THE SYNTENIC DIAMETER OF THE SPACE OF N-CHROMOSOME GENOMES

Jon Kleinberg

David Liben-Nowell

A number of distance measures have recently been proposed for the purpose of determining evolutionary similarity among genomes of different species. For each of these measures, a natural but often difficult problem is to determine the *diameter* of the space it defines: What is the maximum distance between any pair of genomes? In this work we study the *syntenic distance* between genomes, introduced by Ferretti, Nadeau, and Sankoff as a way to approximate evolutionary distance between species for which the gene order within chromosomes is not necessarily known. We show that the diameter of the space of n-chromosome genomes, with respect to the syntenic distance, is exactly $2n-4$. The proof of this result is based on a surprising connection between genome rearrangements and the study of *gossip problems* in communication networks.

1 Introduction

Distances between Genomes. The availability of large volumes of genomic data, across many species, has given rise to a class of computational problems centered around genome comparison. Such problems are motivated by a number of issues. Studying relationships among species at the chromosomal level can provide a highly effective way to infer evolutionary history; and the detailed chromosome-level information being collected for certain "model" species can be used, in conjunction with accurate methods for genome comparison, to gain further insight into genetic properties of related species.

Much of this type of comparison work is based on defining an appropriate *distance function* on pairs of genomes; genomes at smaller distances can then be presumed to have diverged more recently in evolutionary history. A number of distance functions have been proposed, according to the following "transformation-based" paradigm. One considers the space of all possible genomes, and a collection of basic *transformations* that act on this space; each such transformation models a single mutational step that changes one genome into another. The distance between two genomes $\mathcal{G}$ and $\mathcal{G}'$ is then the minimum length of a sequence of basic transformations that converts $\mathcal{G}$ into $\mathcal{G}'$. Such a minimum-length sequence can be used to derive a hypothesized sequence of events in evolutionary history relating $\mathcal{G}$ to $\mathcal{G}'$.

This notion can be seen in the formulation of biological sequence alignment (see e.g. [13, 27]); here, the similarity between two strings is defined based on their *edit distance*, the minimum (weighted) number of insertions, deletions, and substitutions needed to convert one string into the other. However, evolution at the full genome level involves not just insertions and deletions of material; the *rearrangement* of genes in the linear order of a chromosome is a crucial type of mutation, as is the *exchange* of genes between different chromosomes through mechanisms such as reciprocal translocation. These large-scale chromosomal events are rare, but they play a significant role in our reconstruction of evolutionary history among distantly related species [21, 28].

Consequently, beginning with research of Sankoff and his co-workers [23, 24], a number of genome distance functions have been studied in which the basic transformations involve global rearrangements of genes. (See also the survey by Pevzner and Waterman [22].) These functions have provided good approximations to the extent of evolutionary divergence between species, but they have also proved much harder to compute than the basic edit distance — indeed, many have been shown to be NP-complete. As a result, research on this subject has focused on heuristics and approximation algorithms with provable guarantees, as well as combinatorial investigations aimed at gaining more insight into the properties and structure of these distance functions.

Diameter Problems. One of the most basic structural questions associated with any of these distance functions is the *diameter problem*: what is the maximum distance between any pair of genomes? A recurring phenomenon — connected closely with the computational intractability of these functions — is that the associated diameter problems have turned out to be very difficult to resolve, and opened connections with extremal graph theory and combinatorial group theory.

Diameter problems were investigated first for transformation-based distance functions defined on single-chromosome genomes. The most common abstract model here is to consider species with a fixed set of genes labeled $1, 2, 3, \ldots, N$; thus, the genome of a species in this model can be described by the order, or permutation, of these genes on the chromosome. Basic transformations then act by rearranging one permutation into another. We briefly mention three distance functions based on this paradigm — each motivated by a common type of genome rearrangement mechanism — and indicate what is known about their respective diameter problems.

- One obtains the *reversal distance* between permutations by considering the set of all transformations that reverse a contiguous block of genes [1, 5, 6, 19]. Gollan conjectured that the diameter of the space of N-element permutations under this distance function is $N - 1$, and this was proved by Bafna and

Pevzner [1].

- One obtains the *prefix reversal distance* by considering the (smaller) set of all transformations that reverse a prefix of the permutation. The diameter of the space of N-element permutations under this distance function remains an open question; however, Gates and Papadimitriou showed that it lies between $\frac{17}{16}N$ and $\frac{5}{3}N + O(1)$ [12].

