Reconstructing the Pre-Doubling Genome

Nadia El-Mabrouk * David Bryant † David Sankoff ‡

Abstract

Genome duplication is an important source of new gene functions and novel physiological pathways. In the course of evolution, the nucleotide sequences of duplicated genes tend to diverge through mutation, so that one copy loses function (and disappears from view) or develops a new function, encoding a distinct but similar product. Originally a duplicated genome contains two identical copies of each chromosome, but through reciprocal translocation, parallel linkage patterns between the two copies are disrupted. Eventually, all that can be detected are several chromosome segments of greater or lesser length (*blocks*), each of which appears twice in the genome, containing many paralogous genes in parallel orders. We present an exact algorithm for reconstructing the ancestral pre-doubling genome in polynomial time, minimizing in key cases the number of translocations required to derive the observed order and orientation of blocks along the present-day chromosomes. We apply this to the genome duplication which has been described for *Saccharomyces cerevisiae*.

1 Genome duplication

Perhaps the most spectacular cause of gene duplication is tetraploidization of the genome. Normally a lethal accident of meiosis or other reproductive step, if this doubling of the genome can be resolved in the organism and eventually fixed as a normalized diploid state in a population, it represents a simultaneous duplication of the entire genetic complement. It transcends other mechanisms for gene duplication in that not only is one copy of each gene free to evolve its own function, but it can evolve in concert with any subset of the thousands of other extra gene copies (cf [4] for accounts of gene family coevolution). Whole new physiological pathways may emerge, involving novel functions for many of these genes. Genome duplication is thus a likely source of rapid and far-reaching evolutionary progress. Its rarity does not detract from its importance.

Evidence for its effects has shown up across the eukaryote spectrum. More than two hundred million years ago the vertebrate genome underwent two duplications [2, 7, 12]. Although numerous chromosome rearrangements such as inversions and reciprocal translocations have subsequently occurred, the number of rearrangements has been sufficiently modest that hundreds of conserved paralogous segments can be detected in the human genome since the ancient duplications; similar observations hold for the murine genome [10, 11] and for less intensively mapped vertebrate genomes. More recent genome duplications are known to have occurred in some vertebrate lines, such as the frogs [19], the salmoniform fish [12] and zebrafish [14].

Comparison of chromatin-eliminating *Ascaridae* with other nematodes suggest that somatic cells of these worms have discarded a good proportion of the genes present in germ cells, possible because these are redundant duplicates arising through genomic doubling some 200 million years ago [8].

Genome duplication is particularly prevalent in plants. Comparison of the well-studied rice [1], oats (wild and domestic), corn [1, 5] and wheat [9] genomes indicate several occurrences in the cereal lineage. Soybeans [17], rapeseed [15], and other cultivars have genome duplications in their ancestry. Paterson *et al.* have presented convincing evidence that one or more genome duplications also occurred much earlier in plant evolution [13].

Recently, following the complete sequencing of all *Saccharomyces cerevisiae* chromosomes, the prevalence of gene duplication has led to the conclusion that this yeast genome is also the product of an ancient doubling [18].

Subsequent to genome duplication, duplicated genes tend to diverge through mutation, so that one copy loses function (becomes a pseudogene) or develops a new function, encoding a distinct but similar product. Originally a duplicated genome contains two identical copies of each chromosome, but through inversion or other intrachromosomal movement, the gene orders in each pair of chromosomes change independently, and through reciprocal translocation, parallel linkage patterns between the two copies are disrupted. Eventually, all that can be detected are several chromosome segments of greater or lesser length (*blocks*), each of which appears twice in the genome, containing many paralogous genes in parallel

*Département d'Informatique et de recherche opérationnelle, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: mabrouk@iro.umontreal.ca.

†Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: bryant@crm.umontreal.ca.

‡Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: sankoff@ere.umontreal.ca.

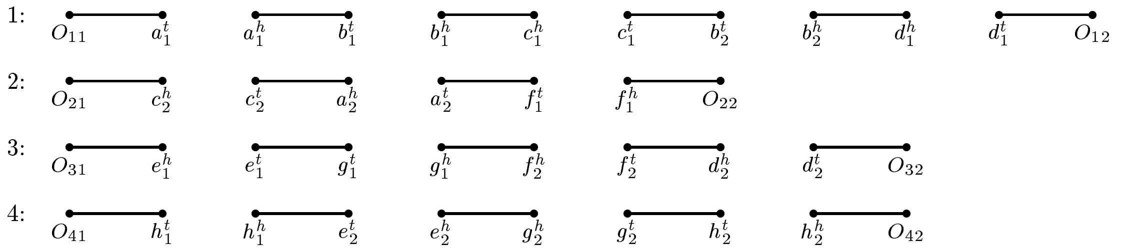

Figure 1: Partial graph corresponding to Example 1

orders. We previously proposed a suite of "Genome halving" problems [3] and offered an algorithm for one of them involving (set-theoretical) relations of synteny only. Here we propose to deal with the evolution of gene *order* and *transcriptional orientation* on each chromosome. We present a polynomial algorithm for finding an exact solution to most of the interesting instances of the problem.

