Conserved Interval Distance Computation between Non-trivial Genomes[⋆]

Guillaume Blin[1] and Romeo Rizzi[2]

[1] LINA FRE CNRS 2729 Université de Nantes, 2 rue de la Houssinière
BP 92208 - 44322 Nantes Cedex 3 - FRANCE
blin@lina.univ-nantes.fr
[2] Universit degli Studi di Trento - Dipartimento di Informatica e Telecomunicazioni
Via Sommarive, 14 - I38050 Povo - Trento (TN) - ITALY
Romeo.Rizzi@unitn.it

Abstract. Recently, several studies taking into account the ability for a gene to be absent or to have some copies in genomes have been proposed, as the examplar distance [11, 6] or the gene matching computation between two genomes [10, 3]. In this paper, we study the time complexity of the conserved interval distance computation considering duplicated genes using both those two strategies.

Keywords: Conserved interval distance, Exemplar string, Matching, Computational complexity, Longest Common Substring, Duplicated genes.

1 Introduction

In comparative genomics, gene order study in a set of organisms has been intensively led essentially in phylogenetic research field [5, 2, 4]. Most of the methods associated to gene order study are based on a distance computation. This distance has to reflect the number of genetic operations needed to transform a source genome into a target genome. For this purpose, a set of distances and associated methods have been developed in the past decade. Among others, we can mention three intensively studied distances: *edit* [9, 12], *breakpoint* [3], and *conserved interval* [1] distances.

From an algorithmic point, distances can roughly be defined as follows: given a set $\mathcal{F}$ of *gene families*, two *genomes* G and H, represented as sequences of signed elements (genes) from $\mathcal{F}$, and a set of evolutionary operations that operate on segments of genes (like reversals, transpositions, insertions, duplications, deletions for example), the distance between G and H is the minimum number of operations needed to transform G into H.

Until recently, the assumption that *in a genome there is no copy of a gene* was a requirement of most of the methods associated to gene order study. This restriction reduces the problem to the comparison of signed permutations [8]. It is known that this assumption is very restrictive and is only justified in small

[⋆] This work was partially supported by the French-Italian PAI Galileo project number 08484VH

virus genomes, therefore one needs to consider genomes containing duplicated genes.

In [11], Sankoff has proposed a method to select, from the set of copies of a gene, the common ancestor gene such that the distance between the reduced genomes is minimized. In [6], Bryant proved that the corresponding problem, so called *exemplar string*, was **NP**-complete for two distances: the signed reversals and the breakpoint distances. Marron *et al.* have proposed in [12] methods relying on a matching between genes of two genomes. Provided with a matching between genes of the two genomes, one can, by a rewriting of the genomes according to the matching, create genomes without duplicated genes and solve the reduced problem.

In this paper, we investigate the complexity of both the use of exemplar strategy and of matchings to compute the conserved interval distance between genomes containing duplicated genes. First we prove that the use of both strategies unfortunately induces **NP**-completeness. To overstep **NP**-hardness of problems, many techniques have been developed: heuristic, parameterized complexity and approximation algorithm. For biological problems those alternative techniques have been intensively used, since in most cases specific properties of the problem are not taken into account in the **NP**-hardness proof.

This paper is organized as follows. After presenting some preliminaries in Section 2, we show in Section 3 that both the use of exemplar strategy and of matchings to compute the conserved interval distance between genomes containing duplicated genes induces **NP**-completeness. Then in Section 4, we present a heuristic approach based on the Longest Common Substring which have been implemented and tested over a set of 20 bacteria.

2 Preliminaries

Genomes, gene families and gene. Following terminology introduced in [11], a *genome* G is a sequence of elements of an alphabet $\mathcal{F}$ (referred as the set of *gene families*) such that each element is provided with a sign (+ or $-$). Each occurrence of a gene family from $\mathcal{F}$ in G is called a *gene*. Given a genome $G = g_1 g_2 \ldots g_n$, we say that gene g_i precedes gene g_{i+1}. For two genomes G and H and a gene family $\mathbf{f}$, the number of occurrences of $\mathbf{f}$ in G and H is called the *cardinality* of family $\mathbf{f}$. A gene family $\mathbf{f}$ is said to be *trivial* if $\mathbf{f}$ has cardinality exactly 1 or 2. Otherwise, $\mathbf{f}$ is said to be *non-trivial*. A gene belonging to a trivial (resp. non-trivial) family is said to be trivial (resp. non-trivial). A segment (*i.e.* a substring) of G that contains only non-trivial genes is called a *non-trivial segment*. We say that two genomes G and H are *balanced* if, for any gene family $\mathbf{f}$, there are as many occurrences of $\mathbf{f}$ in G as in H.