- Finally, one obtains the *transposition distance* by considering the set of all transformations that "splice out" a contiguous block of genes and "re-insert" it (in the same order) elsewhere in the sequence. The diameter problem for this distance function is also an open question; Bafna and Pevzner showed that the diameter of the space of N-element permutations under this distance lies between $\frac{1}{2}N$ and $\frac{3}{4}N$ [2].

The Syntenic Distance. The problems above are expressed in terms of genes in a single linear order — in other words, belonging to a single chromosome. When we consider rearrangement scenarios for genomes consisting of multiple chromosomes, we must take into account basic transformations that operate in this more complex setting: *fissions*, in which a single chromosome splits into two pieces; *fusions*, in which two chromosomes merge into a single one; and *translocations*, in which two chromosomes exchange contiguous blocks (generally, prefixes or suffixes) of their genes. Distance functions arising from this type of model were studied by Kececioglu and Ravi [18] and Hannenhalli and Pevzner [15, 16].

This type of analysis can only be performed on the genomes of species for which relatively detailed maps are available; for a larger number of species, we may only have information about which genes belong to which chromosomes, but not about their actual *order* on the chromosomes. Also, it does not necessarily make sense to treat all of the basic transformations (reversals, transpositions, translocations, fissions, and fusions) as equally "costly" in computing a distance function [10, 25].

Motivated by these types of considerations, Ferretti, Nadeau, and Sankoff proposed a more abstract measure of genomic distance, known as *syntenic distance* [11]. This model assumes no knowledge of the order of the genes within chromosomes, treating each chromosome as an unordered set of genes. Thus a genome becomes simply a collection of n sets — the synteny sets of the chromosomes — each a subset of an underlying set $\{1, 2, \ldots, N\}$ of genes. We allow a gene to appear in more than one set, and we allow a genome to contain multiple copies of the same set. We define the *support* of a genome to be the collection of genes it contains; in other words, the support is the union of all its synteny sets. The basic transformations underpinning the model are

- *fissions*, in which one set A splits into two sets B and C so that $B \cup C = A$;

- *fusions*, in which two sets B and C merge into a single set $A = B \cup C$; and

- *translocations*, in which two sets A and B exchange arbitrary subsets of their genes, yielding new sets A' and B' with the property that $A' \cup B' = A \cup B$.

In a translocation, we require that A' and B' both be non-empty; otherwise, the operation is simply a fusion.

In the style of previous models, the *syntenic distance* between two genomes with the same support is then the minimum number of these basic transformations needed to convert one into the other. We will denote this quantity by $d(\mathcal{G}, \mathcal{G}')$, for a pair of genomes $\mathcal{G}, \mathcal{G}'$. This definition can easily be extended to pairs of genomes with unequal supports, by simply focusing on the genes that are common to the two: if $\mathcal{G}$ has support U, and $\mathcal{G}'$ has support U', we first remove all genes other than those in $U \cap U'$, and then compute the syntenic distance on the resulting pair of genomes.

The Present Work. In this paper, we settle the diameter problem for the syntenic distance over n-chromosome genomes. Note that because we allow for an arbitrary number N of genes, the space of all possible n-chromosome genomes is in fact infinite; so it is not immediately apparent that the diameter of this space should be finite. However, it is not difficult to show that the diameter can be bounded by a function of n, as we will see in the next section. Indeed, in previous work, the second author [20] established an upper bound on the diameter, showing that the maximum distance between two n-chromosome genomes is at most $2n-3$. It is not difficult to show that this bound is tight for $n = 2, 3$. Subsequently, however, Cormode and Paterson [8] observed that for $n \geq 4$, this upper bound could be reduced by 1, to $2n - 4$.

Here we show that this latter bound is tight.

(1.1) *The diameter of the set of n-chromosome genomes, under the syntenic distance, is $2n - 4$ for each $n \geq 4$.*

We prove this result by showing that there exists a pair of n-chromosome genomes for which it requires $2n - 4$ basic transformations to convert one into the other. In the process, we obtain a new proof that $2n - 4$ is an upper bound as well.

Our proof of (1.1) is based on a surprising connection between the syntenic distance and the study of *gossip problems* in communication networks. Gossip problems are concerned with the flow of information among individuals communicating in a network, and their investigation has grown into a sub-field of combinatorics [17]. We find that the flow of discrete pieces of information in this setting can be related, in a precise sense, to the passage of genes through synteny sets. We are thus able to exploit certain non-trivial lower bound results for gossip problems in order to provide a lower bound for the syntenic diameter.