2 Genome halving of signed, ordered chromosomes.

A **block string** is a string of signed (+ or −) terms (**blocks**) from a set $\mathcal{B}$. A **rearranged duplicated genome** G is a collection of non-null block strings, $C_1, \cdots, C_{2N}$, (**chromosomes**), such that each block in $\mathcal{B}$ is present exactly twice, i.e., once in each of two different chromosomes, or twice in a single chromosome.

Example 1 *Let $\mathcal{B} = \{a, b, c, d, e, f, g, h\}$ be a set of 8 different blocks, and G a genome consisting of four chromosomes:*
1: $+a\ +b\ -c\ +b\ -d$;   2: $-c\ -a\ +f$;
3: $-e\ +g\ -f\ -d$;   4: $+h\ +e\ -g\ +h$.
G is a rearranged duplicated genome. Each block appears exactly twice in the set of chromosomes. E.g. block b appears twice in chromosome 1. Signs represent block orientation.

For block string $X = x_1 x_2 \cdots x_r$, denote by $-X$ the **reverse** string $-x_r - x_{r-1} \cdots - x_1$.

The problem is to calculate the minimum number of translocations required to transform a given rearranged duplicated genome G into some **perfect duplicated genome** H (to be found), consisting of $K_1, \cdots, K_{2M}$ chromosomes, where for each $i \in \{1, \cdots, 2M\}$, we have $K_i = K_j$ for exactly one $j \in \{1, \cdots, 2M\} \backslash \{i\}$.

Let X_1, X_2, Y_1 and Y_2 be non-null block strings. A **reciprocal translocation** between two chromosomes $X = X_1X_2$ and $Y = Y_1Y_2$ is of form $X_1X_2, Y_1Y_2 \longrightarrow X_1Y_2, Y_1X_2$ (prefix-prefix) or of form $X_1X_2, Y_1Y_2 \longrightarrow X_1 - Y_1, -Y_2X_2$ (prefix-suffix).

3 The Hannenhalli graph.

Given two genomes $H_1 = C_{1,1}, \cdots, C_{1,N}$ and $H_2 = C_{2,1}, \cdots, C_{2,N}$ such that H_1 and H_2 contain the same blocks, each of the $|\mathcal{B}|$ blocks appears exactly once in each genome, and the set containing the $2N$ initial and final blocks in all the chromosomes of H_1 is the same as in H_2. How many reciprocal translocations, as described in Section 2, does it take to transform H_1 into H_2?

Hannenhalli [6] solved this using $\mathcal{G}_{12}$, the bicoloured **cycle graph** of H_1 with respect to H_2. If block x_i in chromosome $X = x_1 \cdots x_k$ of H_1 has positive sign, replace it by the pair $x_i^t x_i^h$, and if it is negative, by $x_i^h x_i^t$. Then the vertices of $\mathcal{G}_{12}$ are just the x^t and the x^h for all x in $\mathcal{B}$. Any two vertices which are adjacent in some chromosome in H_1, other than x_i^t and x_i^h from the same x, are connected by a black edge, and any two adjacent in H_2, by a gray edge. Each vertex is incident to exactly one black and one gray edge, so that there is a unique decomposition of $\mathcal{G}_{12}$ into c_{12} disjoint cycles of alternating edge colours. Note that $c_{21} = c_{12} = c$ is maximized when $H_1 = H_2$, in which case each cycle has one black edge and one gray edge, and $c = |\mathcal{B}| - N$.

Hannenhalli showed that the minimum number of reciprocal translocations necessary to transform H_1 into H_2 is $|\mathcal{B}| - N - c$ in all but certain cases. The exceptional cases contain *subpermutations*, a number of contiguous, but differently ordered blocks in both H_1 and H_2. Note that these are precisely the cases where, from a biological viewpoint, a comparison of the genomes would seem to require *inversions* (reversals) or *transpositions* (interchanging two adjacent block strings), as well as translocations.

