Gossip is Synteny: Incomplete Gossip and the Syntenic Distance between Genomes*

David Liben-Nowell
Department of Computer Science
Carleton College
Northfield, MN 55057 USA
dlibenno@carleton.edu

Abstract

The *syntenic distance* between two genomes is given by the minimum number of fusions, fissions, and translocations required to transform one into the other, ignoring the order of genes within chromosomes. Computing this distance is NP-hard. In the present work, we give a tight connection between syntenic distance and the *incomplete gossip problem*, a novel generalization of the classical gossip problem. In this problem, there are n gossipers, each with a unique piece of initial information; they communicate by phone calls in which the two participants exchange all their information. The goal is to minimize the total number of phone calls necessary to inform each gossiper of his set of *relevant gossip* which he desires to learn.

As an application of the connection between syntenic distance and incomplete gossip, we derive an $O(2^{O(n \log n)})$ algorithm to exactly compute the syntenic distance between two genomes with at most n chromosomes each. Our algorithm requires $O(n^2 + 2^{O(d \log d)})$ time when this distance is d, improving the $O(n^2 + 2^{O(d^2)})$ running time of the best previous exact algorithm.

1 Introduction

Recently there has been considerable interest in computational models measuring the genetic distance between two species. Such models can be used in the construction of trees of evolutionary history, or—if such a tree is known through other means—in estimating the rate of genomic evolution. These measures are generally based on a hypothesized set of *transformations* that can alter a genome; the *distance* between the genomes of two species is then the minimum number of these steps necessary to transform one into the other.

In addition to local mutations like insertions, deletions, and substitutions in the DNA sequence, a realistic distance measure must account for non-local transformations that alter the placement of genes within or among chromosomes. These rearrangements may include *reversals*, which invert the order of a section of a chromosome, and *transpositions*, which extract a segment of a chromosome and reinsert it elsewhere in the chromosome. The corresponding distance measures—*reversal*

*Appears in *Journal of Algorithms*, Volume 43, Number 2, May 2002, pp. 264–283. A preliminary version of this paper appeared in *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'01)* under the title "Gossip is Synteny: Incomplete Gossip and an Exact Algorithm for Syntenic Distance." Minor changes have been made in this document; the last update was on 4 August 2005. Comments are welcome.

distance, *transposition distance*, and the combined *reversal and transposition distance*—have been explored by the theoretical computer science community [1, 2, 7, 9, 13, 14, 16, 24].

When comparing genomes containing multiple chromosomes, one must consider transformations acting between chromosomes in addition to those acting within a single chromosome. These transformations include *fissions*, in which one chromosome splits into two, *fusions*, in which two chromosomes merge into one, and *translocations*, in which two chromosomes exchange contiguous blocks (usually prefixes or suffixes) of genes.

Initial mathematical investigations of multi-chromosomal distance functions considered translocations in isolation or combined only with reversals [18, 23], and disregarded fusions and fissions. This limits the model to pairs of species with the same number of chromosomes. Later research [19] extended the model to include fusions and fissions.

1.1 Syntenic Distance

As defined, all of the above models require *gene order* data for their computation. In biological practice, this information can be difficult to obtain—we may have information about the *assignment* of genes to chromosomes, but not the order within them. Despite the recent progress in sequencing the human genome [22], for example, the genomes of a vast majority of species are mostly unanalyzed; gene order data for most species will remain unavailable throughout the foreseeable future. We would like to be able to compare genomes meaningfully even without this data; much recent biological research is devoted to the investigation of this kind of relationship—e.g., [4, 36, 37].

Furthermore, even if gene order information is available, there is some biological reason to believe that it may not be well-suited to use in genomic distance calculations. A number of researchers [29, 31, 34] have found evidence of frequent (small-scale) reversals in chromosomes; in the presence of such reversals, gene order may be poorly preserved even in closely related species, and thus any order-based metric may be a poor estimate of genomic distance.

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 [5, 12, 33]. If, e.g., reversals turn out to be extremely frequent, heavily weighting the rarer interchromosomal rearrangements should give a better estimate of distance.

Motivated by these types of observations, Ferretti, Nadeau, and Sankoff proposed a more abstract measure of genomic distance, known as *syntenic distance* [15]. (Two genes are *syntenic* if they appear in the same chromosome.) Any information about the order of genes within chromosomes is ignored; a chromosome is then simply an unordered set of genes. A genome can be transformed by fusions, fissions, and translocations—i.e., exactly those transformations that alter the assignments of genes to chromosomes. Computing the syntenic distance between species has been shown to be NP-complete, though there are efficient 2-approximations [11, 15, 27]. Other recent work has also explored some of the rich combinatorial structure of this model [25, 28, 30].

1.2 Incomplete Gossip

In the early 1970s, the following puzzle (popularized by Paul Erdös) was circulated among mathematicians. There are n gossipers, each of whom knows a unique piece of initial information. They communicate by telephone calls, and whenever two speak they share all the gossip that they know. The goal is to determine the minimum number of calls necessary for all of the participants to learn all of the initial information. A number of researchers have independently proven that $2n-4$ calls are necessary and sufficient to achieve this goal [3, 6, 17, 21, 35].

A voluminous body of work followed these initial proofs, including a wide variety of extensions and variations on this basic problem. (See [20] for a survey.) Most of this work focused on modifications to the communication model—e.g., allowing conference calls involving more than two gossipers, or placing restrictions on who can talk to whom.

In this paper, we generalize the gossip problem in a different way, by allowing gossipers to have interest in only a subset of the initial information. For each gossiper i, suppose that there is a set of *relevant gossip* S_i that he wishes to learn. In the *incomplete gossip problem*, the gossipers communicate by phone calls as before, but the goal is now to minimize the total number of calls necessary so that each gossiper i learns all of his relevant gossip S_i. We will formally introduce the incomplete gossip problem in Section 4.

Incomplete gossip generalizes a number of previous variants in the gossip literature. The complete gossip problem is simply the case in which all gossipers want to learn all information. The *broadcasting problem* is the case in which all participants only want to learn the single piece of information initially known to the *originator*. In the *set-to-set gossiping* (or *set-to-set broadcasting*) problem [26, 32], we are given two (possibly intersecting) sets A and B of gossipers, and the goal is to minimize the number of calls necessary to inform all gossipers in A of all the gossip known to the members of B. This is the special case of incomplete gossip when every $a \in A$ wishes to learn the initial information of every $b \in B$.

Other variations on the gossiping problem have some similarity to the incomplete gossip problem, but differ in the details. In the *partial gossiping problem* [8, 32], each participant wishes to learn at least $k \leq n$ pieces of gossip, but does not care which k tidbits he learns. Brief consideration has also been given to the situation in which each gossiper initially knows several pieces of gossip (not necessarily distinct from the initial information of the others) and everyone wishes to learn all the information [10].

1.3 Our Results: Relating Gossip and Synteny

In the present work, we derive and explore a tight connection between syntenic distance and incomplete gossip. Our main contribution is this unexpected technical link between genomic distance measures and problems of information flow. The connection to the incomplete gossip problem—a conceptually simpler problem—also yields increased combinatorial insight into the syntenic distance problem. We also believe that the incomplete gossip problem is interesting in its own right. Finally, as an application of this connection, we derive a new gossip-based exact algorithm for syntenic distance. Our algorithm is a significant asymptotic improvement over the best previous algorithm.

1.3.1 Similarities between gossip and synteny

One can view the syntenic distance problem between genomes $\mathcal{G}_1$ and $\mathcal{G}_2$ as follows. Let the *target* of a chromosome C in $\mathcal{G}_1$ denote the set of chromosomes of $\mathcal{G}_2$ which share a gene with C. Then the goal for the syntenic distance problem is to exchange genes among the chromosomes of $\mathcal{G}_1$ so that the target of each chromosome becomes a unique single chromosome of $\mathcal{G}_2$.

