. By induction, it implies that $\mathcal{S}^{\{a,b\}}$ does not dominate one of the instances $\mathcal{C}(m,n-1)$ and $\mathcal{D}(m,n-1)$. This leads to a contradiction with the fact that $\mathcal{S}^{\{a,b\}}$ is a reduction of an instance that dominates one the two instances $\mathcal{C}(m,n)$ and $\mathcal{D}(m,n)$. □

Remark 4 *Theorems 2 and 3 give a characterization of maximal instances $(\mathcal{G}_1,\mathcal{G}_2)$ for the syntenic distance with $n\geq 6$ (where n is the size of $\mathcal{G}_2$). However, the syntenic diameter is know to be equal to $2n-4$*

for any $n \geq 4$, thus the cases $n = 4$ and $n = 5$ also need to be considered. This work has been done using a computer program based on Liben-Nowell's algorithm [11], and is described in the Appendix A.

Remark 5 *Our results (Theorems 2 and 3, and Appendix A) relate the maximality of an instance of the syntenic distance in terms of domination of this instance over a small number of simple instances. It implies immediately that for $n, m \geq 6$ and $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_m\}$ an instance of synteny, one can decide in time polynomial in $n + m$ whether $\mathcal{S}$ is a (m, n)-maximal instance.*

6 Conclusion

In this paper, we answered two open questions from Kleinberg and Liben-Nowell [8] about the syntenic distance - we described a characterization of square maximal instances that induces a polynomial time algorithm deciding whether a square instance is maximal - and we extended our results to the bidimensional case.

An interesting point is the usefulness of the relationship between translocations and calls in the gossiping problem. This fact is central in the proofs of our results.

Thanks to our study, it is also possible to confirm the fact that maximal square instances can be solved by translocations only (that is, fusions and fissions do not necessarily help for instances that are at distance equal to the diameter). This extends a similar result in the case of the n-square instance containing n copies of $[n]$ due to [8]. However, it has been pointed out in [11] that for some instances, fusions and fissions are necessary in order to get the minimum distance between two genomes. Hence, among several open problems, we would like to point out the following one: is it possible to characterize those instances that can be solved by translocations only ?

References

[1] V. Bafna and P. Pevzner. Genome rearrangements and sorting by reversals. *SIAM Journal on Computing*, 25(2):272-289, 1996.

[2] V. Bafna and P. Pevzner. Sorting by transpositions. *SIAM Journal on Discrete Mathematics*, 11(2):224-240, 1996.

[3] A. Bergeron. A very elementary presentation of the Hannenhalli-Pevzner theory. In *Combinatorial Pattern Matching (CPM'01)*, volume 2089 of *Lecture Notes in Computer Science*, pages 106–117. Springer, 2001.

[4] B. DasGupta, T. Jiang, S. Kannan, M. Li and E. Sweedyk. On the complexity and approximation of syntenic distance. *Discrete Applied Mathematics*, 88(1-3):59–82, 1998.

[5] V. Ferretti, J. H. Nadeau and D. Sankoff. Original synteny. In *Combinatorial Pattern Matching (CPM'96)*, volume 1075 of *Lecture Notes in Comput. Sci.*, pages 159–167. Springer, 1996.

[6] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 46(1):1-27, 1999.

[7] S. Hedetniemi, S. Hedetniemi and A. Liestman. A survey of gossiping and broadcasting in communication networks. *Networks*, 18:319-349, 1988.

[8] J. Kleinberg and D. Liben-Nowell. The syntenic diameter of the space of N-chromosomes genomes. In *Conference on Gene Order Dynamics, Comparative Maps, and Multigene Families (DCAF)*, pages 185–197. Kluwer Academic Press, 2000.

[9] D. Liben-Nowell and J. Kleinberg. Structural properties and tractability results for linear synteny. In *Combinatorial Pattern Matching (CPM'00)*, volume 1848 of *Lecture Notes in Computer Science*, pages 248–263. Springer, 2000.