4 Maximizing the number of cycles.

4.1 Preliminaries

To make use of the Hannenhalli graph structure for the genome halving problem, we first introduce, arbitrarily, a distinction within each pair of identical blocks in the rearranged duplicated genome G, labeling one occurrence x_1 and the other x_2 for all x in $\mathcal{B}$.

Next, to each chromosome C_i, we add new initial and final terms $+O_{i1}$ and $+O_{i2}$. This releases the erstwhile initial and final blocks on each chromosome from their constraint in the Hannenhalli formulation and ensures that all translocations, including those which reduce (by **fusion**, e.g. null X_1Y_2) or augment (by **fission**, e.g. null X_1X_2) the number of chromosomes in the genome, can be treated as reciprocal translocations. Chromosomes consisting of just one initial and one final O are dummies. They allow $M \neq N$, and G to have an odd number of chromosomes, in the formulation of the problem in Section 2 while still making use of the Hannenhalli graph in which H_1 and H_2 have the same number of chromosomes.

In each chromosome, each x_j (except the O_{ij}) is replaced by x_j^t and x_j^h as in the Hannenhalli construction. Define:

$$O = \{O_{i1}, O_{i2}\}_{i=1,\cdots,2N}, \; V = \{x_j^s\}^{s \in \{h,t\}}_{\substack{x \in \mathcal{B} \\ j=1,2}}, \; \mathbf{V} = O \cup V.$$

We use the notation $\overline{1} = 2$, $\overline{2} = 1$, $\widetilde{t} = h$, $\widetilde{h} = t$. For $u = x_j^s \in V$, its **counterpart**, denoted $\overline{u}$, is $x_{\overline{j}}^s$, and its **obverse**, denoted $\widetilde{u}$, is $x_j^{\widetilde{s}}$. Note that $\overline{\overline{u}} = \widetilde{\widetilde{u}} = u$.

The **partial graph** $\mathcal{G}(\mathbf{V}, A)$ associated with G, has the edge set A of (black) undirected edges linking adjacent terms (other than obverses) in G. The partial graph associated

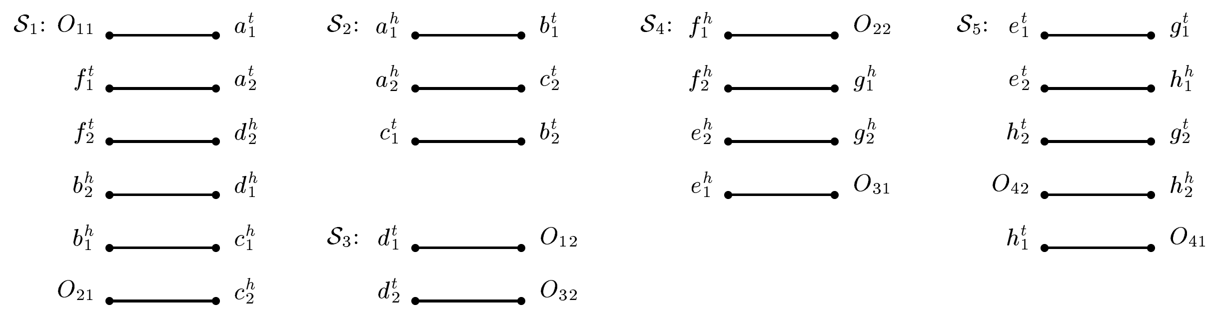

Figure 2: The natural subgraphs of $\mathcal{G}(\mathbf{V}, A)$.

with the genome in Example 1 is shown in Figure 1. To differentiate the two occurrences of each block x, one is subscripted "1", its counterpart "2".

The addition to the partial graph $\mathcal{G}(\mathbf{V}, A)$ of a set D of gray undirected edges, corresponding to some perfect duplicated genome H, produces a **completed graph** $\mathcal{G}(\mathbf{V}, A, D)$. Note that every vertex in $\mathbf{V}$ will then be incident to exactly one black edge and one gray edge. Our goal is to find a perfect duplicated genome H, with edge set D, which maximizes the number of cycles in $\mathcal{G}(\mathbf{V}, A, D)$; we call this a **maximal completed graph**.

Lemma 1 follows directly from the above definitions.

Lemma 1 *In a completed graph $\mathcal{G}(\mathbf{V}, A, D)$,*

1. *D contains no edge of form $(u, \overline{u})$, for any $u \in V$.*
2. *Suppose $(u, v) \in D$ and $v \in V$.*
   (a) *If $u \in V$ then $(\overline{u}, \overline{v}) \in D$.*
   (b) *If $u \in O$ then $\overline{v}$ is also linked by a gray edge to some element of O.*

Let $\mathcal{G}(\mathbf{V}, A)$ contain a subgraph $\mathcal{G}(\mathbf{V}', A')$, representing a set of fragments of the $2N$ chromosomes of G. Lemma 2 states conditions on the vertices in $\mathbf{V}'$ for it to be possible to add gray edges satisfying Lemma 1.