Under this target view, the analogy between incomplete gossip and syntenic distance is as follows. Gossipers and information in the gossip problem correspond, respectively, to chromosomes of $\mathcal{G}_1$ and chromosomes of $\mathcal{G}_2$ in the syntenic distance problem; a sequence of phone calls corresponds to a sequence of translocations, viewed in reversed order. For the gossip problem, we begin with n unique pieces of initial knowledge and complete a series of phone calls to spread the information. For the syntenic distance problem, we aim to complete a series of translocations to merge targets so that every chromosome of $\mathcal{G}_1$ has a unique target chromosome in $\mathcal{G}_2$. A phone call is in essence the reverse of a translocation: in a phone call, two gossipers take their information sets A and B and exchange information so that both know $A \cup B$; in a translocation, two chromosomes with targets contained in $A \cup B$ exchange genes so that they have targets A and B afterwards.

In previous work with Jon Kleinberg, we used this notion to establish a connection between the *syntenic diameter*—the syntenic distance between the two n-chromosome species maximally different under the syntenic distance model—and the (complete) gossip problem. We proved that the number of calls necessary for the complete gossip problem is exactly the syntenic diameter [25]. In the present paper, we explore the relationship between gossip and the syntenic distance on general instances, using the incomplete gossip problem. Roughly, the relevant gossip S_i for gossiper i is the target of the ith chromosome of genome $\mathcal{G}_1$; as we shall see, however, there are complications in the analogy.

For ease of exposition, we will sometimes denote as *sets* the chromosomes of $\mathcal{G}_1$ and the gossipers, and as *elements* the chromosomes of $\mathcal{G}_2$ and the pieces of gossip.

1.3.2 Differences between gossip and synteny

There are three major obstacles to the equivalence of incomplete gossip and translocation syntenic distance: (1) whether the number of sets and number of elements may differ, (2) monotonicity, and (3) whether the sets are ordered with respect to the elements. All three of these hurdles are hidden by properties of the particular instance that we considered in [25].

Implicit in the gossip formulation is that the number of gossipers is the same as the number of pieces of initial information. There is no such constraint in the syntenic distance problem, where the number of chromosomes in $\mathcal{G}_1$ and $\mathcal{G}_2$ can differ arbitrarily. Accordingly, we introduce the *translocation syntenic distance*, analogous to the translocation distance in the ordered case [18, 23], in which fusions and fissions are forbidden. This restricts us to the situation where $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same number of chromosomes. (Furthermore, it is not clear how phone calls could be analogous to fusions and fissions, since such moves would change the number of gossipers.)

Incomplete gossip is monotonic in the sense that if we add more relevant gossip for gossiper i, then we can only increase the number of phone calls required to solve the instance. However, if we increase the target for a chromosome C_i in $\mathcal{G}_1$, we cannot prove—and, in fact, believe to be false—the claim that the translocation syntenic distance will not decrease. We overcome this difference by introducing a variant on the translocation syntenic distance which enforces monotonicity. We prove that this variant is equivalent to the general syntenic distance, including fusions and fissions.

The most troublesome difference between incomplete gossip and translocation syntenic distance is that of the ordering of the sets. In the gossip problem, there is a strict relationship between the gossiper i and the piece of information i—namely, information i is exactly that piece of gossip initially known by the ith gossiper. There is no such relationship between the ith chromosome of $\mathcal{G}_1$ and the ith chromosome of $\mathcal{G}_2$: if two genomes "differ" only by the numbers assigned to each chromosome, then we want to report that the two species are identical. This difference forces us into a brute force examination of all permuted orderings of the sets.

1.3.3 An improved exact algorithm for syntenic distance

We exploit the connection between translocation syntenic distance and incomplete gossip to develop an improved exact algorithm for the general syntenic distance problem. We first handle the easier case in which translocations are the only legal transformations, and then we add fusions and fissions to the model. Consider two genomes $\mathcal{G}_1$ and $\mathcal{G}_2$ with n and k chromosomes, respectively. When the syntenic distance between $\mathcal{G}_1$ and $\mathcal{G}_2$ is d, our algorithm requires $O(nk + 2^{O(d \log d)})$ time, which improves on the $O(nk + 2^{O(d^2)})$ running time of the best previous exact algorithm, of DasGupta et al. [11]. (Recall that the syntenic distance problem is NP-complete.)

Intuitively, the speed-up in our algorithm is derived from the following. The algorithm of DasGupta et al. essentially enumerates all possible sequences of transformations of length d, and checks whether any of these sequences transform $\mathcal{G}_1$ into $\mathcal{G}_2$. The vast majority of the time spent in this algorithm is on translocations: for a translocation, we not only must select the input chromosomes, but also for each gene g in either input chromosome we must specify which output chromosome will contain g. In the gossip-based approach, we only need to select which people participate in each call. Once we have selected the participants, both learn whatever new information the other knows; there is no choice of different output sets.

2 Syntenic Distance

We first formally introduce the syntenic distance model, and mention a few of the properties that we will need in the remainder of this work. For the purposes of this paper, a *chromosome* is an unordered set of *genes*, and a *genome* is an unordered collection of chromosomes. (We limit our consideration to genomes with no duplicated genes, though we do not forbid duplication in the model; this simplifies the notation of the compact representation, defined below.) A genome can be transformed by any of the following operations:

- a *fusion* $(S, T) \longrightarrow U$, in which two chromosomes S and T merge into a single chromosome U, where $U = S \cup T$.
- a *fission* $U \longrightarrow (S, T)$, in which a chromosome U splits into two chromosomes S and T, where $U = S \cup T$.
- a *translocation* $(S, T) \longrightarrow (S', T')$, in which two chromosomes S and T exchange arbitrary subsets of their genes, producing two new chromosomes S' and T', where $S \cup T = S' \cup T'$.

We require these operations to take non-empty chromosomes as input, and produce non-empty chromosomes as output. Fissions and translocations can cause the duplication of genes; e.g., the move $(\{1\}, \{2\}) \longrightarrow (\{1, 2\}, \{2\})$ is a legal translocation.

Definition 2.1. *The* syntenic distance $d(\mathcal{G}_1, \mathcal{G}_2)$ *between two genomes* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *is the minimum number of fusions, fissions, and translocations required to transform* $\mathcal{G}_1$ *into* $\mathcal{G}_2$*, ignoring all genes that appear in only one of the two genomes.*

Suppose we have an instance of the problem specified by genomes $\mathcal{S} = S_1, \ldots, S_k$ and $\mathcal{T} = T_1, \ldots, T_n$. The *compact representation* of the instance [11, 15] is obtained as follows: for each chromosome S_i and each gene $g \in S_i$, replace g by the indices of the chromosomes of $\mathcal{T}$ in which it appears. That is, the ith chromosome of $\mathcal{S}$ is replaced by $S'_i = \bigcup_{g \in S_i} \{j : g \in T_j\}$. Thus, in the compact representation, $\mathcal{S}$ has been replaced by the genome $\mathcal{S}' = S'_1, \ldots, S'_k$. It is not difficult to show that $d(\mathcal{S}, \mathcal{T}) = d(\mathcal{S}', \mathcal{T}')$, where $\mathcal{T}' = \{1\}, \ldots, \{n\}$ [11, 15]. The compact representation allows us to limit the number of genes to n (the number of chromosomes in the second genome) while also considering a more "uniform" target genome $\mathcal{T}'$.