The remainder of the paper is organized as follows. We review some technical background on the syntenic distance in Sections 2 and 3, and indicate how to obtain an upper bound of $2n - 4$ for the syntenic diameter. In Section 4, we develop a connection between the diameter and certain move sequences that consist entirely of translocations. This allows us to develop connections to a particular gossip problem that we introduce in Section 5, and this in turn leads to a proof of (1.1) in Section 6. Finally, we survey some further directions in Section 7.

2 Preliminaries

We begin by reviewing an equivalent form of the syntenic distance problem that is easier to work with. Suppose we have an instance of the problem specified by genomes $\mathcal{G}_1 = \{S_1, \ldots, S_k\}$ and $\mathcal{G}_2 = \{T_1, \ldots, T_n\}$. (We note that $\mathcal{G}_1$ and $\mathcal{G}_2$ may be multisets, in that we may have $S_i = S_j$ or $T_i = T_j$ for certain pairs of indices i, j.) As above, we may assume that $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same support. Now, the *compact representation* of the instance, as defined in [9, 11], is obtained as follows: for each synteny set $S_i \in \mathcal{G}_1$, and each gene $g \in S_i$, we replace g by the indices of the synteny sets of $\mathcal{G}_2$ in which it appears. That is, the i^{th} synteny set of $\mathcal{G}_1$ becomes $S'_i = \cup_{g \in S_i} \{j : g \in T_j\}$.

Thus, in the compact representation, $\mathcal{G}_1$ has been replaced by the genome $\mathcal{G}'_1 = \{S'_1, \ldots, S'_k\}$. Now, if we define $\mathcal{G}'_2$ to be the collection of singleton synteny sets $\{\{1\}, \{2\}, \ldots, \{n\}\}$, it is not difficult to show the following [9, 11].

(2.1) *[9, 11]* $d(\mathcal{G}_1, \mathcal{G}_2) = d(\mathcal{G}'_1, \mathcal{G}'_2)$

Hence the compact representation is truly an equivalent instance for purposes of computing the syntenic distance.

The advantage of working with the compact representation is that it allows us to consider a more "uniform" target genome $\mathcal{G}'_2$; and, crucially, it says we may assume without loss of generality that the underlying set of genes is $\{1, 2, \ldots, n\}$, the same as the number of chromosomes in the second genome. As an example of the compact representation, consider the following instance:

$$\begin{array}{lcll} \mathcal{G}_1 & = & \{a, b\}, & \text{(Chromosome 1)} \\ & & \{c, d, e\}, & \text{(Chromosome 2)} \\ & & \{f, g\}, & \text{(Chromosome 3)} \\ & & \{h, i\} & \text{(Chromosome 4)} \end{array} \qquad \begin{array}{lcll} \mathcal{G}_2 & = & \{a, c, d\}, & \text{(Chromosome 1)} \\ & & \{b, e, f, g, h\}, & \text{(Chromosome 2).} \\ & & \{i\} & \text{(Chromosome 3).} \end{array}$$

In the compact representation, we are trying to transform the collection of sets $\{1, 2\}, \{1, 2\}, \{2\}, \{2, 3\}$ into the collection of sets $\{1\}, \{2\}, \{3\}$. One possible solution to this instance is depicted in Figure 1.

Here are two additional properties of the syntenic distance; the first is due to Dasgupta et al. [9].

(2.2) *[9] For any pair of genomes $\mathcal{G}_1$ and $\mathcal{G}_2$, there is a minimum-length sequence of transformations converting $\mathcal{G}_1$ into $\mathcal{G}_2$ that has the following special form: it consists of a sequence of fusions, followed by a sequence of translocations, followed by a sequence of fissions.*

The second fact is from [20] and establishes a monotonicity property of the distance function — enlarging the synteny sets in a genome cannot reduce the syntenic distance.

(2.3) *[20] Consider two genomes $\mathcal{G}_1 = \{S_1, \ldots, S_k\}$ and $\mathcal{G}'_1 = \{S'_1, \ldots, S'_k\}$ with the property that $S_i \subseteq S'_i$ for each i. Define $\mathcal{G}'_2 = \{\{1\}, \{2\}, \ldots, \{n\}\}$. Then $d(\mathcal{G}_1, \mathcal{G}'_2) \leq d(\mathcal{G}'_1, \mathcal{G}'_2\}$.*

$$\begin{array}{llll@{\qquad}l}
\{1,2\} & \{1,2\} & \{2\} & \{2,3\} & \\
 & & & & \text{fusion} \\
\{1,2\} & \{1,2\} & \{2,\,3\} & & \\
 & & & & \text{transloc.} \\
\{1\} & \{2\} & \{2,\,3\} & & \\
 & & & & \text{transloc.} \\
\{1\} & \{2\} & \{3\} & &
\end{array}$$

Figure 1: Transforming $\{1,2\},\{1,2\},\{2\},\{2,3\}$ into $\{1\},\{2\},\{3\}$.