Lemma 2 1. *If $u \in \mathbf{V}' \cap V$, then $\overline{u} \in \mathbf{V}'$.*

2. *$\mathbf{V}'$ contains an even number (possibly zero) of elements of O.*
3. *Let $\mathbf{V}''$ be the subset of $\mathbf{V}'$ containing* **fragment endpoints**, *i.e., vertices u satisfying one of:*
   - *$u \in O$.*
   - *If $u \in V$, then $\widetilde{u} \notin \mathbf{V}'$.*

   Let $p = |\mathbf{V}''|$ be the number of elements of $\mathbf{V}''$. p must be a multiple of four.

Proof: Points (1) and (2) follow from Lemma 1, points (2a) and (2b), respectively.

It can be seen that $p/2$ is the number of chromosome fragments represented by $\mathcal{G}(\mathbf{V}', A')$. In order that some sequence of reciprocal translocations can transform these fragments into a set of duplicated fragments, we require that $p/2$ be even. $\Box$

A subgraph $\mathcal{G}(\mathbf{V}', A')$ of $\mathcal{G}(\mathbf{V}, A)$ satisfying Lemma 2 is called a **completable subgraph**.

4.2 Decomposition into completable subgraphs

Definition : Let $e = (u, v) \in A$. Define A_e recursively by:

- $(u, v) \in A_e$;
- If $(x, y) \in A_e$ and $x \notin O$ then the edge of A adjacent to $\overline{x}$ is also in A_e. Similarly, if $y \notin O$ then the edge of A adjacent to $\overline{y}$ is also in A_e.

Let $\mathbf{V}_e$ be the subset of $\mathbf{V}$ made up of vertices incident to the edges in A_e. Then $\mathcal{G}(\mathbf{V}_e, A_e)$ is the **natural subgraph** (of **size** $|A_e|$) of $\mathcal{G}(\mathbf{V}, A)$ generated by e. Note that if $f \in A_e$, then $A_f = A_e$

Consider the genome in Example 1. The natural subgraphs of $\mathcal{G}(\mathbf{V}, A)$ are as in Figure 2.

Theorem 1 *A natural subgraph is completable iff it is of even size.*

Proof: Let $\mathcal{G}(\mathbf{V}_e, A_e)$ be a natural subgraph of $\mathcal{G}(\mathbf{V}, A)$ of size n (i.e. $|\mathbf{V}_e| = 2n$). By definition, $\mathcal{G}(\mathbf{V}_e, A_e)$ satisfies condition (1) of Lemma 2. Moreover, by construction, it contains either zero or two elements of O, and so satisfies condition (2). To see that it also satisfies condition (3) iff n is even, let $\mathbf{V}_{e,1}$ be the subgraph of $\mathbf{V}_e$ defined as in part (3) of Lemma 2, and $\mathbf{V}_{e,2} = \mathbf{V}_e \setminus \mathbf{V}_{e,1}$. Let $q = |\mathbf{V}_{e,2}|$ and $p = |\mathbf{V}_{e,1}|$. By the definitions of $\mathbf{V}_{e,1}$ and $\mathbf{V}_{e,2}$, $p = 2n - q$ and q must be a multiple of 4. Thus, $p/2$ is even iff n is even. $\Box$

We then divide the set GC of natural subgraphs of $\mathcal{G}(\mathbf{V}, A)$ into the following subsets:

- GCE is the subset of GC containing the completable natural subgraphs (i.e. of even size).
- GCO is the subset of GC containing the natural subgraphs of odd size. We further subdivide GCO into GCO_+ and GCO_- according to whether the natural subgraphs include vertices in O or not.

The set A contains $2(|\mathcal{B}| + N)$ edges, and subgraphs in GCE contain an even number of edges. Then GCO must also contain an even number of edges, and thus an even number of subgraphs. We can then pair off all the subgraphs in GCO as follows, and amalgamate the two subgraphs in each pair in order to produce completable subgraphs of $\mathcal{G}(\mathbf{V}, A)$:

- Arbitrarily choose pairs of subgraphs in GCO_+ to amalgamate. The set of larger subgraphs thus formed is denoted $\mathcal{CO}_+$.