As an example of the compact representation, consider the following instance:

$$
\begin{array}{rcll}
\mathcal{S} & = & \{a, b\}, & \text{(Chromosome 1)} \\
 & & \{c, d, e\}, & \text{(Chromosome 2)} \\
 & & \{f, g\}, & \text{(Chromosome 3)} \\
 & & \{h, i, j\} & \text{(Chromosome 4)} \\
 & & & \\
\mathcal{T} & = & \{a, c, d\}, & \text{(Chromosome 1)} \\
 & & \{b, e, f, g, h\}, & \text{(Chromosome 2)} \\
 & & \{i, k\} & \text{(Chromosome 3).}
\end{array}
$$

In the compact representation, we wish to transform the collection of sets $\{1, 2\}$, $\{1, 2\}$, $\{2\}$, $\{2, 3\}$ into the collection $\{1\}, \{2\}, \{3\}$.

For the remainder of this paper, we will only consider instances in the compact representation. Also, for an instance $\mathcal{S} = S_1, \ldots, S_n$ where $S_i = \{\}$ for some i, we will understand $\mathcal{S}$ to denote the instance $\mathcal{S}' = S_1, \ldots, S_{i-1}, S_{i+1}, \ldots, S_n$. Thus we will presume that all sets are initially non-empty.

Definition 2.2. *Let $\mathcal{S} = S_1, \ldots, S_k$ be a collection of sets such that $\bigcup_i S_i = \{1, \ldots, n\}$. Then the* syntenic distance *of $\mathcal{S}$ is*

$$d(\mathcal{S}) \ := \ d(\mathcal{S}, \overline{\mathcal{G}}_n),$$

where $\overline{\mathcal{G}}_n = \{1\}, \ldots, \{n\}$.

We will use σ to denote a sequence of fusions, fissions, and translocations, and we say that σ *solves* $\mathcal{S}$ if it transforms $\mathcal{S}$ into $\overline{\mathcal{G}}_n$. Note that the sets of $\overline{\mathcal{G}}_n$ do not have to be produced in any particular order.

We write $\mathcal{S} \sqsubseteq \widehat{\mathcal{S}}$ for $\mathcal{S} = S_1, \ldots, S_n$ and $\widehat{\mathcal{S}} = \widehat{S}_1, \ldots, \widehat{S}_n$, if, for all i, we have $S_i \subseteq \widehat{S}_i \subseteq \bigcup_i S_i$. We will make use of the following known properties of the syntenic distance:

Theorem 2.3 (Canonicalization [11]). *For any instance $\mathcal{S}$, there is an optimal move sequence σ solving $\mathcal{S}$ in which all fusions precede all translocations precede all fissions.*

Theorem 2.4 (Syntenic Monotonicity [27]). *For any instance $\mathcal{S}$ and any $\mathcal{S}' \sqsubseteq \mathcal{S}$, we have $d(\mathcal{S}') \leq d(\mathcal{S})$.*

In addition, we will use following observation of DasGupta et al. [11]: for any instance $\mathcal{S} = S_1, \ldots, S_n$ where $\bigcup_i S_i = \{1, \ldots, n\}$, there is a move sequence that solves $\mathcal{S}$ in at most $2n-2$ moves.

3 Translocation Syntenic Distance

We begin by restricting our attention to move sequences that consist solely of translocations. We explore some of the properties of the *translocation syntenic distance*, and relate this measure and a variant to the general syntenic distance. For this restricted model, we must limit ourselves to instances in which the number of elements is equal to the number of sets.

Definition 3.1. *A collection of non-empty sets $\mathcal{S} = S_1, \ldots, S_n$ is* square *iff* $\bigcup_i S_i = \{1, \ldots, n\}$.

Translocation-only move sequences can only solve square instances; since a translocation transforms two non-empty sets into two other non-empty sets, a sequence of translocations cannot alter the number of non-empty sets, as is required for instances that are not square. Note that any move sequence solving a square instance must contain the same number of fusions and fissions—the number of non-empty sets increases by one with a fission and decreases by one with a fusion, and a square instance initially contains exactly the right number of non-empty sets.

Definition 3.2. *For square $\mathcal{S} = S_1, \ldots, S_n$, let $\chi(\mathcal{S})$ be the* translocation syntenic distance *of $\mathcal{S}$:*

$$\chi(\mathcal{S}) := \min_{\rho \text{ solves } \mathcal{S}} |\rho|$$

where ρ contains only translocations.

Just as for the syntenic distance, ρ *solves* $\mathcal{S}$ by transforming it into $\overline{\mathcal{G}}_n = \{1\}, \ldots, \{n\}$. We will use ρ to denote sequences of translocations.

Proposition 3.3. *For all square $\mathcal{S}$, $d(\mathcal{S}) \leq \chi(\mathcal{S})$.*

Proof. Any translocation move sequence is also a legal fusion-fission-translocation move sequence.□

We would like to show that fusions and fissions never help in solving a square instance. However, for example, the instance $\{1,2,3,4\}$, $\{1,2,3,4\}$, $\{1,2,3,4\}$, $\{5,6,7\}$, $\{5,6,7\}$, $\{5,6,7\}$, $\{5,6,7\}$ appears to require 8 translocations if fusions and fissions are forbidden, but can be solved with the following 7 moves (1 fusion, 5 translocations, and 1 fission):

$$\begin{array}{rcl}
(\{1,2,3,4\},\{1,2,3,4\}) & \longrightarrow & \{1,2,3,4\} \\
(\{1,2,3,4\},\{1,2,3,4\}) & \longrightarrow & (\{1\},\{2,3,4\}) \\
(\{2,3,4\},\{5,6,7\}) & \longrightarrow & (\{2\},\{3,4,5,6,7\}) \\
(\{3,4,5,6,7\},\{5,6,7\}) & \longrightarrow & (\{3\},\{4,5,6,7\}) \\
(\{4,5,6,7\},\{5,6,7\}) & \longrightarrow & (\{4\},\{5,6,7\}) \\
(\{5,6,7\},\{5,6,7\}) & \longrightarrow & (\{5\},\{6,7\}) \\
\{6,7\} & \longrightarrow & (\{6\},\{7\}).
\end{array}$$

Fortunately, the following weaker statement will suffice for our purposes:

Lemma 3.4. *For all square $\mathcal{S}$, there exists $\widehat{\mathcal{S}} \sqsupseteq \mathcal{S}$ such that $d(\mathcal{S}) \geq \chi(\widehat{\mathcal{S}})$.*

Proof. Let σ be an optimal canonical move sequence solving $\mathcal{S}$ using the fewest possible fusions, and let α be the number of fusions in σ. We proceed by induction on α.

For $\alpha = 0$, the sequence σ contains only translocations. Therefore $\chi(\mathcal{S}) \leq |\sigma| = d(\mathcal{S})$ by the optimality of σ.

For $\alpha \geq 1$, consider the last fusion $\sigma_\alpha = (A, B) \longrightarrow A \cup B$ and the first fission $C \cup D \longrightarrow (C, D)$ of σ. Note that since only other fusions precede σ_α, the set A must consist of the union of some number of original input sets. Let S_i be one such set.

We define σ' and $\mathcal{S}'$ as follows. Add $C \cup D$ to the input set S_i, and carry $C \cup D$ along with the elements of S_i until σ_α. Instead of σ_α, we complete the translocation $(A \cup C \cup D, B) \longrightarrow (A \cup B, C \cup D)$. Therefore $\sigma'_1, \ldots, \sigma'_\alpha$ applied to $\mathcal{S}'$ yields exactly the same sets as $\sigma_1, \ldots, \sigma_\alpha$ applied to $\mathcal{S}$, except that the former produces an additional set $C \cup D$.

