On maximal instances for the original syntenic distance

Cedric Chauve
LaCIM, Université du Québec à Montréal
CP 8888 Succ. Centre-Ville
H3C 3P8, Montréal (QC), Canada
chauve@lacim.uqam.ca

Guillaume Fertin
IRIN UPRES-EA 2157, Université de Nantes
2 rue de la Houssinière
BP 92208 - F44322 Nantes Cedex 3, France
fertin@irin.univ-nantes.fr

Abstract

The syntenic distance was introduced by Ferretti Nadeau and Sankoff as a mesure of evolutionary distance between multichromosomal genomes in terms of rearrangements at the genes level. In this work, we give a combinatorial description of the pairs of genomes, with a given number of chromosomes and genes, that are at maximal syntenic distance.

1. Introduction.

In the past few years, the interest in determining evolutionary distance between species, in the framework of genomes rearrangements, has been continuously growing. In this context, the models that have been defined 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, this gives rise to several distinct problems, like the *reversal distance* [2], where mutations are described in terms of reversals of chromosomes segments, or the *transposition distance* [1], where a transposition 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 work, we are interested in the *syntenic distance* (introduced by Ferretti, Nadeau and Sankoff [5]), defined below. For this model, we propose a combinatorial characterization of pairs of genomes that are at maximal syntenic distance, called *maximal instances*. This characterization allows to decide efficiently if a pair of genomes is a maximal instance.

2. Syntenic distance: definitions and previous works.

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, and a gene can not appear into several chromosomes. The mutations considered are the *fusions* of two chromosomes (they are joined to form one chromosome), the *fissions* of a chromosome 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 chromosome.

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 and containing respectively n and m chromosomes (we call such a pair of genomes an (m, n)-*instance*), the *syntenic distance* is the minimum number of mutations (fusions, fissions and translocations) needed to transform $\mathcal{G}_1$ into $\mathcal{G}_2$.

Among others, it has been shown that computing the syntenic distance between two arbitrary genomes is NP-hard [4] and some approximation algorithms with approximation ratio 2 are known [4, 5, 7]. However, the algorithmic issue of approximating precisely the general syntenic distance between two genomes still asks for a better solution, and thus for a better understanding of the combinatorial nature of this problem (see [4, 7, 9]). 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). The diameter for the syntenic distance (*syntenic diameter*) between pairs of n-chromosomes genomes has been shown to be equal, for $n \geq 4$, to $2n - 4$ [6, 9]. The computation of the syntenic diameter

was later generalized by Pisanti and Sagot [9] who proved that the maximal distance between an n-chromosomes genome and an m-chromosomes genome is $n + m - 4$, for $n \geq m \geq 4$.

3. Characterization of maximal instances.

As a natural extension of their results on the syntenic diameter, 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 $2n - 4$) [6]. We provide here such a characterization, both for (n, n)-instances and (n, m)-instances when $n > m$.

Our results are defined in terms of *compact representation* of an instance for the syntenic distance [5] and of *domination* of an instance over a second one [7]. 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 a 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 call such an instance a *compact instance*. A compact instance $\mathcal{S} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ dominates a compact instance $\mathcal{S}' = \{\mathcal{S}'_1, \dots, \mathcal{S}'_{m'}\}$ if there is a permutation $i_1, \dots, i_{m'}$ of $[m']$ such that for if for every $j \in [m']$, one has $\mathcal{S}'_{i_j} \subseteq \mathcal{S}_i$.

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

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

Remark 1. In [3] characterizations for maximal instances are given in the case where $m = n \leq 6$, $m < n \leq 6$ or $n < m$.

Remark 2. One can deduce immediately from our two main results (Theorems 1 and 2) a polynomial time (in $n + m$) algorithm deciding whether an (m, n)-instance for the syntenic distance is maximal.

Remark 3. The proofs of our results rely on the importance of the translocations and the equivalence between sequences of translocations and a problem of communication in networks, the *incomplete gossiping* [6, 8]. These two points could be of great interest to identify classes of instances for the syntenic distance that can be solved in polynomial time.

References

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

[2] A. Bergeron. A very elementary presentation of the Hannenhalli-Pevzner theory. In *CPM'01* (A. Amir and G. M. Landau, editors), volume 2089 of *Lecture Notes in Computer Science*, pages 106–117. Springer, 2001.

[3] C. Chauve and G. Fertin On maximal instances for the original syntenic distance. Report 03.03, IRIN, Nantes, 2003.

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

[5] V. Ferretti, J. H. Nadeau and D. Sankoff. Original synteny. In *CPM'96* (D. Hirschberg and G. Myers, editors), volume 1075 of *Lecture Notes in Computer Science*, pages 159–167. Springer, 1996.

[6] J. Kleinberg and D. Liben-Nowell. The syntenic diameter of the space of N-chromosomes genomes. In *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families* (D. Sankoff and J. H. Nadeau, editors), pages 185–197. Kluwer Academic Publishers, 2000.

[7] D. Liben-Nowell. On the structure of syntenic distance. *J. Comput. Biol.*, 8(1):53–67, 2001.

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

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