[10] D. Liben-Nowell. On the structure of syntenic distance. *Journal of Computational Biology*, 8(1):53–67, 2001.

[11] D. Liben-Nowell. Gossip is synteny: incomplete gossip and syntenic distance between genomes. *Journal of Algorithms*, 43(2):264–283, 2002.

[12] N. Pisanti and M. F. Sagot. Further thoughts on the syntenic distance between genomes. *Algorithmica*, 34:157–180, 2002.

Appendix A: solving the small square and bidimensional cases.

Square Instances. Theorem 2 gives a characterization of n-square instances with $n \geq 6$. However, the syntenic diameter is known to be equal to $2n-4$ for any $n \geq 4$, thus the cases $n = 4$ and $n = 5$ also need to be considered. This is the purpose of Propositions 7 and 8 below. Before proving those propositions, we need to introduce the following three notations for small instances of the syntenic distance:

$\mathcal{C}(4) = \{\{1,2,3\},\{1,2,3\},\{1,2,3\},\{1,4\}\}$ ($n=4$), $\mathcal{D}(4) = \{\{1,2,3\},\{1,2,3\},\{1,2,4\},\{3,4\}\}$ ($n=4$), $\mathcal{E}(5) = \{[5],[5],\{1,2,3,4\},\{1,2,5\},\{3,4,5\}\}$ ($n=5$).

Proposition 7 *A 4-square instance on $[n]$ is a maximal 4-square instance if and only if it dominates at least one of the five 4-square instances $\mathcal{A}(4)$, $\mathcal{C}(4)$, $\mathrm{Dual}(\mathcal{C}(4))$ or $\mathcal{D}(4)$.*

Proof: Let $\mathcal{F}(4) = \{[4],[4],\{1,2\},\{3,4\}\}$ and $\mathcal{G}(4) = \{\{1,2,3\},\{1,2,3\},\{1,2,3\},\{4\}\}$. Lemma 2 applies when $n = 4$, thus, due to the equivalence and monotonicity properties, any 4-square instance not dominated by one of the seven following instances is not maximal: $\mathcal{A}(4)$, $\mathcal{C}(4)$, $\mathrm{Dual}(\mathcal{C}(4))$, $\mathcal{D}(4)$, $\mathcal{F}(4)$, $\mathrm{Dual}(\mathcal{F}(4))$ and $\mathcal{G}(4)$. We are now going to prove that 3 operations are sufficient to solve $\mathcal{F}(4)$ (and consequently, $\mathrm{Dual}(\mathcal{F}(4))$ by Proposition 1) and $\mathcal{G}(4)$. Indeed, 3 translocations are enough to solve $\mathcal{F}(4)$: consider the gossiping problem corresponding to $\mathcal{F}(4)$, and make the three following calls: $(1,2)$; $(1,3)$; $(2,3)$. The same occurs to $\mathcal{G}(4)$ by making the three following calls: $(1,2)$; $(3,4)$; $(2,4)$. Hence the four remaining candidates are $\mathcal{A}(4)$, $\mathcal{C}(4)$, $\mathrm{Dual}(\mathcal{C}(4))$ and $\mathcal{D}(4)$. By an exhaustive computer search done by a program based on Liben-Nowell's algorithm (see Appendix B and [11]) we get that 3 mutations are not sufficient to solve $\mathcal{A}(4)$, $\mathcal{C}(4)$ or $\mathcal{D}(4)$. Hence the result. □

Proposition 8 *A 5-square instance is a maximal 5-square instance if and only if it dominates at least one of the five 5-square instances $\mathcal{A}(5)$, $\mathcal{B}(5)$, $\mathrm{Dual}(\mathcal{B}(5))$, $\mathcal{E}(5)$ or $\mathrm{Dual}(\mathcal{E}(5))$.*