We duplicate exactly the translocation phase of σ in σ', ignoring the presence of the extra set $C \cup D$. The result of these moves of σ' on $\mathcal{S}'$ is identical to the result of these moves of σ on $\mathcal{S}$, again except for the extra $C \cup D$. Now in place of the fission $C \cup D \longrightarrow (C, D)$, we can complete the translocation $(C \cup D, C \cup D) \longrightarrow (C, D)$. Now the resulting instances are exactly identical. Let the remainder of σ' match that of σ. Thus σ' solves $\mathcal{S}'$ and $|\sigma'| = |\sigma|$.

We have produced an instance $\mathcal{S}' \sqsupseteq \mathcal{S}$, where $\mathcal{S}'$ is solved by a move sequence σ' containing $\alpha - 1$ fusions and $|\sigma|$ total moves. Note that $|\sigma| = |\sigma'| \geq d(\mathcal{S}') \geq d(\mathcal{S}) = |\sigma|$ by syntenic monotonicity, so $d(\mathcal{S}') = d(\mathcal{S})$. By the induction hypothesis, there is an instance $\widehat{\mathcal{S}} \sqsupseteq \mathcal{S}'$ so that $\chi(\widehat{\mathcal{S}}) \leq d(\mathcal{S}')$. Therefore $\widehat{\mathcal{S}} \sqsupseteq \mathcal{S}' \sqsupseteq \mathcal{S}$ and $\chi(\widehat{\mathcal{S}}) \leq d(\mathcal{S}') = d(\mathcal{S})$. □

With a monotonicity property for translocation syntenic distance—that is, $\chi(\mathcal{S}) \leq \chi(\widehat{\mathcal{S}})$ whenever $\mathcal{S} \sqsubseteq \widehat{\mathcal{S}}$—this lemma would imply $\chi(\mathcal{S}) = d(\mathcal{S})$. Unfortunately, we are unable to prove this property, and, in fact, believe that it is false. This situation inspires the definition of a variant on translocation syntenic distance with enforced monotonicity:

Definition 3.5. *For a square instance $\mathcal{S}$ the* expanded translocation syntenic distance $\chi^*(\mathcal{S})$ *is*

$$\chi^*(\mathcal{S}) := \min_{\mathcal{S}' \sqsupseteq \mathcal{S}} \chi(\mathcal{S}').$$

Proposition 3.6. *For any square instance $\mathcal{S}$ and any $\widehat{\mathcal{S}} \sqsupseteq \mathcal{S}$, we have $\chi^*(\mathcal{S}) \leq \chi^*(\widehat{\mathcal{S}}) \leq \chi(\widehat{\mathcal{S}})$.*

Proof. Immediate from the definition of $\chi^*(\mathcal{S})$. □

We can now prove the equality of expanded translocation syntenic distance and syntenic distance.

Theorem 3.7. *If $\mathcal{S}$ is square then $\chi^*(\mathcal{S}) = d(\mathcal{S})$.*

Proof. By Proposition 3.3 and syntenic monotonicity,

$$\chi^*(\mathcal{S}) \quad = \quad \min_{\mathcal{S}' \sqsupseteq \mathcal{S}} \chi(\mathcal{S}') \quad \geq \quad \min_{\mathcal{S}' \sqsupseteq \mathcal{S}} d(\mathcal{S}') \quad = \quad d(\mathcal{S}).$$

For the other direction, let $\widehat{\mathcal{S}} \sqsupseteq \mathcal{S}$ be such that $d(\mathcal{S}) \geq \chi(\widehat{\mathcal{S}})$ according to Lemma 3.4. Then $d(\mathcal{S}) \geq \chi(\widehat{\mathcal{S}}) \geq \chi^*(\mathcal{S})$ by Proposition 3.6. □

4 Incomplete Gossip and the Translocation Syntenic Distance

Formally, the *(complete) gossip problem* is as follows. We are given a number n of gossipers, where gossiper i initially knows a unique piece of information i. Any two gossipers can communicate via a *phone call*, during which the two participants each tell the other all of the information that they know at the time. We seek a minimum-length sequence ϕ of phone calls so that every gossiper learns all the information $\{1, \ldots, n\}$.

For the *incomplete gossip problem*, we are also given sets $S_1, \ldots, S_n$, where $S_i \subseteq \{1, \ldots, n\}$ is the set of *relevant gossip* for gossiper i. Each gossiper wishes only to learn his set of relevant gossip.

We will say that a sequence ϕ of phone calls *spreads* $\mathcal{S} = S_1, \ldots, S_n$ if, after ϕ is complete, each gossiper i has learned $\widehat{S}_i \supseteq S_i$, and *exactly spreads* $\mathcal{S}$ if every gossiper i has learned exactly $\widehat{S}_i = S_i$.

Definition 4.1. *For a square instance $\mathcal{S} = S_1, \ldots, S_n$, the* incomplete gossip number $\gamma(\mathcal{S})$ *is*

$$\gamma(\mathcal{S}) \quad := \quad \min_{\phi \text{ spreads } \mathcal{S}} |\phi|.$$

Note that this quantity is defined only when $\bigcup_i S_i = \{1, \ldots, n\}$—that is, we insist that each initial piece of information is interesting to some gossiper (possibly the same gossiper who knows it initially), and that no gossiper wants to learn anything other than what the other participants know.

Proposition 4.2 (Gossip Monotonicity). *For any square $\mathcal{S}$, if $\mathcal{S} \sqsubseteq \mathcal{S}'$ then $\gamma(\mathcal{S}) \leq \gamma(\mathcal{S}')$.*

Proof. Every call sequence ϕ that spreads $\mathcal{S}'$ also spreads $\mathcal{S}$. $\square$

We now relate the incomplete gossip problem to the (expanded) translocation syntenic distance. First we need to introduce some notation.

Suppose ρ is a sequence of translocations solving square $\mathcal{S} = S_1, \ldots, S_n$. Let $\mathsf{S}[\rho, \mathcal{S}]_i^t$ denote the contents of the ith set after the first t translocations of ρ have been applied to $\mathcal{S}$. Initially, $\mathsf{S}[\rho, \mathcal{S}]_i^0 = S_i$. The $(t+1)$th translocation in ρ is between $\mathsf{S}[\rho, \mathcal{S}]_x^t$ and $\mathsf{S}[\rho, \mathcal{S}]_y^t$, for some x and y, and produces two non-empty sets A and B, where $A \cup B = \mathsf{S}[\rho, \mathcal{S}]_x^t \cup \mathsf{S}[\rho, \mathcal{S}]_y^t$. Define

$$\begin{array}{rcl} \mathsf{S}[\rho, \mathcal{S}]_x^{t+1} & = & A \\ \mathsf{S}[\rho, \mathcal{S}]_y^{t+1} & = & B \\ \mathsf{S}[\rho, \mathcal{S}]_i^{t+1} & = & \mathsf{S}[\rho, \mathcal{S}]_i^t \quad \text{for all } i \notin \{x, y\}. \end{array}$$

Let ϕ be a call sequence. Let $\mathsf{K}[\phi]_i^t$ denote the set of all pieces of gossip that person i knows after the first t calls of ϕ. Thus $\mathsf{K}[\phi]_i^0 = \{i\}$ for any person i and any call sequence ϕ. If the $(t+1)$th phone call in ϕ is between gossipers p and q, then

$$\mathsf{K}[\phi]_p^{t+1} \quad = \quad \mathsf{K}[\phi]_q^{t+1} \quad = \quad \mathsf{K}[\phi]_p^t \cup \mathsf{K}[\phi]_q^t$$

and $\mathsf{K}[\phi]_i^{t+1} = \mathsf{K}[\phi]_i^t$ for all $i \notin \{p, q\}$.