Conserved interval, conserved interval distance. Following terminology introduced in [1], given a set of n genomes $\mathcal{G}$ and two genes $a, b \in \mathcal{F}$, an interval $[a, b]$ is a *conserved interval* of $\mathcal{G}$ if (1) either a precedes b, or -b precedes -a in each genome of $\mathcal{G}$ and (2) the set of unsigned genes (*i.e.* not considering signs) appearing between genes a and b is the same for all genomes of $\mathcal{G}$. For example, given two genomes $G_1 = a\ b\ c\ g\ e\ f$ -$d\ h$ and $G_2 = a\ g$ -c -$b\ e$ -f -$d\ h$, there are seven conserved intervals between G_1 and G_2: $[a,\text{-}d]$, $[a,e]$, $[a,h]$, $[b,c]$, $[e,\text{-}d]$, $[e,h]$ and $[\text{-}d,h]$.

Given two set of genomes $\mathcal{G}$ and $\mathcal{H}$, the *conserved interval distance* between $\mathcal{G}$ and $\mathcal{H}$ is defined by $d(\mathcal{G},\mathcal{H}) = N_{\mathcal{G}} + N_{\mathcal{H}} - 2N_{\mathcal{G}\bigcup\mathcal{H}}$ where $N_{\mathcal{G}}$ (resp. $N_{\mathcal{H}}$ and $N_{\mathcal{G}\bigcup\mathcal{H}}$) is the number of conserved intervals in $\mathcal{G}$ (resp. $\mathcal{H}$ and $\mathcal{G}\bigcup\mathcal{H}$). For example, let $\mathcal{G} = \{G_1, G_2\}$ and $\mathcal{H} = \{H_1, H_2\}$ be two sets of genomes where G_1 and G_2 are as above and $H_1 = a\ e$ -$f\ b\ g\ c$ -$d\ h$ and $H_2 = a\ f$ -c -$g\ b$ -e -$d\ h$. We obtain $d(\mathcal{G},\mathcal{H}) = 7 + 3 - 4 = 6$. In the rest of the paper, for readability, we denote the conserved interval distance between two singleton sets $d(\{G\},\{H\})$ by $d(G,H)$.

Gene matching. Let $G = g_1g_2\dots g_n$ and $H = h_1h_2\dots h_m$ be two genomes on $\mathcal{F}$. A *gene matching* $\mathcal{M}$ between G and H is a maximal matching between genes of G and H such that, for every pair $(g_i,h_j) \in \mathcal{M}$, g_i and h_j belong to the same family. By maximal matching, we mean that for any gene family **f**, it is forbidden to have at the same time an occurrence of **f** in G and one in H that do not belong to $\mathcal{M}$. It follows from the maximality condition of matchings that in any matching $\mathcal{M}$ between balanced genomes G and H, every gene of G is matched to a gene of H and conversely. Given a matching $\mathcal{M}$ and a segment $s = s_1s_2\dots s_m$ of G if, for all $1 \le i \le m$, $(s_i,t_i) \in \mathcal{M}$ such that: (1) $s_i = t_i$ and $t = t_1t_2\dots t_m$ is a segment of H or (2) $s_i = -t_i$ and $t = t_mt_{m-1}\dots t_1$ is a segment of H then s is *perfectly matched* in $\mathcal{M}$; *not-perfectly matched* otherwise.

Minimum Conserved Interval Matching. Given two genomes G, H and a gene matching $\mathcal{M}$, we denote by $d(G,H,\mathcal{M})$ the conserved interval distance between G and H with respect to $\mathcal{M}$, and by $d(G,H)$ the conserved interval distance between G and H, defined as the minimum $d(G,H,\mathcal{M})$ among all matchings $\mathcal{M}$. The matching $\mathcal{M}$ such that $d(G,H,\mathcal{M}) = d(G,H)$ is a *minimum conserved interval matching.*

3 Hardness results

In this section, we first prove that, even if there is no non-trivial segment containing more than one gene, Exemplar Conserved Interval Distance (ECID) problem (formalized hereafter) is **NP**-complete. Then, we prove that, even with just one non-trivial gene family, the problem of finding a Minimum Conserved Interval Matching (MCIM) is **NP**-complete. This last result is based on a polynomial time reduction which is inspired from the one presented in [3] which proves that a connected problem, Minimum Breakpoint Matching, is **NP**-complete.