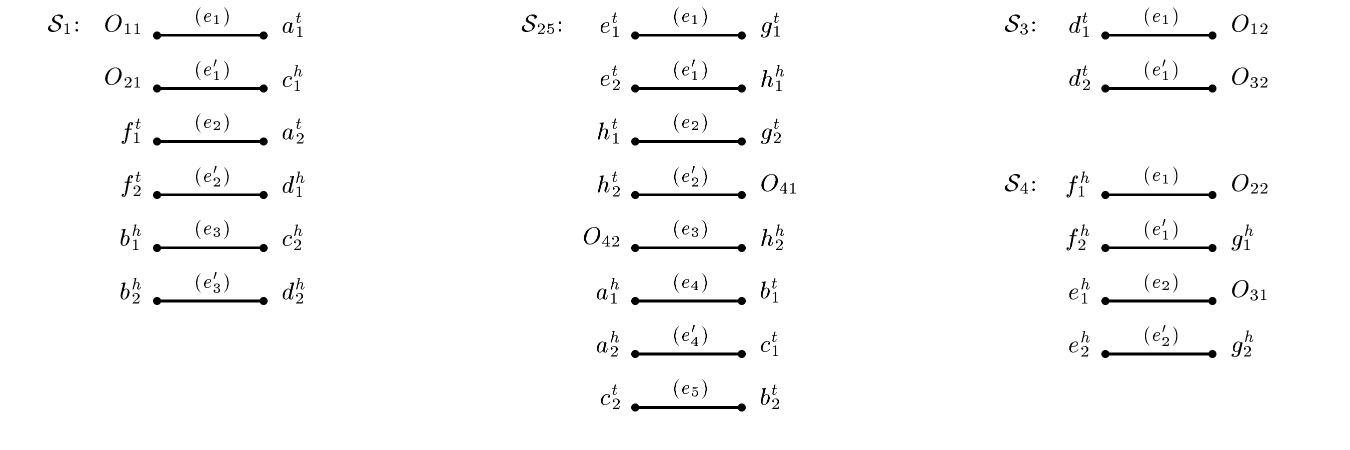

Figure 3: A suitable order for the edges of the subgraphs. In our notation, $\overline{O_{11}} = O_{21}$, $\overline{O_{41}} = O_{42}$, $\overline{O_{12}} = O_{32}$ and $\overline{O_{22}} = O_{31}$. For subgraphs $\mathcal{S}_1, \mathcal{S}_3$ and $\mathcal{S}_4$, the sets of left and right vertices are made up of the vertices on the left and right, respectively, of all their edges. In the case of $\mathcal{S}_{25}$ the right vertices are the vertices on the right of its edges, plus O_{42}^t and c_2^t. The remaining vertices are left vertices.

- Arbitrarily choose pairs of the remaining subgraphs in GCO to amalgamate. This includes subgraphs in GCO_- plus, if applicable, the remaining one in GCO_+. The set of subgraphs thus formed is denoted $\mathcal{CO}_-$.

The subgraphs in $GCE \cup \mathcal{CO}_+ \cup \mathcal{CO}_-$ are called **supernatural** subgraphs. We denote $\mathcal{CE} = GCE \cup \mathcal{CO}_+$.

Example 2 *Consider the natural subgraphs $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$ and $\mathcal{S}_5$ of Figure 2. Note that $\mathcal{S}_1$, $\mathcal{S}_3$, $\mathcal{S}_4 \in GCE$, $\mathcal{S}_2 \in GCO_-$ and $\mathcal{S}_5 \in GCO_+$.*

Let $\mathcal{S}_{25}$ be the supernatural subgraph in $\mathcal{CO}_-$ obtained by amalgamating $\mathcal{S}_2$ and $\mathcal{S}_5$. Then the set $\{\mathcal{S}_1, \mathcal{S}_{25}, \mathcal{S}_3, \mathcal{S}_4\}$ is a decomposition of $\mathcal{G}(\mathbf{V}, A)$ into supernatural subgraphs.

Notation :

- In a supernatural subgraph $\mathcal{G}(\mathbf{V}', A')$ in $GCE \cup \mathcal{CO}_-$, for each vertex u in $\mathbf{V}' \cap O$, if there is one, $\overline{u}$ designates the (only) other vertex in $\mathbf{V}' \cap O$.
- Let $\mathcal{G}_1(\mathbf{V}'_1, A'_1)$ and $\mathcal{G}_2(\mathbf{V}'_2, A'_2)$ be the two natural subgraphs in GCO_+ which make up a subgraph $\mathcal{G}(\mathbf{V}', A')$ of $\mathcal{CO}_+$. If $u \in \mathbf{V}'_1 \cap O$, then we arbitrarily choose one of the two vertices of $\mathbf{V}'_2 \cap O$ to be $\overline{u}$.