Lemma 4.3. *Let $\mathcal{S}$ be square. Then $\gamma(\mathcal{S}) \geq \chi^*(\mathcal{S})$.*

Proof. Let ϕ optimally spread $\mathcal{S} = S_1, \ldots, S_n$, and let $\widehat{\mathcal{S}} = \widehat{S}_1, \ldots, \widehat{S}_n$, where gossiper i actually learns $\widehat{S}_i \supseteq S_i$ after ϕ. Note that $\mathcal{S} \sqsubseteq \widehat{\mathcal{S}}$ and that ϕ exactly spreads $\widehat{\mathcal{S}}$. By gossip monotonicity and the fact that there exists a call sequence optimally spreading $\mathcal{S}$ that also spreads $\widehat{\mathcal{S}}$, we have $\gamma(\mathcal{S}) = \gamma(\widehat{\mathcal{S}})$. We will show:

($*$) If ϕ exactly spreads $\widehat{\mathcal{S}} = \widehat{S}_1, \ldots, \widehat{S}_n$ and $\mathsf{K}[\phi]_i^{|\phi|} = \widehat{S}_i$, then $|\phi| \geq \chi(\widehat{\mathcal{S}})$.

This proves the theorem because $\gamma(\mathcal{S}) = \gamma(\widehat{\mathcal{S}}) = |\phi| \geq \chi(\widehat{\mathcal{S}}) \geq \chi^*(\mathcal{S})$ by Proposition 3.6.

We prove ($*$) by induction on $|\phi|$, implicitly constructing a sequence of $|\phi|$ translocations solving $\widehat{S}$. For the base case $|\phi| = 0$, we must have $\widehat{S}_i \subseteq \{i\}$ because no information is exchanged and $\widehat{S}_i \supseteq \{i\}$ since gossiper i initially knows information i and ϕ exactly spreads $\widehat{\mathcal{S}}$. Therefore $\chi(\widehat{\mathcal{S}}) = 0$.

For the inductive case $|\phi| \geq 1$, suppose the last call in ϕ is between gossipers p and q. Let $\mathcal{S}' = S'_1, \ldots, S'_n$ be the collection of sets just before the last call of ϕ, i.e., $S'_i = \mathsf{K}[\phi]_i^{|\phi|-1}$, and let $\phi' = \phi_1, \ldots, \phi_{|\phi|-1}$. Then ϕ' exactly spreads $\mathcal{S}'$, and we have $\mathsf{K}[\phi]_i^{|\phi'|} = S'_i = \mathsf{K}[\phi']_i^{|\phi'|}$. Applying the inductive hypothesis gives us $\chi(\mathcal{S}') \leq |\phi'| = |\phi| - 1$.

To complete the proof, it suffices to produce $\mathcal{S}'$ from $\widehat{\mathcal{S}}$ with one translocation. Define the input sets to be $\widehat{S}_p = \mathsf{K}[\phi]_p^{|\phi|}$ and $\widehat{S}_q = \mathsf{K}[\phi]_q^{|\phi|}$, and the output sets to be $S'_p = \mathsf{K}[\phi]_p^{|\phi|-1}$ and $S'_q = \mathsf{K}[\phi]_q^{|\phi|-1}$. This is a legal translocation by the definition of a phone call:

$$\begin{aligned} \mathsf{K}[\phi]_p^{|\phi|} \cup \mathsf{K}[\phi]_q^{|\phi|} &= \mathsf{K}[\phi]_p^{|\phi|} = \mathsf{K}[\phi]_q^{|\phi|} \\ &= \mathsf{K}[\phi]_p^{|\phi|-1} \cup \mathsf{K}[\phi]_q^{|\phi|-1} \end{aligned}$$

and, for all $i \notin \{p, q\}$, we have $\widehat{S}_i = \mathsf{K}[\phi]_i^{|\phi|} = \mathsf{K}[\phi]_i^{|\phi|-1} = S'_i$. □

Note that it is possible for $\gamma(\mathcal{S})$ to be strictly greater than $\chi^*(\mathcal{S})$: consider the instance $\mathcal{S} = S_1, \ldots, S_n$ where $S_i = \{(i+1) \bmod n\}$. Then $\chi^*(\mathcal{S}) = \chi(\mathcal{S}) = 0$ because each of the singletons $\{1\}, \ldots, \{n\}$ appears exactly once in the n sets. However, $\gamma(\mathcal{S}) = n - 1$—although each person only cares about one piece of gossip, it is unfortunately not the piece of gossip that he initially possesses! However, under a relatively weak assumption, we can show something akin to the other direction:

Lemma 4.4. *Suppose there exists a translocation sequence ρ solving $\mathcal{S} = S_1, \ldots, S_n$ such that after ρ, we have $S_i = \{i\}$. Then $\gamma(\mathcal{S}) \leq |\rho|$.*

Proof. Suppose we have a sequence ρ of translocations solving $\mathcal{S}$. We construct a sequence ϕ of $|\rho|$ phone calls, in the process defining the sets $\mathsf{K}[\phi]_i^t$. In our construction, we will maintain the following property:

(†) For each $1 \leq i \leq n$ and $0 \leq t \leq |\rho|$, we have $\mathsf{K}[\phi]_i^{|\rho|-t} \supseteq \mathsf{S}[\rho, \mathcal{S}]_i^t$.

We will prove this property holds by induction on $|\rho| - t$, together with our construction of the phone calls ϕ.

For the base case $t = |\rho|$, we have $\mathsf{K}[\phi]_i^0 = \{i\} = \mathsf{S}[\rho, \mathcal{S}]_i^{|\rho|}$, for any ϕ.

For the inductive case, suppose that we have defined $|\rho| - t$ phone calls, and (†) holds for all i and all $t' \geq t$. Now, suppose that the (t)th translocation in ρ involves the sets $\mathsf{S}[\rho, \mathcal{S}]_x^{t-1}$ and $\mathsf{S}[\rho, \mathcal{S}]_y^{t-1}$; we define $\phi_{|\rho|-t+1}$ to be between gossipers x and y. We must show that (†) now holds for $t - 1$. We have

$$\begin{aligned} \mathsf{K}[\phi]_x^{|\rho|-t+1} &= \mathsf{K}[\phi]_x^{|\rho|-t} \cup \mathsf{K}[\phi]_y^{|\rho|-t} \\ &\supseteq \mathsf{S}[\rho, \mathcal{S}]_x^t \cup \mathsf{S}[\rho, \mathcal{S}]_y^t \\ &\supseteq \mathsf{S}[\rho, \mathcal{S}]_x^{t-1} \end{aligned}$$

with a completely symmetric argument holding for y. For $i \notin \{x, y\}$, we have $\mathsf{K}[\phi]_i^{|\rho|-t+1} = \mathsf{K}[\phi]_i^{|\rho|-t} \supseteq \mathsf{S}[\rho, \mathcal{S}]_i^t = \mathsf{S}[\rho, \mathcal{S}]_i^{t-1}$.

This completes the proof of (†). As a consequence, we have $\mathsf{K}[\phi]_i^{|\rho|} \supseteq \mathsf{S}[\rho, \mathcal{S}]_i^0 = S_i$, and thus ϕ has spread $\mathcal{S}$ in $|\rho|$ phone calls. □

The proof is virtually identical to the proof in [25] that $\chi(\mathcal{G}^*_{n,n}) \geq 2n-4$, where $\mathcal{G}^*_{n,n}$ is the instance consisting of n copies of the set $\{1,\ldots,n\}$. The only crucial property of the instance is that after the move sequence ρ, we have $S_i = \{i\}$.