Proof: Lemmas 1, 2, 3 and 4 apply when $n = 5$, thus any 5-square instance not dominated by one of the five above mentioned instances is not maximal. Since Lemmas 5, 6 and 7 apply for $n = 5$, we only need to prove that $\mathcal{E}(5)$ and $\mathrm{Dual}(\mathcal{E}(5))$ are maximal. By Proposition 1, we know that proving that $\mathcal{E}(5)$ is maximal is sufficient. By an exhaustive computer search (see Appendix B), we get that 5 mutations are not sufficient to solve $\mathcal{E}(5)$. □

Bidimensional Instances. Theorem 3 gives a characterization of (m,n)-instances with $n \geq 6$. However, the syntenic diameter is known to be equal to $m+n-4$ for any $m,n \geq 4$, thus the remaining cases also need to be considered. When the constraint $n > m \geq 6$ of Theorem 3 is not satisfied, we have two subcases: (1) $m = 4$ and $n \geq 5$ and (2) $m = 5$ and $n \geq 6$. Those cases will be considered in this order in Propositions 9 and 10.

In the case where $m = 4$, we need some notations. Indeed, we introduce the instances $\mathcal{A}'(4,n)$, $\mathcal{G}(4,n)$, $\mathcal{I}(4,n)$ and $\mathcal{J}(4,n)$ ($n \geq 5$), which are described in Figure 1. The figure represents different binary matrices with 4 rows, from which every instance $\mathcal{A}'(4,n)$, $\mathcal{G}(4,n)$, $\mathcal{I}(4,n)$ and $\mathcal{J}(4,n)$ will be built. Take for example $\mathcal{A}'(4,n)$: if an element $M_{A'}[i][j]$ in row i and column j of matrix $M_{A'}$ is equal to 1, then the element j is present in ${\mathcal{A}'}_i$. If $M_{A'}[i][j]$ is equal to 0, then j is not present in ${\mathcal{A}'}_i$. The label "**(1)**" (resp. "**(0)**") means that all the elements in this part of the matrix are equal to 1 (resp. 0). We consider that there are a certain number k_i ($1 \leq i \leq 4$) copies of columns with a zero in he i-th row, where each k_i can be as big as $n-4$. Now, any $(4,n)$-instance $\mathcal{A}'(4,n)$ is built from the matrix $M_{A'}$ of Figure 1 as follows: take the four leftmost columns of $M_{A'}$, and pick any $n-4$ columns from the rest of the matrix. Instances $\mathcal{G}(4,n)$, $\mathcal{I}(4,n)$ and $\mathcal{J}(4,n)$ are built from the matrices given in the figure using a similar construction.

Proposition 9 *For any $n \geq 5$, a $(4,n)$-instance is a maximal $(4,n)$-instance if and only if it dominates at least one of the four $(4,n)$-instances $\mathcal{A}'(4,n)$, $\mathcal{G}(4,n)$, $\mathcal{I}(4,n)$ or $\mathcal{J}(4,n)$.*

Proof: It is only a time-consuming exercise to see that the above mentioned instances are the only instances for which any X-reduction leading to a 4-square instance gives one of the 4-square instances listed in Proposition 7. □