3 An Extremal Pair of Genomes

We have now seen that in order to study the syntenic diameter of the space of n-chromosome genomes, we can restrict our attention to instances in the compact representation; and this immediately shows that the diameter is finite. We will use $\Delta(n)$ to denote this quantity.

Now, as observed by the second author in [20], the monotonicity theorem (2.3) tells us a specific pair of genomes on which the diameter is actually attained. Let us define $\mathcal{G}^*_{k,n}$ to be the genome consisting of k copies of the set $\{1,2,\ldots,n\}$, and define $\overline{\mathcal{G}}_n = \{\{1\},\{2\},\ldots,\{n\}\}$. Then by (2.3), $d(\mathcal{G}^*_{n,n},\overline{\mathcal{G}}_n) \geq d(\mathcal{G},\overline{\mathcal{G}})$ for every n-chromosome genome $\mathcal{G}$, and so we have

(3.1) *[20]* $\Delta(n) = d(\mathcal{G}^*_{n,n},\overline{\mathcal{G}}_n)$.

To obtain an upper bound on this quantity, we need only describe a short sequence of transformations converting $\mathcal{G}^*_{n,n}$ into $\overline{\mathcal{G}}_n$. Let us review what can be said about such upper bounds. One transformation sequence is to apply $n-2$ fusions, obtaining the genome $\mathcal{G}^*_{2,n}$; a single translocation then produces $\{\{1\},\{2,3,4,\ldots,n\}\}$; and then $n-2$ fissions produce $\overline{\mathcal{G}}_n$. This is a total of $2n-3$ transformations.

However, it is possible to convert $\mathcal{G}^*_{4,4}$ to $\overline{\mathcal{G}}_4$ using only four transformations, as described by Christie [7]: with two translocations we obtain the genome $\{\{1,2\},\{1,2\},\{3,4\},\{3,4\}\}$, and with two more we obtain $\overline{\mathcal{G}}_4$. Cormode and Paterson observed that for any $n \geq 4$, one can use this to improve the upper bound on $d(\mathcal{G}^*_{n,n},\overline{\mathcal{G}}_n)$ to $2n-4$ [8]. Specifically, one first applies $n-4$ fusions, obtaining $\mathcal{G}^*_{4,n}$; with four translocations, one can then obtain $\{\{1\},\{2\},\{3\},\{4,5,\ldots,n\}\}$; and finally completing $n-4$ fissions yields $\overline{\mathcal{G}}_n$. Thus we have

(3.2) *[8]* $d(\mathcal{G}^*_{n,n},\overline{\mathcal{G}}_n) \leq 2n-4$ *for* $n \geq 4$.

We now find ourselves in a situation that is common across many diameter problems: one has a natural upper bound on the diameter, and a pair of genomes believed (in this case, *known*) to achieve the diameter. What remains is the combinatorially non-trivial task of producing a lower bound on the distance between this pair of genomes — proving that a conjectured number of transformations is indeed *required* to convert one into the other. We turn to this matter now.

4 A Reduction to the Case of Translocations

The crucial issue in proving the lower bound is dealing with the effect of translocations, and to study this we define the following quantity. Let $\chi(n)$ denote the minimum number of *translocations* needed to convert $\mathcal{G}^*_{n,n}$ into $\overline{\mathcal{G}}_n$; that is, we restrict our attention to sequences in which fissions and fusions are disallowed.

We begin by observing that such a sequence need not be longer than the bound of (3.2). Note that this provides an alternate proof of (3.2), by a different sequence of transformations.

(4.1) *$\chi(n) \leq 2n - 4$ for $n \geq 4$.*

Proof. We prove this by induction on n, constructing recursively a sequence of $2n-4$ translocations. The case $n = 4$ has been considered above. For a larger value of n, we write $\mathcal{G}^*_{n,n} = \{S_1, S_2, \ldots, S_n\}$, with each $S_i = \{1, 2, \ldots, n\}$. We first perform a translocation on S_1 and S_2, transforming S_1 to $\{1\}$ and leaving S_2 unchanged. Now we recursively perform a sequence of translocations on $\{S_2, S_3, \ldots, S_n\}$ in which we treat elements 1 and 2 as "glued" together; in this way, we can view $\{S_2, S_3, \ldots, S_n\}$ as a copy of $\mathcal{G}^*_{n-1,n-1}$ and apply our inductive solution for this case. Thus we obtain the sets $\{\{1, 2\}, \{3\}, \{4\}, \ldots, \{n\}\}$ in $2(n - 1) - 4 = 2n - 6$ translocations. Finally, we perform a translocation of $S_1 = \{1\}$ with the copy of the set $\{1, 2\}$, arriving at $\overline{\mathcal{G}}_n$. ■