Let $\mathcal{G}(\mathbf{V}', A')$ be a supernatural subgraph of $\mathcal{G}(\mathbf{V}, A)$ of size $2n$, where $n > 1$. Relabeling the vertices in $\mathbf{V}'$ allows us to define a **suitable order** for the edges in A'.

1. If $\mathcal{G}(\mathbf{V}', A') \in \mathcal{CE}$: $\mathbf{V}' = \mathbf{V}'_l \bigcup \mathbf{V}'_r$, where $\mathbf{V}'_l = \bigcup_{1 \le i \le n} \{a_i, \overline{a_i}\}$ and $\mathbf{V}'_r = \bigcup_{1 \le i \le n} \{b_i, \overline{b_i}\}$ are the sets of **left** and **right vertices** of $\mathbf{V}'$, respectively.
   $A' = \{e_1, e'_1, \cdots, e_n, e'_n\}$ such that
   - $e_1 = (a_1, b_1)$; $e'_1 = (\overline{a_1}, b_2)$.
   - For all i, $1 < i < n$, $e_i = (a_i, \overline{b_{i-1}})$ and $e'_i = (\overline{a_i}, b_{i+1})$.
   - $e_n = (a_n, \overline{b_{n-1}})$; $e'_n = (\overline{a_n}, \overline{b_n})$.
2. If $\mathcal{G}(\mathbf{V}', A') \in \mathcal{CO}_-$, let $\mathcal{G}(\mathbf{V}'_1, A'_1)$ and $\mathcal{G}(\mathbf{V}'_2, A'_2)$ be its two component natural subgraphs, of sizes $2n_1 - 1$ and $2n_2 - 1$, respectively.
   $\mathbf{V}'_1 = \bigcup_{1 \le i \le n_1 - 1} \{a_i, \overline{a_i}, b_i, \overline{b_i}\} \bigcup \{b_{n_1}, \overline{b_{n_1}}\}$ and $A'_1 = \{e_1, e'_1, \cdots, e_{n_1 - 1}, e'_{n_1 - 1}, e_{n_1}\}$ where the e_i and e'_i are defined as above, except $e_{n_1} = (\overline{b_{n_1}}, \overline{b_{n_1 - 1}})$.
   Similarly, $\mathbf{V}'_2 = \bigcup_{n_1 + 1 \le i \le n_1 + n_2 - 1} \{a_i, \overline{a_i}, b_i, \overline{b_i}\} \bigcup \{b_{n_1 + n_2}, \overline{b_{n_1 + n_2}}\}$ and $A'_2 = \{e_{n_1 + 1}, e'_{n_1 + 1}, \cdots, e_{n_1 + n_2 - 1}, e'_{n_1 + n_2 - 1}, e_{n_1 + n_2}\}$ where the e_i and e'_i are defined as above.
   In this case $\mathbf{V}'_l = \bigcup_i \{a_i, \overline{a_i}\}$ is the set of **left vertices**, and $\mathbf{V}'_r = \bigcup_i \{b_i, \overline{b_i}\}$ is the set of **right vertices** of $\mathbf{V}'$. Here it can be seen that there are four more right vertices than left vertices.

Consider the supernatural subgraphs $\{\mathcal{S}_1, \mathcal{S}_{25}, \mathcal{S}_3, \mathcal{S}_4\}$ of Example 2. By means of a relabeling of the vertices (a vertex x_1 could be relabeled as x_2, or vice-versa), one possible suitable order for the edges of the subgraphs is depicted in Figure 3.

In the ensuing discussion, we start with any decomposition of $\mathcal{G}(\mathbf{V}, A)$ into a set $\mathcal{SS}$ of supernatural subgraphs. We then order the vertices and edges of these subgraphs as described above, and partition the vertices of $\mathcal{G}(\mathbf{V}, A)$ into subsets of left and right vertices. (A vertex x is a left vertex in $\mathbf{V}$ if it is a left vertex of a subgraph in $\mathcal{SS}$, otherwise it is a right vertex.)