Theorem 1. Exemplar Conserved Interval Distance *problem is* **NP**-*complete even when all non trivial segments are composed of only one duplicated gene.*

To prove the correctness of Theorem 1, we provide a polynomial time reduction from the **NP**-complete problem Minimum Set Cover [7]. Following terminology introduced in [7], we recall that given a collection C of subsets of a finite set E, a *cover* for E is a subset $C' \subseteq C$ such that every element of E belongs to at least one member of C'. For the sake of clarity, we now state formally the two decision problems we consider: ECID and Minimum Set Cover. Given two genomes G and H, and a positive integer k, ECID problem asks whether it is possible to find two exemplar genomes G' and H' of resp. G and H such that $d(G', H') \leq k$. Given a collection C of subsets of a finite set E and a positive integer k', Minimum Set Cover problem asks whether C contains a cover C' for E s.t. $|C'| \leq k'$.

Hereafter, we consider, w.l.o.g., that in each $C_i \in C$, any $e_j \in C_i$ is also in C_j with $1 \leq i, j \leq m$ and $i \neq j$. In fact, by definition, if an element is covered by only one subset then this subset must be part of C'. In the following, we will prove that, even if G does not contain more than one occurrence of each duplicated gene, ECID problem is **NP**-complete. Given an integer k', two exemplar genomes G' and H', one can compute polynomially $d(G', H')$ and check if $d(G', H') \leq k'$ (see [1]). Therefore, ECID problem is in **NP**. The remainder of the section is devoted to proving that it is also **NP**-hard. For this purpose, we reduce Minimum Set Cover problem to ECID problem. Let $C = \{C_1, C_2 \ldots C_m\}$ be a collection of m subsets of a finite set $E = \{e_1, e_2 \ldots e_n\}$ of n elements.

In the rest of this section, we consider that $\mathcal{F} \subseteq \mathbb{N}$ but any genome is built with elements of $\mathcal{F}$ provided with signs (*i.e.* $\mathbb{R}$). In other words, genes 3 and -3 are of the same family. Let us detail the construction of the two genomes G and H. Let $y = |E| + 2$ if $|E|$ is even, $y = |E| + 1$ otherwise. Let $z_i = (y + 2).(i - 1)$ for any $1 \leq i \leq m + 1$. From (C, E), we construct two genomes G and H as described below (an illustration is given in Figure 1):

$$G_1 = \gamma_{|E|+1}\ \gamma_{|E|+2} \cdots \gamma_{|E|+m-1} \alpha_1\ \beta_1\ \ldots \alpha_m\ \beta_m\ \gamma_1\ \gamma_{|E|+m}\ \gamma_2\ \gamma_{|E|+m+1} \cdots \gamma_{2|E|+m-1}\ \gamma_{|E|}$$
$$H_1 = \alpha_1\ \theta_1\ \gamma_{|E|+1}\ \alpha_2\ \theta_2\ \gamma_{|E|+2} \cdots \gamma_{|E|+m-1}\ \alpha_m\ \theta_m\ \gamma_{|E|+m}\ \gamma_{|E|+m+1}\ \cdots\ \gamma_{2|E|+m-1}$$

We now detail the substrings that compose G_1 and H_1:

- for $1 \leq i \leq m$, we construct the sequences of genes $\alpha_i = z_i$ and $\beta_i = z_i{+}1\ z_i{+}2 \ldots z_i{+}y{+}1$;
- for $1 \leq i \leq 2|E| + m - 1$, we construct a gene $\gamma_i = z_{m+1} + i$;
- for $1 \leq i \leq m$, we construct a gene $\theta_i = z_i{+}2\ z_i{+}4 \ldots z_i{+}y\ z_i{+}1\ z_i{+}3 \ldots z_i{+}y\text{-}1\ z_i{+}y{+}1$.