We are now in a position to state and prove the desired result connecting incomplete gossip and syntenic distance. For $\mathcal{S} = S_1,\ldots,S_n$, write $\mathcal{S}^\pi = S_{\pi_1},\ldots,S_{\pi_n}$ for π a permutation of $(1,\ldots,n)$.

Theorem 4.5. *For any square instance $\mathcal{S} = S_1,\ldots,S_n$, we have $d(\mathcal{S}) = \min_\pi \gamma(\mathcal{S}^\pi)$.*

Proof. For any permutation π, we have $d(\mathcal{S}) = d(\mathcal{S}^\pi)$ since the order of sets is irrelevant to syntenic distance. Thus $d(\mathcal{S}) = d(\mathcal{S}^\pi) = \chi^*(\mathcal{S}^\pi) \leq \gamma(\mathcal{S}^\pi)$ by Theorem 3.7 and Lemma 4.3. Since π was arbitrary, $d(\mathcal{S}) \leq \min_\pi \gamma(\mathcal{S}^\pi)$.

For the other direction, note by Theorem 3.7 we have $d(\mathcal{S}) = \chi^*(\mathcal{S})$. Let $\mathcal{S}^* \sqsupseteq \mathcal{S}$ be an instance such that $\chi^*(\mathcal{S}) = \chi(\mathcal{S}^*)$, where $\mathcal{S}^* = S^*_1,\ldots,S^*_n$. Let ρ^* be an optimal translocation sequence solving $\mathcal{S}^*$, and let π be the permutation of $(1,\ldots,n)$ so that after ρ^*, we have $S^*_{\pi_i} = \{i\}$. Let $\mathcal{S}^{\pi*} = S^{\pi*}_1,\ldots,S^{\pi*}_n$, where $S^{\pi*}_i = S^*_{\pi_i}$ and let $\rho^{\pi*}$ be ρ^* with the moves relabeled in the same way. Since ρ^* was optimal for $\mathcal{S}^*$ and we have only changed the order of the sets, $\chi(\mathcal{S}^{\pi*}) = \chi(\mathcal{S}^*) = |\rho^*| = |\rho^{\pi*}|$. By the definition of π, $\rho^{\pi*}$ is an optimal translocation sequence solving $\mathcal{S}^{\pi*}$ so that after $\rho^{\pi*}$, we have $S^{\pi*}_i = \{i\}$. By Lemma 4.4, then, we have $|\rho^{\pi*}| \geq \gamma(\mathcal{S}^{\pi*})$. Finally, by gossip monotonicity $\gamma(\mathcal{S}^{\pi*}) \geq \gamma(\mathcal{S}^\pi)$. In summary,

$$d(\mathcal{S}) \;=\; \chi^*(\mathcal{S}) \;=\; \chi(\mathcal{S}^*) \;=\; \chi(\mathcal{S}^{\pi*}) \;=\; |\rho^{\pi*}| \;\geq\; \gamma(\mathcal{S}^{\pi*}) \;\geq\; \gamma(\mathcal{S}^\pi).$$

Therefore $d(\mathcal{S}) \geq \gamma(\mathcal{S}^\pi) \geq \min_\pi \gamma(\mathcal{S}^\pi)$. $\square$

5 A Faster Algorithm for Syntenic Distance

As an application of Theorem 4.5, we give a gossip-based exact algorithm to compute the syntenic distance between species with n and k chromosomes, respectively. This new algorithm runs in time $O(nk + 2^{O(d\log d)})$ for syntenic distance d—a significant asymptotic improvement over the previous best of $O(nk + 2^{O(d^2)})$. We present the algorithm in Figure 1.

5.1 Deciding Incomplete Gossip

The decision procedure decide-gossip is the obvious exponential algorithm for deciding if $\gamma(\mathcal{S}) \leq d$ for a square instance $\mathcal{S}$. We simply enumerate all sequences of d calls among n callers, and check it see if any such sequence successfully spreads $\mathcal{S}$.

Lemma 5.1. *For any square instance $\mathcal{S}$, the procedure* decide-gossip *returns* true *iff $\gamma(\mathcal{S}) \leq d$.*

Proof. The procedure returns true iff there is a call sequence ϕ of length d spreading $\mathcal{S}$. Obviously if there is such a ϕ we have $\gamma(\mathcal{S}) \leq d$. If $\gamma(\mathcal{S}) \leq d$, then there is a ϕ of length d spreading $\mathcal{S}$: if ϕ' optimally spreads $\mathcal{S}$ where $|\phi'| < d$, then any extension of ϕ' to d calls also spreads $\mathcal{S}$. $\square$

Lemma 5.2. *The procedure* decide-gossip *runs in $O(dn^{2d+2})$ time.*

Proof. Let $\mathcal{C}_{n,d}$ be the set of call sequences of length d with n gossipers. There are $\binom{n}{2}$ different choices of callers for each of the d calls. Thus we have $|\mathcal{C}_{n,d}| = \prod_{i=1}^{d}\binom{n}{2} < n^{2d}$. We can simulate a single call in $O(n)$ time, so we can simulate any $\phi \in \mathcal{C}_{n,d}$ in $O(dn)$. Testing if $\mathcal{S}^\phi \sqsupseteq \mathcal{S}$ can be done trivially in $O(n^2)$ time, comparing element-by-element. Thus the total time for each call sequence is $O(dn^2)$, yielding $O(dn^{2d+2})$ time for the procedure overall. $\square$

decide-gossip($\mathcal{S} = \langle S_1, \ldots, S_n \rangle, d$)
// decide if $\gamma(\mathcal{S}) \le d$, where $\mathcal{S}$ is square.

1. Let $\mathcal{C}_{n,d}$ be the set of call sequences of length d with n gossipers.
2. For each $\phi \in \mathcal{C}_{n,d}$:
   (a) Simulate ϕ and let S_i^ϕ be what gossiper i learns after ϕ.
   Let $\mathcal{S}^\phi = S_1^\phi, \ldots, S_n^\phi$.
   (b) Return true if $\mathcal{S}^\phi \sqsupseteq \mathcal{S}$.
3. Otherwise return false.

decide-square-synteny($\mathcal{S} = \langle S_1, \ldots, S_n \rangle, d$)
// decide if $d(\mathcal{S}) \le d$, where $\mathcal{S}$ is square.

1. For each permutation π of $(1, \ldots, n)$:
   (a) Let $S_i^\pi = S_{\pi_i}$, and let $\mathcal{S}^\pi = S_1^\pi, \ldots, S_n^\pi$.
   (b) If decide-gossip($\mathcal{S}^\pi, d$) then return true.
2. Otherwise return false.

syntenic-distance($\mathcal{S} = \langle S_1, \ldots, S_k \rangle$)
// compute $d(\mathcal{S})$, where $\bigcup_i S_i = \{1, \ldots, n\}$.

1. If $n > k$ then consider dual($\mathcal{S}$) $= \{j : 1 \in S_j\}, \ldots, \{j : n \in S_j\}$.
2. Remove all *lonely singletons*—an element b appearing only once, in a singleton $\{b\}$—from $\mathcal{S}$.

   Let $\mathcal{S}$ be the resulting instance, and let k and n be the number of sets and elements, respectively, in $\mathcal{S}$.
3. For all sequences f of $k - n$ fusions, let $\mathcal{S}^f$ be the instance resulting from fusing the sets specified in f, starting from $\mathcal{S}$.
4. Search sequentially for the smallest $\delta \in \{0, \ldots, 2n-2\}$ so that, for some f, decide-square-synteny($\mathcal{S}^f, \delta$).
5. Return $\delta + k - n$.