4.3 Upper bound on the number of cycles of a completed graph

Let $\mathcal{G}(\mathbf{V}, A, D)$ be a completed graph based on $\mathcal{G}(\mathbf{V}, A)$, and let $\mathbf{C}$ be the set of cycles of the graph. The **size of a cycle** is the number of black edges (or similarly of gray edges) contained in the cycle. Let $\mathcal{C}$ be a particular cycle of size r in $\mathbf{C}$, with vertex set $\mathbf{V}_\mathcal{C}$ and with sets of black and gray edges $A_\mathcal{C}$ and $D_\mathcal{C}$, respectively. We define the **signature** $S_\mathcal{C}$ of $\mathcal{C}$ to be the subset of $\mathbf{V}_\mathcal{C}$ derived as follows: For every left vertex x in $\mathbf{V}_\mathcal{C}$, if $\overline{x}$ is not already in $S_\mathcal{C}$, then add x to $S_\mathcal{C}$.

Let $\mathcal{S}$ be the set of signatures of all the cycles in $\mathbf{C}$. Define the **signature graph** $\mathcal{SG}(\mathcal{S}, E)$, where $\mathcal{S}$ is the set

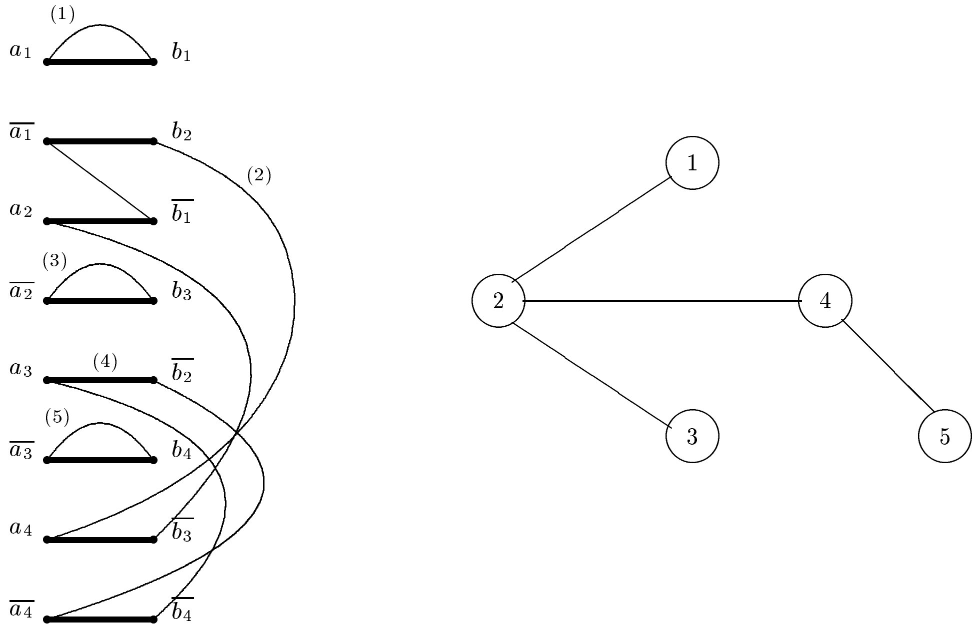

Figure 4: Example of a signature graph

of vertices, and where the set of edges E is defined as follows: for all $S_1, S_2 \in \mathcal{S}$, S_1 and S_2 are linked by an edge in E iff there is a block x such that $x \in S_1$ and $\overline{x} \in S_2$.

In Figure 4, a completed graph is on the left. It represents a completed supernatural subgraph $\mathcal{G}(\mathbf{V}_e, A_e, D_e)$ of some graph $\mathcal{G}(\mathbf{V}, A)$. $\mathcal{G}(\mathbf{V}_e, A_e)$ is a supernatural subgraph in $\mathcal{CE}$.

The left vertices of the graph are the vertices on the left of black edges, that is the a_i and the $\overline{a_i}$, for $1 \leq i \leq 4$.

The completed graph is made up of 5 cycles, whose signatures are as follows:

$$1\text{: } \{a_1\};\quad 2\text{: } \{\overline{a_1}, a_2, a_4\};\quad 3\text{:}\{\overline{a_2}\};\quad 4\text{:}\{a_3, \overline{a_4}\};\quad 5\text{:}\{\overline{a_3}\}.$$

The graph on the right of Figure 4 is the signature graph derived from the graph on the left.

For vertex $S_\mathcal{C}$ in $\mathcal{S}$, denote by $t(S_\mathcal{C})$ the number of elements in $S_\mathcal{C}$ and by $\delta(S_\mathcal{C})$ the number of edges outgoing from $S_\mathcal{C}$.