The following technical lemma will be useful in the proof of (4.3); it identifies a family of instances that behave essentially identically to the instance $(\mathcal{G}^*_{k,k}, \overline{\mathcal{G}}_k)$.

(4.2) *Let $S_1, S_2, \ldots, S_k$ form a partition of the set $\{1, 2, \ldots, n\}$, and let $\mathcal{G} = \{S_1, \ldots, S_k\}$. Then the minimum length of a sequence of translocations converting $\mathcal{G}^*_{k,n}$ into $\mathcal{G}$ is equal to $\chi(k)$.*

Proof. The proof is based on the observation that the pair of genomes $(\mathcal{G}^*_{k,k}, \overline{\mathcal{G}}_k)$ is the compact representation of $(\mathcal{G}^*_{k,n}, \mathcal{G})$. We cannot directly invoke (2.1), however, since we are considering optimal sequences with the added constraint that only translocations are allowed. Nevertheless, the proof is a straightforward application of the idea that underlies (2.1).

Given a sequence of translocations converting $\mathcal{G}^*_{k,k}$ into $\overline{\mathcal{G}}_k$, we can produce a sequence of translocations of the same length converting $\mathcal{G}^*_{k,n}$ into $\mathcal{G}$ by treating the elements of each set S_i as "glued" together throughout the process. Conversely, given a sequence of translocations converting $\mathcal{G}^*_{k,n}$ into $\mathcal{G}$, we can modify it if necessary to maintain the invariant that at any point in the process, any set containing an element $g \in S_i$ contains all the elements of S_i. ¿From a sequence of translocations that obey this invariant, we can directly obtain a sequence of translocations of the same length converting $\mathcal{G}^*_{k,k}$ into $\overline{\mathcal{G}}_k$. ■

We can now establish a first relationship between the quantities $\Delta(\cdot)$ and $\chi(\cdot)$.

(4.3) $\Delta(n) = \min\limits_{1 \leq k \leq n} \left[2(n-k) + \chi(k)\right].$

Proof. By (3.1), $\Delta(n)$ is the length of an optimal sequence converting $\mathcal{G}^*_{n,n}$ to $\overline{\mathcal{G}}_n$; and by (2.2), we may assume it has the following form. For a number k, it begins with a sequence of $n-k$ fusions, resulting in the genome $\mathcal{G}^*_{k,n}$. It then performs a sequence of translocations. Since each translocation takes two non-empty sets and produces two non-empty sets; the end result of these translocations is a genome with k sets, $\mathcal{G}' = \{S_1, S_2, \ldots, S_k\}$. Finally, the optimal sequence performs a number of fissions.

Since each fission increases the number of sets by one, and since we end with $\overline{\mathcal{G}}_n$, we know the following: $\mathcal{G}'$ must in fact be a partition of $\{1, 2, \ldots, n\}$; and the optimal sequence concludes with $n-k$ fissions. By (4.2), the portion of the sequence that converts $\mathcal{G}^*_{k,n}$ to $\mathcal{G}'$ involves at least $\chi(k)$ transformations, and thus we have $\Delta(n) \geq \min\limits_{1 \leq k \leq n} \left[2(n-k) + \chi(k)\right].$

For any k, we can convert $\mathcal{G}^*_{n,n}$ to $\overline{\mathcal{G}}_n$ by performing $n-k$ fusions, $\chi(k)$ translocations, and $n-k$ fissions; hence $\Delta(n) \leq \min\limits_{1 \leq k \leq n} \left[2(n-k) + \chi(k)\right].$ ■

Via (4.3), we can reduce the problem of finding a lower bound on $\Delta(n)$ to that of finding a lower bound on $\chi(n)$.

5 A Gossip Problem

We will see that converting $\mathcal{G}^*_{n,n}$ to $\overline{\mathcal{G}}_n$ via translocations is closely connected to the following combinatorial gossip problem. Suppose there are n people, each of whom initially knows a distinct piece of gossip. They then proceed to call each other on the telephone; each time two people talk on the phone, they exchange all the gossip they know at that point in time. The question is: what is the minimum total number of phone calls needed in order for everyone to learn all the pieces of gossip? Let us denote this number by $\Gamma(n)$.