Figure 1: An gossip-based algorithm to exactly compute syntenic distance.

5.2 Deciding the Syntenic Distance of Square Instances

We use the connection between gossip and synteny to build a decision procedure decide-square-synteny for the syntenic distance of a square instance $\mathcal{S}$. Because of the differences between the two problems with respect to the correspondence between the orderings of the sets and of the elements, we use brute force to examine all permutations of the sets.

Lemma 5.3. *For any square instance $\mathcal{S}$, the procedure* decide-square-synteny *returns* true *iff $d(\mathcal{S}) \leq d$.*

Proof. By Theorem 4.5, we have $d(\mathcal{S}) = \min_\pi \gamma(\mathcal{S}^\pi)$. The procedure decide-square-synteny returns true iff we have $\gamma(\mathcal{S}^\pi) \leq d$ for some π, by Lemma 5.1. □

Lemma 5.4. *The procedure* decide-square-synteny *requires $O(n!dn^{2d+2})$ time.*

Proof. There are $n!$ permutations of $(1, \ldots, n)$, and thus $n!$ calls to decide-gossip, each of which requires $O(dn^{2d+2})$ time by Lemma 5.2. □

5.3 Computing Syntenic Distance

In syntenic-distance, we use decide-square-synteny to compute the actual syntenic distance of an instance $\mathcal{S}$, where $\mathcal{S}$ is arbitrary (i.e., not necessarily square). First, however, we do some preprocessing based on the results of DasGupta et al. [11].

For an instance $\mathcal{S} = S_1, \ldots, S_k$ and $\bigcup_i S_i = \{1, \ldots, n\}$, the *dual* of $\mathcal{S}$ is the instance $\mathcal{S}' = S'_1, \ldots, S'_n$, where $j \in S'_i$ iff $i \in S_j$. Call a set S_i a *lonely singleton* iff $S_i = \{b\}$ and $b \notin \bigcup_{j \neq i} S_j$.

Lemma 5.5 (Duality [11]). *For any instance $\mathcal{S}$ with dual $\mathcal{S}'$, we have $d(\mathcal{S}) = d(\mathcal{S}')$.*

Lemma 5.6 (Lonely Singleton Removal [11]). *Let $\mathcal{T}$ be an instance with a lonely singleton $T_i = \{b\}$. Let $\mathcal{S}$ be the instance obtained by removing the element b and the set T_i from $\mathcal{T}$. Then $d(\mathcal{T}) = d(\mathcal{S})$.*

Thus we can limit ourselves to the case where $k \geq n$—otherwise we consider the dual instance—and there are no lonely singletons. We can compute the dual and remove all lonely singletons in $O(nk)$ time.

Theorem 5.7. *The procedure* syntenic-distance *computes $d(\mathcal{S})$.*

Proof. Note that, by duality, considering the dual instance does not change the distance. Also note that removing the lonely singletons does not alter the distance, by Lemma 5.6. Therefore we presume that $k \geq n$ and that there are no lonely singletons.

For a sequence f of $n-k$ fusions, write $\mathcal{S}^f$ to denote the instance resulting from the application of f to $\mathcal{S}$. By Lemma 5.3 (and the observation that $2n-2$ moves suffice to solve any $\mathcal{S}^f$), we have $\delta = \min_f d(\mathcal{S}^f)$. After any sequence f of $k-n$ fusions, the resulting instance $\mathcal{S}^f$ is square. Let f be a fusion sequence for which $\delta = d(\mathcal{S}^f)$. Doing f followed by an optimal move sequence for $\mathcal{S}^f$ solves $\mathcal{S}$, so $d(\mathcal{S}) \leq k - n + d(\mathcal{S}^f) = k - n + \delta$.

For the other direction, it suffices to identify a fusion sequence f such that $d(\mathcal{S}) = k-n+d(\mathcal{S}^f) \geq k - n + \min_f d(\mathcal{S}^f) = k - n + \delta$. In any move sequence that solves $\mathcal{S}$, we must decrease the number of sets from k to n, which requires $k - n$ fusions. Let σ be an optimal canonical move sequence solving $\mathcal{S}$, and let f denote the first $k - n$ fusions in σ. Let $\mathcal{S}^f$ be the instance resulting from completing them. Since σ was presumed to be optimal, we have $d(\mathcal{S}) = k - n + d(\mathcal{S}^f)$. □

We turn to the running time, after noting the following fact:

Lemma 5.8. *For any instance $\mathcal{S} = S_1, \ldots, S_k$ with $d(\mathcal{S}) < k/2$, there is a lonely singleton S_i.*

Proof. A move sequence σ can only touch at most $2|\sigma|$ input sets, since each fusion, fission, and translocation takes no more than two sets as input. Since $d(\mathcal{S}) < k/2$, an optimal move sequence σ touches fewer than k input sets, leaving at least one set S_i unaltered by σ. If the untouched set S_i were not a lonely singleton, then σ would not solve $\mathcal{S}$. □

This is the motivation for the removal of lonely singletons: repeatedly doing so results in an instance where the distance is on the order of the number of chromosomes.

Theorem 5.9. *For any $\mathcal{S} = S_1, \ldots, S_k$ with $\bigcup S_i = \{1, \ldots, n\}$ and $d(\mathcal{S}) = d$, the procedure* syntenic-distance *requires $O(nk + 2^{(k+2d+3)\log k})$ time. This is $O(nk + 2^{(5k-1)\log k})$ and $O(nk + 2^{(4d+3)\log 2d})$.*

Proof. After the preprocessing steps, we have that $k \geq n$ and there are no lonely singletons. This preprocessing requires $O(nk)$ time. We now consider the running time of steps 3–5.

First we count the number of sequences of fusions. There are $\binom{k-i+1}{2}$ choices of sets to involve in the ith fusion, and $k - n$ total fusions. Therefore there are $\prod_{i=1}^{k-n} \binom{k-i+1}{2} \leq \prod_{i=1}^{k-n} \binom{k}{2} < k^{2k-2n}$ fusion sequences f.

Let $k - n + \delta = d$, so that $\delta = \min_f d(\mathcal{S}^f)$. We run δ iterations of the sequential search; by Lemma 5.4, the time to complete iteration i is $O(n!in^{2i+2})$ time. This is asymptotically dominated by the last iteration, so the total time is $O(n!\delta n^{2\delta+2})$.

Thus the total running time is at most $O(k^{2k-2n} \cdot n!\delta n^{2\delta+2})$. Since $k \geq n$ and $k! \leq k^k$, this is

$$O\left(k^{2k-2n} \cdot n!\delta n^{2\delta+2}\right) \quad = \quad O\left(\delta \cdot k^{k+2(\delta+k-n)+2}\right) \quad = \quad O\left(k^{k+2d+3}\right).$$

since $\delta \leq d \leq 2k - 2$. This is $O(2^{(k+2d+3)\log k})$.

For the other versions of the bounds, we need only note that $d \leq 2k-2$ and that by Lemma 5.8 we can remove lonely singletons until $2d \geq k$. □

6 Conclusion

In this paper, we have defined the incomplete gossip problem—a novel generalization of the classical gossip problem—and shown a tight relationship between it and the syntenic distance between genomes. We believe that incomplete gossip is an interesting problem in its own right, and there are a number of open questions about it. No complexity results are known, and our exact algorithm for incomplete gossip is completely naive; there may be far more efficient solutions.

Using this connection, we have presented a faster exact algorithm for the syntenic distance problem, though it is admittedly practical only for very closely-related species. Whether we can further speed this computation—by refining the techniques presented here, or using some other approach—is an open question.