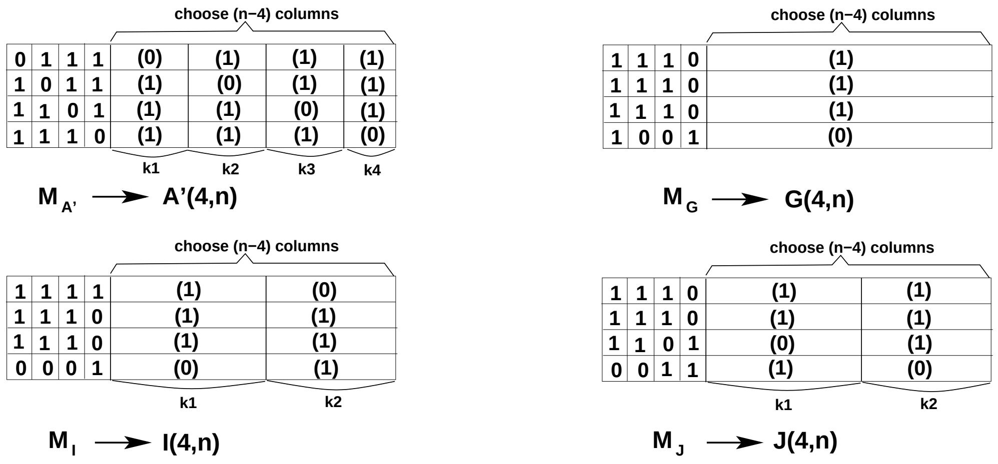

Figure 1: Matrices from which the $(4,n)$-instances $\mathcal{A}'(4,n)$, $\mathcal{G}(4,n)$, $\mathcal{I}(4,n)$ and $\mathcal{J}(4,n)$ are built

In the case where $m = 5$, we introduce a few more notations. They are the following: let $n \geq 6$, $\mathcal{C}(5,n) = \{\mathcal{C}_1, \dots, \mathcal{C}_5\}$ and $\mathcal{F}(5,n) = \{\mathcal{F}_1, \dots, \mathcal{F}_5\}$ be the instances on $[n]$ defined by: $\mathcal{C}_i = [n]/\{i\}$ for $i = 1, \dots, 3$ and $\mathcal{C}_{m-1} = \mathcal{C}_m = [n]/\{4\}$, and $\mathcal{F}_1 = \mathcal{F}_2 = [n]/\{5\}$, $\mathcal{F}_3 = \mathcal{F}_4 = [n]/\{4\}$ and $\mathcal{F}_5 = [n]/\{3\}$.

Moreover, we also introduce the instances $\mathcal{A}''(5,n)$, $\mathcal{B}''(5,n)$ and $\mathcal{E}(5,n)$, which can be constructed from the binary matrices displayed in Figure 2. The representation is similar to the one of Figure 1, but the way to build the corresponding instance is slightly different. For example, take $\mathcal{E}(5,n)$: any $(5,n)$-instance $\mathcal{E}(5,n)$ is built from the matrix M_E of Figure 2 as follows:

- if $n \leq 10$, take the five leftmost columns of M_E, and pick any $n-5$ columns among columns 6 to 10.
- if $n \geq 10$, take the ten leftmost columns of M_E, and pick any $n-10$ columns among the remaining columns.

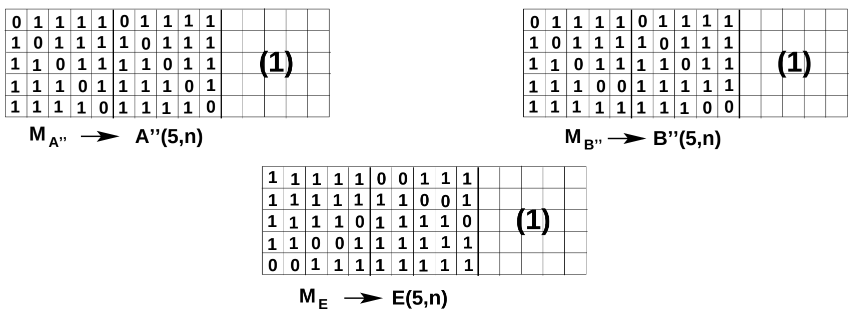

Figure 2: Matrices from which the $(5,n)$-instances $\mathcal{A}''(5,n)$, $\mathcal{B}''(5,n)$ and $\mathcal{E}(5,n)$ are built

