The Complexity of Calculating Exemplar Distances

David Bryant *

Abstract

Traditional methods for estimating rearrangement distances between genomes assume that there is at most one copy of each gene in each genome. In the case that there are multiple genes from the same gene family in a genome, Sankoff (1999) proposes the estimation of *true exemplars*, a selection of one gene from each gene family in both genomes such that the distance between the resulting *exemplar strings* is minimized. This is the *exemplar distance*. Here we show that the calculation of the exemplar distance between two genomes is NP-hard for both the signed reversals distance and the breakpoint distance.

1 Introduction

The comparative study of gene order rearrangements has, for the most part, been restricted to the case when the genes in one genome are homologous to at most one gene in the other genome. In many small virus or mitochondrial genomes, the single homologue assumption is justified. In most cases, however, there can be multiple copies of the same gene, or multiple genes that are highly homologous, and these can be scattered along the length of the genome.

Recently, Sankoff (1999) has proposed a method for estimating which of the multiple copies of a gene in two genomes G and H best reflects the position of the ancestral gene in the common ancestor genome of G and H. The basic idea is that the direct descendent of a gene (called the *true exemplar*) will be marginally less affected by genome rearrangements than the duplicates. The reduced genomes containing only the true exemplars will therefore be less arranged with respect to each other than any other pair of reduced genomes.

The problem then becomes one of selecting genes from gene families such that the distance between the resulting reduced genomes is minimized. This is called the *exemplar distance*. Sankoff formulates two versions of the problem—one based on the signed reversals distance between two gene orders, and the other based on the breakpoint distance. He provides branch and bound algorithms for both versions.

*L.I.R.M.M., 161 rue Ada, Montpellier 34392, Cedex 5, France `bryant@lirmm.fr`

In this paper, we show that both of the exemplar distance problems posed by Sankoff (1999) are NP-hard, even with quite restrictive conditions on the input data.

2 Definitions

We will use the same notation as Sankoff (1999). Given an alphabet $\mathcal{A}$, let G and H be two strings (**genomes**) of signed (+ or -) symbols (representing **genes**) from $\mathcal{A}$, of lengths l_G and l_H, respectively. For each $a \in \mathcal{A}$, let $k_X(a)$ be the number of occurrences (+ or -) of symbol a in genome X. Without loss of generality, we may assume for all $a \in \mathcal{A}$, $k_G(a) > 0$ and $k_H(a) > 0$. All occurrences of the symbol a in both genomes are said to constitute a **gene family**, the "a family". For our purposes, that the genes in a family are not exact copies is immaterial; we simply assume that the families have been constructed correctly.

A gene is a **singleton** in a genome if it is the only member of its family in that genome. A genome is **pegged** if every pair of genes from the same gene family is separated by at least one singleton.

For each genome, an **exemplar** string is constructed by deleting all but one occurrence of each gene family. Call these g and h, respectively. Note that h is just a permutation of the symbols in g. The singletons in a genome G will be in the same relative order in all exemplar strings for G.

Consider two exemplar strings $g = g_1 \dots g_n$ and $h = h_1 \dots h_n$. Note that $n = |\mathcal{A}|$. We say g_i *precedes* g_{i+1} in g. If gene a precedes b in g and neither a precedes b nor $-b$ precedes $-a$ in h, they determine a **breakpoint** in g. Additional breakpoints are posited if $g_1 \neq h_1$ and if $g_n \neq h_n$. The **breakpoint distance** (BD) is the number of breakpoints in g, which is clearly equal to the number of breakpoints in h. The **exemplar breakpoint distance** (EBD) between G and H is the minimum, over all choices of exemplar strings g and h, of the breakpoint distance between g and h.

A **reversal** transforms a string $\cdots xa \cdots by \cdots$ to $\cdots x - b \cdots - ay \cdots$. The reversals distance (RD) between g and h is the minimum number of reversals necessary to transform g into h, or vice-versa. The **exemplar reversals distance** (ERD) between G and H is the minimum, over all choices of exemplar strings g and h, of the reversals distance between g and h.

Example: Let G = -*b* -*a* *b* *a* -*c* *d* *c*, H = *a* -*a* *c* *a* -*c* *b* *d*. Based on the exemplar strings -*b* -*a* -*c* *d* and *c* *a* *b* *d*, the EBD equals 2 and the ERD equals 1.