One possible approach to improving this algorithm is based on the *component bound* [11]. Consider the intersection graph G of the k chromosomes of an instance $\mathcal{S}$ with n elements, say with $n \geq k$. Each move can only increase the number of components of G by one, and we must end with

n components; thus $d(\mathcal{S}) \geq n - p$ where there are p components in G. For values of d that are only slightly larger than $n - p$, the only move sequences that could possibly solve $\mathcal{S}$ in d moves must increase the number of components by one in almost every move. Such *splitting moves* must be fissions or translocations between sets in the same component of G, which can dramatically limit the number of possible moves. (For example, testing if $d = n - p$ is equivalent to testing if there are $n - p$ consecutive splitting moves; since splitting moves are always within components, we can run our exhaustive algorithm only within each component.) Thus the set of d-move sequences possibly solving $\mathcal{S}$ may be much smaller than the set we considered in our algorithm.

There are several known polynomial-time 2-approximations for the syntenic distance problem. Approximating syntenic distance to within a factor better than 2 appears to be difficult, though no inapproximability results are known. We hope that this connection between syntenic distance and incomplete gossip may help to shed light on the problem of approximating syntenic distance, and, more generally, on a variety of questions in the areas of genome rearrangements and information flow.

7 Acknowledgements

I would like to thank Jon Kleinberg for his suggestions on the presentation of this material, and Tarun Agarwal, Chris Batten, Anne Bergeron, and Jeff Vinocur for their comments on previous drafts of this paper. The anonymous referees were also extremely helpful, especially with Section 5 and the overall organization of the paper.

The majority of this work was performed while at the University of Cambridge, supported in part by a Churchill Scholarship from the Winston Churchill Foundation, an NSF Graduate Research Fellowship, and the ONR Young Investigator Award of Jon Kleinberg.

References

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

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

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

[4] W. Bradley Barbazuk, Ian Korf, Candy Kadavi, Joshua Heyen, Stephanie Tate, Edmund Wun, Joseph A. Bedell, John D. McPherson, and Stephen L. Johnson. The syntenic relationship of the zebrafish and human genomes. *Genome Research*, 10(9):1351–1358, September 2000.

[5] Mathieu Blanchette, Takashi Kunisawa, and David Sankoff. Parametric genome rearrangement. *Gene*, 172(1):GC11–GC17, 1996.

[6] Richard T. Bumby. A problem with telephones. *SIAM J. Alg. Disc. Meth.*, 2(1):13–18, March 1981.

[7] Alberto Caprara. Sorting permutations by reversals and Eulerian cycle decompositions. *SIAM J. Discrete Math.*, 12(1):91–110, February 1999.

[8] Gerard J. Chang and Yuh-Jiuan Tsay. The partial gossiping problem. *Discrete Math.*, 148(1–3):9–14, January 1996.

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

[10] Norbert Cot. Extensions of the telephone problem. In *7th SE Conference on Combinatorics, Graph Theory, and Computing*, pages 239–256, Winnipeg, 1976. Utilitas Mathematics.

[11] Bhaskar DasGupta, Tao Jiang, Sampath Kannan, Ming Li, and Elizabeth Sweedyk. On the complexity and approximation of syntenic distance. *Discrete Appl. Math.*, 88(1–3):59–82, November 1998.

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

[13] Niklas Eriksen. $(1 + \varepsilon)$-approximation of sorting by reversals and transpositions. *Theoret. Comp. Sci.* To appear.

[14] Henrik Eriksson, Kimmo Eriksson, Johan Karlander, Lars Svensson, and Johan Wästlund. Sorting a bridge hand. *Discrete Math.*, 241(1–3):289–300, October 2001.

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

[16] Qian-Ping Gu, Shietung Peng, and Ivan Hal Sudborough. A 2-approximation algorithm for genome rearrangements by reversals and transpositions. *Theoret. Comp. Sci.*, 210(2):327–339, January 1999.

[17] A. Hajnal, E. C. Milner, and E. Szemerédi. A cure for the telephone disease. *Canad. Math. Bull.*, 15(3):447–450, September 1972.

[18] Sridnhar Hannenhalli. Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Appl. Math.*, 71(1–3):137–151, December 1996.

[19] Sridnhar Hannenhalli and Pavel Pevzner. Transforming mice into men (polynomial algorithm for genomic distance problem). In *36th IEEE Symposium on Foundations of Computer Science*, pages 581–592, October 1995.

[20] Sandra Hedetniemi, Stephen Hedetniemi, and Arthur Liestman. A survey of gossiping and broadcasting in communication networks. *Networks*, 18(4):319–349, 1988.

[21] C. A. J. Hurkens. Spreading gossip efficiently. *Nieuw Arch. Wiskd.*, 5/1(2):208–210, June 2000.

[22] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, February 2001.

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

[24] J. D. Kececioglu and D. Sankoff. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangment. *Algorithmica*, 13(1/2):180–210, January/February 1995.

[25] Jon Kleinberg and David Liben-Nowell. The syntenic diameter of the space of n-chromosome genomes. In David Sankoff and Joseph H. Nadeau, editors, *Comparative Genomics*, pages 185–197. Kluwer Academic Press, 2000.

[26] Hsun-Ming Lee and Gerard J. Chang. Set to set broadcasting in communication networks. *Discrete Appl. Math.*, 40(3):411–421, September 1992.

[27] David Liben-Nowell. On the structure of syntenic distance. *J. Comp. Bio.*, 8(1):53–67, February 2001.

[28] David Liben-Nowell and Jon Kleinberg. Structural properties and tractability results for linear synteny. In *11th Annual Symposium on Combinatorial Pattern Matching*, pages 248–263, June 2000.

[29] Aoife McLysaght, Cathal Seoighe, and Kenneth H. Wolfe. High frequency of inversions during eukaryote gene order evolution. In David Sankoff and Joseph H. Nadeau, editors, *Comparative Genomics*, pages 47–58. Kluwer Academic Press, 2000.

[30] Nadia Pisanti and Marie-France Sagot. Further thoughts on the syntenic distance between genomes. Submitted for publication.

[31] José María Ranz, Ferran Casals, and Alfredo Ruiz. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Research*, 11(2):230–239, February 2001.

[32] Dana Richards and Arthur L. Liestman. Generalizations of broadcasting and gossiping. *Networks*, 18:125–138, 1988.

[33] David Sankoff and Joseph H. Nadeau. Conserved synteny as a measure of genomic distance. *Discrete Appl. Math.*, 71(1–3):247–257, December 1996.

[34] Cathal Seoighe, Nancy Federspiel, Ted Jones, Nancy Hansen, Vesna Bivolarovic, Ray Surzycki, Raquel Tamse, Caridad Komp, Lucas Huizar, Ronald W. Davis, Stewart Scherer, Evelyn Tait, Duncan J. Shaw, David Harris, Lee Murphy, Karen Oliver, Kate Taylor, Marie-Adele Rajandream, Bart G. Barrell, and Kenneth H. Wolfe. Prevalence of small inversions in yeast gene order evolution. *Proc. Natl. Acad. Sci. USA*, 97(26):14433–14437, December 2000.

[35] R. Tijdeman. On a telephone problem. *Nieuw Arch. Wiskd.*, 19(3):188–192, 1971.

[36] Zdenek Trachtulec and Jiri Forejt. Synteny of orthologous genes in mammals, snake, fly, nematode, and fission yeast. *Mammalian Genome*, 12(3):227–231, March 2001.

[37] S. Randal Voss, Jeramiah J. Smith, David M. Gardiner, and David M. Parichy. Conserved vertebrate chromosome segments in the large salamander genome. *Genetics*, 158(2):735–746, June 2001.