As an example, consider the case of $n = 4$ people, named P_1, P_2, P_3, and P_4. Here is a sequence of four phone calls (written as ordered pairs) by which everyone learns all the gossip: $(P_1, P_2), (P_3, P_4), (P_1, P_3), (P_2, P_4)$. Moreover, one can show that four calls are necessary. If only three calls are made, then some person P_i is involved in at most one phone call, to P_j. For P_i to end up knowing all the gossip, P_j must have learned the rest of gossip before this call, which means that two calls must have preceded (P_i, P_j). But this means that (P_i, P_j) is the last call, hence no one but P_i and P_j learn P_i's gossip. Thus $\Gamma(4) = 4$.

The function $\Gamma(n)$ was studied by a number of researchers in combinatorics, and several distinct proofs were found for the following fundamental result [3, 4, 14, 26].

(5.1) *[3, 4, 14, 26]* $\Gamma(n) = 2n - 4$ *for each* $n \geq 4$*; that is, a total of* $2n - 4$ *calls are necessary and sufficient for everyone to learn all pieces of gossip.*

The form of the bound in (5.1) is clearly suggestive of some connection between gossip and synteny. In the following section, we make this connection precise, and complete the analysis of the syntenic diameter.

6 The Syntenic Diameter

The relationship between gossip and synteny has the following intuitive motivation. When we convert $\mathcal{G}^*_{n,n}$ to $\overline{\mathcal{G}}_n$ by a sequence of translocations, we begin with n copies of the set $\{1, 2, \ldots, n\}$ and try to "reduce" them to copies of the n singleton sets. In each step, we can pick two sets, merge them together, and then partition this merged set into two smaller sets — this is simply the definition of a translocation.

On the other hand, consider n people exchanging gossip through phone calls. In the beginning, person i knows only the piece of gossip i; the goal is to reach a state in which each person knows all the pieces of gossip $\{1, 2, \ldots, n\}$. In each step, we pick two people, take the sets representing what they currently know, and replace each with the union of these two sets. Viewed this way, we see that an optimal sequence of phone calls looks much like an optimal sequence of translocations run in reverse.

We use this idea to prove the following.

(6.1) $\chi(n) = 2n - 4$ *for each* $n \geq 4$.

Proof. Let $b \leq 2n-5$. We suppose that there exists a sequence of b translocations converting $\mathcal{G}^*_{n,n}$ into $\overline{\mathcal{G}}_n$, and construct a way for n people to fully exchange all their pieces of gossip in b phone calls — contradicting (5.1).

We introduce the following notation to describe the hypothesized sequence of b translocations. We begin with sets $S^0_1 = S^0_2 = \cdots = S^0_n = \{1, 2, \ldots, n\}$. After the first t translocations in the sequence (for some $t \geq 0$), we have a genome consisting of sets $S^t_1, S^t_2, \ldots, S^t_n$. The $(t+1)^{\text{st}}$ translocation involves sets $S^t_{x_{t+1}}$ and $S^t_{y_{t+1}}$ (for some indices x_{t+1} and y_{t+1}), and it produces two non-empty sets A and B for which $A \cup B = S^t_{x_{t+1}} \cup S^t_{y_{t+1}}$. We define $S^{t+1}_{x_{t+1}} = A$, $S^{t+1}_{y_{t+1}} = B$, and $S^{t+1}_i = S^t_i$ for each $i \notin \{x_{t+1}, y_{t+1}\}$. Note that the sequence of translocations ends with the genome $\overline{\mathcal{G}}_n$; thus, by labeling the sets appropriately, we can write $S^b_i = \{i\}$.

A sequence of phone calls by which individuals exchange gossip can be described as follows. Let K^t_i denote the set of all pieces of gossip that person i knows after t steps; thus, we have $K^0_i = \{i\}$ for each $i \in \{1, 2, \ldots, n\}$. In step $t+1$, suppose that people p_{t+1} and q_{t+1} speak on the phone, exchanging all the gossip they know. We define $K^{t+1}_{p_{t+1}} = K^{t+1}_{q_{t+1}} = K^t_{p_{t+1}} \cup K^t_{q_{t+1}}$, and $K^{t+1}_i = K^t_i$ for each $i \notin \{p_{t+1}, q_{t+1}\}$. Note the difference between this construction and the previous one: in the gossip problem we know exactly how two sets will be merged, while in the syntenic distance problem with translocations, two sets can be split in an arbitrary way.