3 Calculation of EBD and EBD

Theorem 1 *The calculation of either the EBD or the ERD between two pegged genomes G and H is an NP-hard problem, even when $k_G(a) \leq 2$ and $k_H(a) \leq 2$ for all $a \in \mathcal{A}$.*

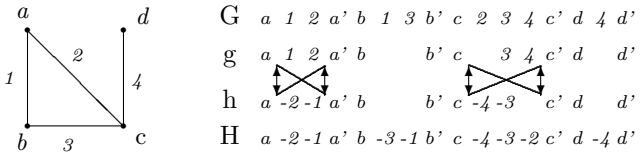

Figure 1: An example of the reduction from VERTEX COVER. On the left a graph $\mathcal{G}$ with four vertices and four edges. On the right we have the genomes G and H, and the exemplar strings corresponding to the vertex cover $\{a, c\}$. We represent the breakpoints by vertical arrows and the two reversals required by dotted lines.

Proof
We provide a reduction from VERTEX COVER:

VERTEX COVER
Instance: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Integer λ.
Question: Is there $\mathcal{V}' \subseteq \mathcal{V}$ such that $|\mathcal{V}'| = \lambda$ and each edge in $\mathcal{E}$ is adjacent to at least one vertex in $\mathcal{V}'$.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and λ make up an arbitrary instance of VERTEX COVER with $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ and $\mathcal{E} = \{e_1, e_2, \ldots, e_m\}$. We construct an alphabet $\mathcal{A}$ of size $2n + m$ given by

$$\mathcal{A} = \mathcal{V} \cup \{v_i' : v_i \in \mathcal{V}\} \cup \mathcal{E}$$

For each $i = 1, \ldots, n$ let $\mathcal{E}_i$ be a string of the symbols e_j for the edges that are adjacent to v_i. Let $-\mathcal{E}_i$ denote the string $\mathcal{E}_i$ with order reversed and opposite signs—the signed reversal of $\mathcal{E}_i$.

Let

$$G = v_1\ \mathcal{E}_1\ v_1'\ v_2\ \mathcal{E}_2\ v_2' \ldots v_n\ \mathcal{E}_n\ v_n'$$

and

$$H = v_1\ -\mathcal{E}_1\ v_1'\ v_2\ -\mathcal{E}_2\ v_2' \ldots v_n\ -\mathcal{E}_n\ v_n'.$$

In figure 1 we give the genomes G and H for a simple graph $\mathcal{G}$ with four vertices and edges.

We claim that

(1) $\mathcal{G}$ has a vertex cover of size λ if and only if the EBD between G and H is at most 2λ.

(2) $\mathcal{G}$ has an vertex cover of size λ if and only if the ERD between G and H is at most λ.

Let $\mathcal{V}' \subseteq \mathcal{V}$ be a vertex cover for $\mathcal{G}$ of size λ. The only non-singletons in G are the symbols e_j, which appear in two places. We remove all copies of symbols

e_j in substrings $\mathcal{E}_i$ such that $v_i \notin \mathcal{V}'$. At least one copy of each e_j remains, since $\mathcal{V}'$ is a vertex cover. We now select an arbitrary exemplar string, giving an exemplar string g for G of the form

$$g = v_1\ \mathcal{E}'_1\ v'_1\ v_2\ \mathcal{E}'_2\ v'_2 \ldots v_n\ \mathcal{E}'_n\ v'_n$$

where each $\mathcal{E}'_i$ is a substring of $\mathcal{E}_i$ and $\mathcal{E}'_i$ equals the empty string for all i such that $v_i \notin \mathcal{V}'$.

For each i let $-\mathcal{E}'_i$ be the signed reversal of $\mathcal{E}'_i$. Put

$$h = v_1\ -\mathcal{E}'_1\ v'_1\ v_2\ -\mathcal{E}'_2\ v'_2 \ldots v_n\ -\mathcal{E}'_n\ v'_n.$$

Then h is an exemplar string for H. See figure 1 for a simple example.