Note that G_1 and H_1 are two exemplar genomes. Genome G is, thus, a copy of G_1. We now turn to transform H_1 into a non-exemplar genome H: for $1 \leq i \leq m$ and $1 \leq j \leq |E|$, if $e_j \in C_i$ then gene γ_j is inserted between the j^{th} and the $j{+}1^{th}$ genes of θ_i. We denote by ECID-construction any construction of this type. An illustration of an ECID-construction is given in Figure 1. Intuitively, θ_i is a shuffle of β_i with some inserted γ_js (*i.e.* no conserved adjacencies) for $1 \leq i \leq m$. Clearly, our construction can be carried out in polynomial time. Moreover, the result of such a construction is indeed an instance of ECID problem.

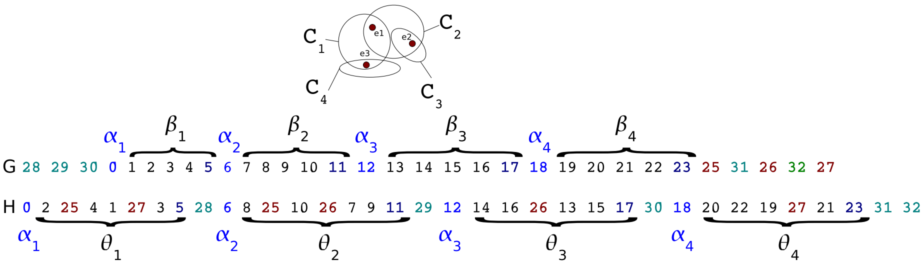
C1
C2
C3
C4
e1
e2
e3
α1 β1 α2 β2 α3 β3 α4 β4
G 28 29 30 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 31 26 32 27
H 0 2 25 4 1 27 3 5 28 6 8 25 10 26 7 9 11 29 12 14 16 26 13 15 17 30 18 20 22 19 27 21 23 31 32
α1 θ1 α2 θ2 α3 θ3 α4 θ4

Fig. 1. Example of an ECID-construction where $E = \{e_1, e_2, e_3\}$ and $y = 4$.

We now turn to proving that our construction is a polynomial time reduction from Minimum Set Cover to ECID problem where G is an exemplar genome whereas H is not. Let first note that, by construction, there are only $|E|$ duplicated gene families in G and H, namely the γ_is for $1 \leq i \leq |E|$.

Lemma 1. *The only conserved intervals that can exist between G and any exemplar genome H' of H are intervals $[\alpha_i, z_i + y + 1]$ such that all the γ_js of $[\alpha_i, z_i + y + 1]$ in H' have been deleted, with $1 \leq i \leq m$ and $1 \leq j \leq |E|$.*

Lemma 2. *Let $I = (C, E, k')$ be an instance of the problem* Minimum Set Cover *with a collection $C = \{C_1, C_2 \ldots C_m\}$ of m subsets of a finite set $E = \{e_1, e_2 \ldots e_n\}$, and $I' = (G, H, k)$ an instance of* ECID *problem obtained by an* ECID*-construction from I. C contains a cover C' of E of size less than or equal to k' iff $d(G, H') \leq k$ where H' is an exemplar genome of H and $k = |G|.|G-1| - 2(m - k')$.*

Proof. ($\Rightarrow$) Suppose C contains a cover C' of E of size less than or equal to k'. Let $f : e_i \rightarrow \{C_1, C_2, \ldots C_m\}$ be the function which, given an element of E, returns the index of the subset covering this element in C'. Let $I' = (G, H, k)$ be the instance obtained from an ECID-construction of I. We look for an exemplar genome H' of H such that $d(G, H') \leq k$. We define H' as follows: for each $e_j \in E$, delete γ_j of θ_p for all $p \in \{1, 2, .., m\}/\{f(e_j)\}$.

By construction, E denotes the set of duplicated gene families and by construction the only duplicated genes in H are the γ_is. Therefore, H' is exemplar since one deletes all occurrences but one of γ_i with $1 \leq i \leq |E|$. Remains us to prove that $d(G, H') \leq k$. By definition, for each $C_j \notin C'$ and each $e_i \in C_j$, $f(e_i) = p$ with $p \neq j$ and γ_i of θ_j has been deleted. Since all the γ_is of θ_j in H' have been deleted, there is a conserved interval $[\alpha_j, z_j + y + 1]$ between G and H'. Globally, there are at least $m - k'$ such subsets. Therefore, there are at least $m - k'$ conserved intervals between G and H'. Thus, $d(G, H') \leq |G|.|G-1| - 2(m - k')$, since the number of conserved intervals between a genome G and itself is $\frac{|G|.|G-1|}{2}$ and $|G| = |H'|$.