Proposition 10 *For any $n \geq 6$, a $(5,n)$-instance is a maximal $(5,n)$-instance if and only if it dominates at least one of the five $(5,n)$-instances $\mathcal{A}''(5,n)$, $\mathcal{B}''(5,n)$, $\mathcal{C}(5,n)$, $\mathcal{E}(5,n)$ or $\mathcal{F}(5,n)$.*

Proof: It is only a time-consuming exercise to see that the above mentioned instances are the only instances for which any X-reduction leading to a 5-square instance gives one the 5-square instances listed in Proposition 8. $\square$

Appendix B: a C program for solving square instances.

```
// Computing the syntenic distance for square instances
// From David Liben-Nowell, Journal of Algorithms 43(2), p 264-283, 2002

#include<stdlib.h>

// Maximal number of genes of an instance
#define P_MAX 100

// An instance S is an array of booleans: S[i][j] = 1 if j is the ith
// chromosome of S.
static int **S1;             // Instance processed by the program
static int M[4*P_MAX];       // A sequence of calls in the gossiping model
static int Distance;         // Final result: the syntenic distance
static int N;                // Size of the input

// ----------------------------------------------------------------------------
// Simulation of a sequence of d/2 calls in the incomplete gossiping
// problem: S2 is the working instance.

static int S2[P_MAX][P_MAX];
static void simulate_calls(int d) {
  int i, j, res;

  // Building the final information to be known after the calls
  for ( i = 0; i < N; i++ ) {
    for ( j = 0; j < N; j++ )
      S2[i][j] = 0;
    S2[i][N-i-1] = 1;
  }

  // Simulation of the sequence of calls
  for ( i = 0; i < d/2; i++ )
    // One exchanges informaion of M[2i-1] and M[2i]
    for ( j = 0; j < N; j++ )
      if ( S2[M[2*(i+1)-1]-1][j] + S2[M[2*(i+1)]-1][j] == 1 )
        S2[M[2*(i+1)]-1][j] = S2[M[2*(i+1)-1]-1][j] = 1;
  // One checks the validity of the sequence of calls with respect to S1
  res = 1;
  for ( i = 0; i < N; i++ )
    for ( j = 0; j < N; j++ )
      if ( S1[i][j] > S2[i][j] )
        res = 0;
  // Updating the minimal distance
  if ( (res == 1) && (d/2 <= Distance) )
    Distance = d/2;
}

// ----------------------------------------------------------------------------
// Exhaustive generation and simulation of all the possible calls sequences.
// Every sequence is a word M of length d at most 2(2DistanceMax-4)
// over the alphabet {0..N-1} such that for every d >= i >= 1,
// M[2i-1]!= M[2i].
```

```
static void gossiping() {
  int i, j, res;

  simulate_calls(0);
  M[0] = 0;
  i    = 1;
  j    = 1;
  while ( i > 0 ) {
    if ( i == (2*Distance)+1 )
      j = M[--i]+1;
    else if ( j >= N+1 )
      j = M[--i]+1;
    else if ( (i % 2 == 0) && (M[i-1] == j) )
      j++;
    else {
      M[i++] = j;
      if ( i % 2 == 1 )
        simulate_calls(i-1);
      j = 1;
    }
  }
}

// -------------------------------------------------------------------------
// Exhaustive generation of all the permutations of the starting instance.
// We use the CAT implementation of the Eades-McKay's algorithm for the
// generation of permutations of a multiset.
// Programmer: Frank Ruskey, 1995.
// Programmer: Joe Sawada, 1997 (translation in C).
// http://www.theory.csc.uvic.ca/~cos/inf/mult/Multiset.html
// For every permutation, we try to solve it with the gossiping procedure.

int A[100], Num[20], Sum[20], Off[20];