Each breakpoint in g with respect to h is of the form $v_i\ x$ for some x and $v_i \in \mathcal{V}'$, or of the form $y v'_i$ for some y and $v_i \in \mathcal{V}'$. Hence the breakpoint distance between g and h is at most $2|\mathcal{V}'| = 2\lambda$. Furthermore, we can obtain h from g by reversing all strings $\mathcal{E}_i$ such that $v_i \in \mathcal{V}'$, so the signed reversals distance between g and h is at most $|\mathcal{V}'| = \lambda$.

Conversely, suppose that the EBD between G and H is at most 2λ. There are exemplar strings g and h of G and H that are breakpoint distance at most 2λ. Put

$$\mathcal{V}' = \{v_i :\ v_i \text{ is not adjacent to } v'_i \text{ in } g\ \}$$

which is a vertex cover for $\mathcal{G}$.

There is a breakpoint between v_i and its successor and a breakpoint between v'_i and its predecessor for each $v_i \in \mathcal{V}'$, so the number of breakpoints between g and h is at least $2|\mathcal{V}'|$. Hence $|\mathcal{V}'| \leq \lambda$, and $\mathcal{G}$ has a vertex cover of size λ.

Now suppose that the ERD between G and H is at most λ. Then there are exemplar strings g and h of G and H that have signed reversals distance at most λ. Waterson *et al.* (1982) prove that the breakpoint distance between g and h is at most 2λ. The result then follows from the EBD case. □

In the case of the EBD, the complexity result can be strengthened by modifying the construction.

Lemma 2 *The calculation of the EBD between two pegged genomes is NP-hard even when $k_G(a) = 1$ and $k_H(a) \leq 2$ for all $a \in \mathcal{A}$.*

Proof
Once again, let $\mathcal{G}$ be an arbitrary graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. We augment the alphabet $\mathcal{A}$ with $m+1$ new elements $x_1, x_2, \ldots, x_{m+1}$. Construct the strings $\mathcal{E}_i$ as before. The two genomes now become

$$G = v_1 v'_1 v_2 v'_2 \ldots v_n v'_n x_1 - e_1 x_2 - e_2 x_3 \ldots x_m - e_m x_{m+1}$$

and

$$H = v_n \mathcal{E}_n v'_n v_{n-1} \mathcal{E}_{n-1} v'_{n-1} \ldots v_1 \mathcal{E}_1 v'_1 x_{m+1} x_m \ldots x_2 x_1$$

(see figure 2). Observe that G contains only singletons, so the only possible exemplar string for G is G itself. Secondly, for any exemplar string h of H the only possible adjacencies between h and G are of the form $v_i v_i'$, and the number of these adjacencies equals the number of strings $\mathcal{E}_i$ which are completely removed when selecting h (see figure 2). By the similar argument to before we have that there is a vertex cover of size $n - \lambda$ for $\mathcal{G}$ if and only if the EBD between G and H is at most $2n + 2m - \lambda$. □

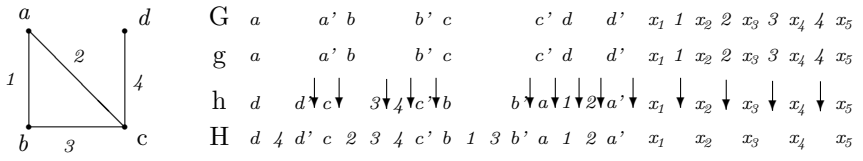

Figure 2: An example of the second reduction from VERTEX COVER. On the left a graph $\mathcal{G}$ with four vertices and four edges. On the right we have the genomes G and H, and the exemplar strings corresponding to the vertex cover $\{a, c\}$. The breakpoints of h with respect to g are marked by arrows.

References

Sankoff, D. (1999) Genome rearrangements with gene families. *Bioinformatics*, 15. In press.

Watterson, G.A., Ewens, W.J., Hall, T.E. and Morgan, A. (1982) The chromosome inversion problem. *Journal of Theoretical Biology*, **99**, 1-7.

Note added much later...

A far easier proof for the hardness of ERD is to construct a reduction from UNSIGNED REVERSAL DISTANCE. Given two unsigned genomes, replace each unsigned gene with the two signed genes, adjacent to each other. The ERD is now the unsigned reversals distance, which was shown to be hard by Caprara. However it does not seem easy to extend this approach to EBD.