($\Leftarrow$) Suppose we have an exemplar genome H' of H such that $d(G, H') \leq k$. Assume, w.l.o.g., that $d(G, H') = d' \leq k$. We define C' as follows: in H', if $\gamma_j \in \theta_p$ then $f(e_j) = p$ and $C_p \in C'$. We now turn to proving that C' defines a cover of E of size at most k'. Since H' is an exemplar genome of H, there is exactly one occurrence of each gene family in H'. Therefore, C' contains a set of subsets that covers E. Remains us to prove that $|C'| \leq k'$.

By definition, since $d' \leq |G|.|G-1| - 2(m-k')$, there are at least $m-k'$ conserved intervals between G and H'. By Lemma 1, the only conserved intervals that can exist between G and any exemplar genome H' of H are intervals $[\alpha_i, z_i + y + 1]$ such that all the γ_js of $[\alpha_i, z_i + y + 1]$ in H' have been deleted. Therefore, by construction, there are at least $m-k'$ such intervals $[\alpha_i, z_i + y + 1]$ in H'. Correctness of Theorem 1 follows. □

Given Theorem 1, one can ask if instead of deleting the duplicated genes, one can compute the interval distance taking into account duplicated genes [3, 12]. For this purpose, we propose the MCIM problem: finding a minimum conserved interval matching between two genomes. Unfortunately, as we will show hereafter, this problem is also **NP**-complete.

Theorem 2. Minimum Conserved Interval Matching *problem is* **NP**-*complete.*

To prove the correctness of Theorem 2, we provide a polynomial time reduction from the **NP**-complete problem Minimum Bin Packing [7]. For the sake of clarity, we now state formally the two decision problems we consider: MCIM and Minimum Bin Packing. Given two genomes G and H, and an integer k, MCIM problem asks whether it is possible to find a matching between G and H such that $d(G,H) \leq k$. Given a finite set $U = \{u_1, u_2, \dots, u_n\}$, a size $s(u) \in \mathbb{Z}^+$ for each $u \in U$ and two positive integers k' and $\mathcal{C}$, Minimum Bin Packing problem asks whether there is a partition of U into k' disjoint sets $U_1, U_2, \dots, U_{k'}$ such that $\sum(s(u)|u \in U_i) \leq \mathcal{C}$ for each U_i.

It is easily seen that MCIM is in **NP** since given an integer k and a set of matchings between two genomes we can polynomially compute the number of conserved intervals between G and H and thus check if the distance is less than or equal to k (see [1]). The remainder of the section is devoted to proving that MCIM is also **NP**-hard even when there is only one non trivial family in $\mathcal{F}$, which implies Theorem 2. For this, we reduce Minimum Bin Packing problem to MCIM problem. Let $\mathcal{N} = k'.\mathcal{C} - \sum_{i=1}^{n} s(u_i)$.

Let us first detail the construction of genomes G and H from a Minimum Bin Packing instance $(U, k', \mathcal{C})$. The gene families are $\mathcal{F} = \{\alpha, \beta, \mathbf{x}, A_1, A_2 \dots, A_{n+\mathcal{N}}, B_1, B_2, \dots, B_{k'+1}\}$. On the whole, there are $k' + \mathcal{N} + n + 4$ families of genes, and $\mathbf{x}$ is the unique non-trivial family. For $1 \leq i \leq n$ (resp. $n+1 \leq i \leq n+\mathcal{N}$), we denote by $\mathbf{u'_i}$ a sequence of $s(u_i)$ consecutive genes $\mathbf{x}$ (resp. one gene $\mathbf{x}$). For $1 \leq j \leq k'$, $\mathbf{U'_j}$ represents a sequence of $\mathcal{C}$ consecutive genes $\mathbf{x}$. G and H are defined as follows:

$G = \alpha\ \mathbf{u'_1}\ A_1\ \mathbf{u'_2}\ A_2\ \dots\ \mathbf{u'_{n+\mathcal{N}}}\ A_{n+\mathcal{N}}\ B_1\ B_2\ \dots\ B_{k'+1}\ \beta$
$H = \alpha\ B_1\ \mathbf{U'_1}\ B_2\ \mathbf{U'_2}\ \dots\ B_{k'}\ \mathbf{U'_{k'}}\ B_{k'+1}\ A_1\ A_2\ \dots\ A_{n+\mathcal{N}}\ \beta$

An illustration of such a construction, that can obviously be achieved in polynomial time, is given in Figure 2. To complete the construction of the instance of MCIM, it remains to us to define k: $k = \frac{|G|.(|G|-1)}{2} + \frac{|H|.(|H|-1)}{2} - 2q$ with $q = 1 + \sum_{i=1}^{n} \frac{s(u_i).(s(u_i)-1)}{2}$

In the next three lemmas, that establish that MCIM is **NP**-complete, we consider an instance $(U, k', \mathcal{C})$ of Minimum Bin Packing and the corresponding instance (G, H, k) of MCIM according to the above construction.

Lemma 3. *Given a matching $\mathcal{M}$, a non trivial segment of size p perfectly matched in $\mathcal{M}$ induces more conserved intervals (*i.e. $\frac{p(p-1)}{2}$*) than a non-trivial segment of size k not-perfectly matched.*

Lemma 4. *$d(G,H) \geq k$ and in any matching $\mathcal{M}$ between G and H, any conserved interval I with respect to $\mathcal{M}$ is either $[\alpha, \beta]$ or $I = [p,q]$ with $S[p..q]$ being a non trivial segment.*

Lemma 5. *There is a partition of U into k' disjoint sets $U_1, U_2, \ldots, U_{k'}$ such that the sum of the sizes of the elements in each U_i is at most $\mathcal{C}$ if and only if $d(G,H) \leq k$.*

Proof. ($\Leftarrow$) Suppose that $d(G,H) \leq k$. By Lemma 4, we know that $d(G,H) = k$, and any conserved interval I with respect to a minimum conserved interval matching between G and H is either $[\alpha, \beta]$ or $I = [p,q]$ with $S[p..q]$ being a non trivial segment. Moreover, if $d(G,H) = k$ then the number of conserved intervals should be maximal ($i.e.$ $1 + \sum_{i=1}^{n} \frac{s(u_i).(s(u_i)-1)}{2}$). Therefore, by Lemma 3, any non-trivial segment u'_i, with $1 \leq i \leq n$, in G should be matched with a sequence of consecutive genes $\mathbf{x}$ in H. Precisely, for any $1 \leq i \leq n$, there is a given $1 \leq j \leq k'$ such that the sequence u'_i is perfectly matched with a substring of U'_j as illustrated in Figure 2.

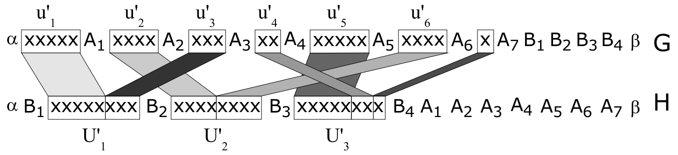
u'1
u'2
u'3
u'4
u'5
u'6
α xxxxx A1 xxxx A2 xxx A3 xx A4 xxxxx A5 xxxx A6 x A7 B1 B2 B3 B4 β
G
α B1 xxxxxxxx B2 xxxxxxxx B3 xxxxxxxx B4 A1 A2 A3 A4 A5 A6 A7 β
H
U'1
U'2
U'3

Fig. 2. Instance of MCIM associated to the Minimum Bin Packing instance where $k' = 3$, $\mathcal{C} = 8$ and $U = \{u_1, \ldots, u_6\}$ with $s(u_1) = s(u_5) = 5$, $s(u_2) = s(u_6) = 4$, $s(u_3) = 3$ and $s(u_4) = 2$ and the gene to gene matching corresponding to the following partition of $U : U_1 = \{u_1, u_3\}$, $U_2 = \{u_2, u_6\}$ and $U_3 = \{u_4, u_5\}$

Therefore, such a matching induces a partition P of the set of sequences $\{u'_1, u'_2, \ldots, u'_n\}$ into at most k' disjoint sequences $U'_1, U'_2, \ldots, U'_{k'}$. As, by construction, $|U'_i| = \mathcal{C}$ for $1 \leq i \leq k'$, P corresponds to an answer to the corresponding Minimum Bin Packing instance.