int Dir[20];

static void neg(int t, int n, int k);
static void gen(int t, int n, int k);

static void BigGen(int t) {
  if ( Dir[t] )
    gen(t, Sum[t-1], Sum[t]);
  else
    neg(t, Sum[t-1], Sum[t]);

  if ( t > 1 )
    BigGen(t-1);
  Dir[t] = (Dir[t] + 1) % 2;
  if ( Dir[t] )
    Off[t] = Off[t-1] + Num[t-1];
  else
    Off[t] = Off[t-1];
}
```

```
static void swap(int t, int x, int y) {
  int b, temp1, *temp2, i, j;

  if ( t > 1 )
    BigGen(t-1);
  b      = Off[t-1];
  temp1  = A[x+b]; temp2     = S1[x+b-1];
  A[x+b] = A[y+b]; S1[x+b-1] = S1[y+b-1];
  A[y+b] = temp1;  S1[y+b-1] = temp2;

  gossiping();
}

static void gen(int t, int n, int k) {
  int i;

  if ( ( 1 < k ) && ( k < n ) ) {
    gen( t, n-2, k-2 );  swap( t, n-1, k-1 );
    neg( t, n-2, k-1 );  swap( t,  n, n-1 );
    gen( t, n-1, k );
  }
  else if ( k == 1 )
    for ( i = n-1; i >= 1; i-- )
      swap(t,i,i+1);
}

static void neg(int t, int n, int k) {
  int i;

  if ( ( 1 < k ) && ( k < n ) ) {
    neg( t, n-1, k );    swap( t,  n, n-1 );
    gen( t, n-2, k-1 );  swap( t, n-1, k-1 );
    neg( t, n-2, k-2 );
  }
  else if ( k == 1 )
    for ( i = 1; i <= n-1; i++)
      swap(t,i,i+1);
}

// -------------------------------------------------------------------------
// Main procedure
int main(int argc, char *argv[]) {
  int i, j, k, l, t;

  N   = atoi(argv[1]); // Size of the square instance
  t   = atoi(argv[2]); // Number of different subsets of [N]
  j   = 3; // first argument describing the instance S1
  l   = 0;
  Distance = (2*N)-4;

  // Definition the instance S1
  S1 = (int **) malloc(N * sizeof(int *));
  for ( i = 0; i < N; i++ )
    S1[i]  = (int *) calloc(N, sizeof(int));
```

```
  // Reading the instance S1
  for ( i = 0; i <= t; i++ ) {
    Num[i] = atoi(argv[j++]);
    Dir[i] = 1;
    while ( atoi(argv[j]) != 0 )
      S1[l][atoi(argv[j++])-1] = 1;
    j++;
    l++;
    for ( k = 0; k < Num[i]-1; k++ )
      S1[l] = S1[(l++)-1];
  }

  Off[0] = 0;
  for( i = 1; i <= t; i++ )
    Off[i] = Off[i-1] + Num[i-1];
  Dir[t+1] = 1;
  Sum[t]   = Num[t];
  for( i = t-1; i >= 0; i-- )
    Sum[i] = Sum[i+1] + Num[i];

  // Solving the instance S1
  gossiping();
  BigGen(t);
  printf("Distance : %d\n", Distance);
}
```

Institut de Recherche en Informatique de Nantes

On maximal instances for the original syntenic distance

Cedric Chauve and Guillaume Fertin

Abstract

The syntenic distance between two multichromosomal genomes has been introduced by Ferretti, Nadeau and Sankoff as an approximation of the evolutionary distance between genomes for which the gene order is not known. This distance is the minimum number of fusions, fissions and translocations required to transform a genome into the other. Kleinberg and Liben-Nowell proved that for n-chromosomes genomes the diameter for this distance is $2n - 4$ and asked for a characterization of maximal instances for the syntenic distance (pairs of n-chromosomes genomes at a distance of $2n - 4$). Pisanti and Sagot generalized the result of Liben-Nowell and Kleinberg and showed that the maximal distance between a genome with m chromosomes and a genome with n chromosomes is $n + m - 4$. In this paper, we give a characterization of pairs of genomes with respectively n and m chromosomes that are at maximal distance.

IRIN, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3