We now construct a sequence of phone calls from our sequence of translocations, in the process defining the sets K_i^t. In our construction, we will maintain the following property:

$(*)$ For each $i \in \{1, 2, \ldots, n\}$ and $t \in \{0, 1, \ldots, b\}$, we have $K_i^{b-t} \supseteq S_i^t$.

We will prove this property holds by induction on $b - t$, together with our construction of the phone calls. We begin with the basis $t = b$: in this case, we have $K_i^0 = \{i\} = S_i^b$.

For the inductive step, suppose that we have defined $b - t$ phone calls, and $(*)$ holds for all i and all $t' \geq t$. Now, the t^{th} translocation in our sequence involves the sets $S_{x_t}^{t-1}$ and $S_{y_t}^{t-1}$; we define the $(b-t+1)^{\text{st}}$ phone call to take place between people x_t and y_t. Let us verify that with this additional phone call, $(*)$ now holds for $t-1$. For $i \notin \{x_t, y_t\}$, we have $K_i^{b-t+1} = K_i^{b-t} \supseteq S_i^t = S_i^{t-1}$. We also have

$$K_{x_t}^{b-t+1} = K_{x_t}^{b-t} \cup K_{y_t}^{b-t} \supseteq S_{x_t}^t \cup S_{y_t}^t \supseteq S_{x_t}^{t-1},$$

with a completely symmetric argument holding for y_t.

This completes the inductive argument. As a consequence, we have $K_i^b \supseteq S_i^0 = \{1, 2, \ldots, n\}$ for each i, establishing that the sequence of b phone calls we have constructed fully exchanges the pieces of gossip among all the individuals. ■

Finally, we have

Proof of (1.1). By (3.1) and (3.2), we know that $\Delta(n) \leq 2n - 4$. Thus we need only show the lower bound.

One can work out directly that $\chi(1) = 0$, $\chi(2) = 1$, and $\chi(3) = 3$. Thus, for all $n \geq 1$, we have $\chi(n) \geq 2n - 4$.

Now, applying (4.3), we have

$$\begin{aligned} \Delta(n) &= \min_{1 \leq k \leq n} [2(n-k) + \chi(k)] \\ &\geq \min_{1 \leq k \leq n} [2(n-k) + 2k - 4] \\ &= 2n - 4. \qquad ■ \end{aligned}$$

7 Further Directions

Given an arbitrary genome $\mathcal{G}$, determining $d(\mathcal{G}, \overline{\mathcal{G}}_n)$ is known to be NP-complete [9]. However, in the spirit of the analysis in this paper, it would be interesting to consider the extreme end of the distance scale and try characterizing those n-chromosome genomes $\mathcal{G}$ for which $d(\mathcal{G}, \overline{\mathcal{G}}_n) = 2n - 4$. As a related question, we can ask: Given genomes $\mathcal{G}_1$ and $\mathcal{G}_2$, each with n chromosomes, what is the computational complexity of determining whether $d(\mathcal{G}_1, \mathcal{G}_2) = 2n - 4$?

Approximation algorithms for the syntenic distance problem — in particular, for determining $d(\mathcal{G}, \overline{\mathcal{G}}_n)$ in the case of an arbitrary n-chromosome genome $\mathcal{G}$ — have been studied in [9, 11, 20]. Finding a polynomial-time algorithm that can approximate $d(\mathcal{G}, \overline{\mathcal{G}}_n)$ to within a factor better than 2 appears to be a difficult problem.

Determining $d(\mathcal{G}, \overline{\mathcal{G}}_n)$ for an arbitrary n-chromosome genome $\mathcal{G}$ is closely related to the following general gossip problem. There are n people, each of whom initially knows a distinct piece of gossip. Each person i is seeking to learn only a *subset* S_i of all the pieces of gossip. What is the minimum number of phone calls needed for each person i to learn all the items in the associated set S_i? It would be interesting to see if considering this formulation of the problem might suggest new types of approximation algorithms for the syntenic distance, or for close variants.

More generally, we believe it would be fruitful to look for further relationships between these two areas; any connections that can be developed may be valuable for resolving problems both in the analysis of information flow and in the study of genome rearrangements.

Acknowledgments

The work of the first author was supported in part by a David and Lucile Packard Foundation Fellowship, an Alfred P. Sloan Research Fellowship, an ONR Young Investigator Award, and NSF Faculty Early Career Development Award CCR-9701399. The second author is on leave from graduate studies at MIT and is currently studying at Cambridge University, supported by a Churchill Scholarship from the Winston Churchill Foundation. The work of this author was performed in part at Cornell University, supported by the ONR Young Investigator Award of the first author.