($\Rightarrow$) Suppose we have a partition P of U into disjoint sets $U_1, U_2, \ldots, U_{k'}$ each of cardinality at most $\mathcal{C}$. We compute a gene matching $\mathcal{M}$ between G and

H as follows: (1) each trivial gene in G is matched with its corresponding gene in H and (2) for $1 \leq j \leq k'$, for each $u_i \in U$, if $u_i \in U_j$, then the sequence of genes $\mathbf{x}$ of u'_i in G is perfectly matched with the first free (*i.e.* not already matched) sequence of genes $\mathbf{x}$ of U'_j in H.

Since $\mathcal{M}$ is built according to P, we claim that each non-trivial segment u'_i, with $1 \leq i \leq n$, is matched to a contiguous sequence of genes $\mathbf{x}$ in H. Thus, any non-trivial segment u'_i, with $1 \leq i \leq n$, in G induces $\frac{s(u_i).(s(u_i)-1)}{2}$ conserved intervals. Therefore, $\mathcal{M}$ induces $1 + \sum_{i=1}^{n} \frac{s(u_i).(s(u_i)-1)}{2}$ conserved intervals (see proof of Lemma 4). This leads to $d(G, H, \mathcal{M}) \leq k$, and so to $d(G, H) \leq k$. Correctness of Theorem 2 follows. □

4 Using the L.C.Substring to approximate MCIM

In this section, we present a heuristic approach to solve MCIM problem that performs well on real data. This approach uses the following intuition: *long segments of genes that match, up to a complete reversal, in two genomes are likely to belong to a Minimum Conserved Interval Matching.* Remark that given two trivial genomes this intuition gives the optimal solution. Unfortunately, this is not always the case when considering non trivial genomes. Even so, this approach works very well on real genomes. In the rest of this section, we consider two genomes G and H build with elements of $\mathcal{F} \subseteq \mathbb{N}$ and provided with signs.

Our approach to solve MCIM problem is based on the following loop:

1. Identify a longest common segment s of genes between G and H (by common segment, we mean a segment appearing, up to a complete reversal, both in G and H);
2. Replace s in both genomes by an integer g_c, further call *compressed gene*, s.t. $g_c \notin \mathcal{F}$ (this induces that g_c is a trivial gene);
3. Mark g_c as treated;
4. Store the number of genes of s in $N_s[g_c]$ and the set of genes of s in $C[g_c]$.

While a common segment of unmarked genes exists, the algorithm performs the loop described above on the modified genomes. In the following, we will refer to the modified versions of the genomes G and H as G' and H'. Once the algorithm exits of the loop, any unmarked gene g_u of G' and H' is deleted since, by definition, g_u is not common (*i.e.* there are more genes g_u in one of the genomes). This first step of the algorithm leads to the computation of two exemplar genomes G' and H' of resp. G and H. Clearly, this step gives the gene to gene matching of G and H in a compressed version: the corresponding matching $\mathcal{M}$ is obtain by perfectly matching the segments of genes corresponding to each compressed gene of G as illustrated in Figure 3.

In a second step, the algorithm computes the interval distance induced by $\mathcal{M}$. By Lemma 3, each compressed gene g_c of G' induces $\frac{k(k-1)}{2}$ conserved intervals between G and H with $k = N_s[g_c]$. Moreover, some conserved intervals between G' and H' may exist. Therefore, since G' and H' are trivial genomes, the algorithm computes the set of conserved intervals S_{ci} between G' and H' in

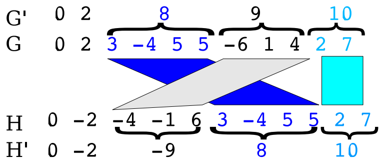

Fig. 3. The matching of the genes of G and H deduced from the exemplar genomes G' and H'.

polynomial time using the algorithm defined in [1]. Since G' and H' are composed of compressed genes, for each conserved interval $[g_{c1}, g_{c2}] \in S_{ci}$, $N_s[g_{c1}].N_s[g_{c2}]$ conserved intervals between G and H are induced. Indeed, if $[g_{c1}, g_{c2}] \in S_{ci}$ then a segment of genes $g_{c1}\lambda g_{c2}$ appears in G' and either a segment of genes $g_{c1}\lambda' g_{c2}$ or $-g_{c2}\lambda' - g_{c1}$ appears in H' with λ and λ' being similar segments of genes not considering genes order and sign. Therefore, considering $\mathcal{M}$, for any genes $g_i \in C[g_{c1}]$ and $g_j \in C[g_{c2}]$, $[g_i, g_j]$ is a conserved interval between G and H. This step returns $d(G,H,\mathcal{M}) = |G|.|G-1| - 2(\sum_{g_c \in G'} \frac{N_s[g_c].(N_s[g_c]-1)}{2} + \sum_{[g_{c1},g_{c2}] \in S_{ci}} N_s[g_{c1}].N_s[g_{c2}])$

We implemented our approach using a suffix tree. Indeed, longest common segments between G and H can be found in linear time by browsing a suffix tree built on G, H and the reversed of H. To test our algorithm and get an estimate of its performance in practice, we applied our heuristic approach to a set of 20 bacteria from NCBI.

Data and programs used and mentioned in this article can be found at
`http://www.sciences.univ-nantes.fr/info/perso/permanents/blin/Cocoon05/`

Interesting characteristics of this set are given on the web page. We implemented the brute force algorithm which consists in computing all possible matchings and we compared the obtained results. In average, the gap between the optimal solution *opt* and the solution given by our algorithm is less than $0,12\%$ of *opt*. We noticed that more than $\frac{2}{3}$ of the bacteria have duplicated genes with, for most of them, duplicated families of cardinality 2. The effectiveness of our algorithm relies on the fact that the number of duplicated genes are not significant compared to the size of the genomes. Moreover, since our algorithm gives the optimal solution for trivial genomes, duplicated genes have a very little impact on the results.

5 Conclusion

The assumption of uniqueness of each gene in a genome has been a requirement of traditional methods in comparative genomics but is only justified in small virus genomes, since in general, there are more than one copy of a gene in a genome. In this paper, we investigate the time complexity of the conserved interval distance computation considering duplicated genes. We proved that both use of exemplar and matching methods leads to **NP**-completeness. We are doing thorough experimental testing which will determine how well our algorithm does in practice under different regimes of duplication, but our preliminary results are extremely encouraging.

Note that, since the Brute Force Algorithm is in $O(k^{k.l}n)$ with k being the maximal cardinality of any non-trivial gene family, l being the number of non-trivial families and n being the size of the genomes, MCIM problem is in **FPT** for parameter k and l. In order to be usable in many reconstruction algorithms, it would be of interest to determine if the problem is in **FPT** for other parameters.

References

1. A. Bergeron and J. Stoye. On the similarity of sets of permutations and its applications to genome comparison. *Proceedings of COCOON 03*, 2697 of LNCS:68–79, 2003.
2. M. Blanchette, T. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.*, 49(2):193–203, 1999.
3. G. Blin, C. Chauve, and G. Fertin. The breakpoints distance for signed sequences. In *Actes de* CompBioNets 2004, volume 3 of *Texts in Algorithms*, pages 3–16. KCL Publications, 2004.
4. G. Bourque and P. A. Pevzner. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res.*, 12(1):26–36, 2002.
5. G. Bourque, P.A. Pevzner, and G. Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse and rat genomes. *Genome Res.*, 14(4):507–516, 2004.
6. D. Bryant. The complexity of calculating exemplar distances. In D. Sankoff and J. Nadeau, editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*, pages 207–212. Kluwer Acad. Pub., 2000.
7. M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.
8. O. Gascuel, editor. *Mathematics of Evolution and Phylogeny.* Oxford Univ. Press, 2004. To appear.
9. M. Marron, K.M. Swenson, and B.M.E. Moret. Genomic distances under deletions and inversions. *Proceedings of COCOON 03*, 2697 of LNCS:537–547, 2003.
10. B. M. E. Moret, A. C. Siepel, J. Tang, and T. Liu. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In *WABI 2002*, volume 2452 volume of LNCS, pages 521–536. Springer Verlag, 2002.
11. D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15(11):909–917, 1999.
12. K.M. Swenson, M. Marron, J.E Earnest-DeYoung, and B.M.E. Moret. Approximating the true evolutionary distance between two genomes. Technical Report TR-CS2004-15, Department of Computer Science, University of New Mexico, 2004.