References

[1] Vineet Bafna and Pavel A. Pevzner. Genome rearrangements and sorting by reversals. *SIAM Journal on Computing*, 25(2):272–289, April 1996.

[2] Vineet Bafna and Pavel A. Pevzner. Sorting by transpositions. *SIAM Journal on Discrete Mathematics*, 11(2):224–240, 1998.

[3] Brenda Baker and Robert Shostak. Gossips and telephones. *Discrete Mathematics*, 2:191–193, 1972.

[4] Richard T. Bumby. A problem with telephones. *SIAM Journal on Algebraic and Discrete Methods*, 2(1):13–18, March 1981.

[5] Alberto Caprara. Sorting by reversals is difficult. In *1st Annual International Conference on Computational Molecular Biology*, pages 75–83, 1997.

[6] D. A. Christie. A 3/2-approximation algorithm for sorting by reversals. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 244–252, 1998.

[7] D. A. Christie. *Genome Rearrangement Problems*. PhD thesis, Univeristy of Glasgow, 1999.

[8] G. Cormode and M. Paterson. Personal Communication, July 1999.

[9] Bhaskar DasGupta, Tao Jiang, Sampath Kannan, Ming Li, and Elizabeth Sweedyk. On the complexity and approximation of syntenic distance. *Discrete Applied Mathematics (special issue on computational biology)*, 88(1–3):59–82, November 1998.

[10] Jason Ehrlich, David Sankoff, and Joseph H. Nadeau. Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, 147(1):289–296, September 1997.

[11] Vincent Ferretti, Joseph H. Nadeau, and David Sankoff. Original synteny. In *7th Annual Symposium on Combinatorial Pattern Matching*, pages 159–167, 1996.

[12] W. H. Gates and C. H. Papadimitriou. Bounds for sorting by prefix reversals. *Discrete Mathematics*, 27:47–57, 1979.

[13] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press, 1997.

[14] A. Hajnal, E. C. Milner, and E. Szemerédi. A cure for the telephone disease. *Canadian Mathematical Bulletin*, 15(3):447–450, 1972.

[15] S. Hannenhalli. *Transforming Mice into Men (A Computational Theory of Genome Rearrangements).* PhD thesis, Pennsylvania State University, 1995.

[16] S. Hannenhalli and P. Pevzner. Transforming mice into men (polynomial algorithm for genomic distance problem). In *Proceedings of the 36th IEEE Symposium on Foundations of Computer Science*, pages 581–592, 1995.

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

[18] J. D. Kececioglu and R. Ravi. Of mice and men: Algorithms for evolutionary distance between genomes with translocations. In *Proceedings of 6th ACM-SIAM Symposium on Discrete Algorithms*, pages 604–613, 1995.

[19] J. D. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutations. *Algorithmica*, 13:180–210, 1995.

[20] David Liben-Nowell. On the structure of syntenic distance. In *10th Annual Symposium on Combinatorial Pattern Matching*, pages 43–56, 1999.

[21] J. Nadeau and B. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. In *Proc. Natl. Acad. Sci. USA*, volume 81, pages 814–818, 1984.

[22] Pavel Pevzner and Michael Waterman. Open combinatorial problems in computational molecular biology. In *Proceedings of the Third Israel Symposium on Theory of Computing and Systems*, pages 158–173, January 1995.

[23] D. Sankoff. Edit distance for genome comparison based on non-local operations. In *3rd Annual Symposium on Combinatorial Pattern Matching*, pages 121–135, 1992.

[24] D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B. Lang, and R. Cedergren. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. In *Proc. Natl. Acad. Sci. USA*, volume 89, pages 6575–6579, 1992.

[25] David Sankoff and Joseph H. Nadeau. Conserved synteny as a measure of genomic distance. *Discrete Applied Mathematics (special issue on computational biology)*, 71(1–3):247–257, December 1996.

[26] R. Tijdeman. On a telephone problem. *Nieuw Archief voor Wiskunde*, 9(3):188–192, 1971.

[27] M. Waterman. *Introduction to Computational Biology.* Chapman-Hall, 1995.

[28] G. Watterson, W. Ewens, T. Hall, and A. Morgan. The chromosome inversion problem. *J. Theoretical Biology*, 99:1–7, 1982.

Department of Computer Science
Cornell University
Ithaca NY 14853 USA
E-mail: kleinber@cs.cornell.edu

Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge MA 02139 USA
E-mail: dln@cs.cornell.edu