Lemma 3 *Let $\mathcal{G}(\mathbf{V}_e, A_e) \in \mathcal{SS}$ be a supernatural subgraph of $\mathcal{G}(\mathbf{V}, A)$ of size $2n$, where $n > 0$. Let $\mathcal{G}(\mathbf{V}_e, A_e, D_e)$ be a completed graph and let c_e be the number of cycles in it. Then:*

- *If $\mathcal{G}(\mathbf{V}_e, A_e) \in \mathcal{CE}$, then $c_e \leq n + 1$.*
- *If $\mathcal{G}(\mathbf{V}_e, A_e) \in \mathcal{CO}_-$, then $c_e \leq n$.*

Proof: Let $\mathcal{SG}(\mathcal{S}, E)$ be the signature graph of $\mathcal{G}(\mathbf{V}_e, A_e, D_e)$. Then $c_e = |\mathcal{S}|$.

For all $S_\mathcal{C} \in \mathcal{S}$, $\delta(S_\mathcal{C}) \leq t(S_\mathcal{C})$. Now $\sum_{S_\mathcal{C} \in \mathcal{S}} t(S_\mathcal{C}) \leq 2n$, so that $|E| = \frac{1}{2}\sum_{S_\mathcal{C} \in \mathcal{S}} \delta(S_\mathcal{C}) \leq \frac{1}{2}\sum_{S_\mathcal{C} \in \mathcal{S}} t(S_\mathcal{C}) \leq n$.

A supernatural subgraph is connected, so that

$$|\mathcal{S}| \leq |E| + 1 \leq n + 1.$$

For the case $\mathcal{G}(\mathbf{V}_e, A_e) \in \mathcal{CO}_-$, $\sum_{S_\mathcal{C} \in \mathcal{S}} t(S_\mathcal{C}) \leq 2n - 2$. Indeed, the vertices $\overline{b_{n_1}}$ and a'_{n_1+1} belong to no signature $S_\mathcal{C}$ in $\mathcal{S}$. By the same argument as above,

$$|\mathcal{S}| \leq |E| + 1 = \frac{1}{2}\sum_{S_\mathcal{C} \in \mathcal{S}} \delta(S_\mathcal{C}) + 1 \leq \frac{1}{2}\sum_{S_\mathcal{C} \in \mathcal{S}} t(S_\mathcal{C}) + 1 \leq n. \ \square$$

Theorem 2 *Let $\mathcal{G}(\mathbf{V}, A)$ be a partial graph and $\mathcal{G}(\mathbf{V}, A, D)$ be a completed graph. Let $N_A = \frac{1}{2}|A|$ and c_D be the number of cycles in $\mathcal{G}(\mathbf{V}, A, D)$. Denote by α_p the number of supernatural subgraphs of $\mathcal{CE}$. Then:*

$$c_D \leq \alpha_p + N_A$$

Proof: Let $\mathbf{C}$ be the set of cycles in $\mathcal{G}(\mathbf{V}, A, D)$, and $\mathcal{SG}(\mathcal{S}, E)$ the signature graph associated with $\mathbf{C}$. The set of r connected components of $\mathcal{SG}(\mathcal{S}, E)$ decomposes $\mathcal{G}(\mathbf{V}, A, D)$ into even-sized subgraphs $\{\mathcal{J}_i\}_{1 \leq i \leq r}$, where $\mathcal{J}_i = \mathcal{J}_i(\mathbf{V}_i, A_i, D_i)$. For each of the $\mathcal{J}_i$, let t_i be the sum of the sizes of the signatures of all of its cycles, and let $n_i = \frac{1}{2}|A_i|$.

Let k be the number of the $\mathcal{J}_i$ satisfying $t_i < 2n_i$. Then by the same argument used to prove Lemma 3, we can show $c_D \leq N_A + (r - k)$. Now, $k' = r - k$ is the number of subgraphs satisfying $t_i = 2n_i$. But the maximum number of such graphs is α_p. Thus $c_D \leq k' + N_A \leq \alpha_p + N_A$. $\square$

4.4 Maximal completed graph.

Based on the decomposition of $\mathcal{G}(\mathbf{V}, A)$ into supernatural subgraphs, can we construct a completed graph $\mathcal{G}(\mathbf{V}, A, D)$ having $c_D = \alpha_p + N_A$ cycles? By Theorem 2, this would necessarily be maximal.

We will complete the supernatural subgraphs in $\mathcal{SS}$ one at a time in producing a duplicated genome H. At each step, we denote by $F = \cup_{1 \leq i \leq r}\{f_i, \overline{f_i}\}$ the set of fragments of the genome H resulting from the preceding steps. At the outset, F is made up of the **unitary fragments**, which include not only $x^t x^h$, for all $x \in \mathcal{B}$, but also the $2N$ elements of O.

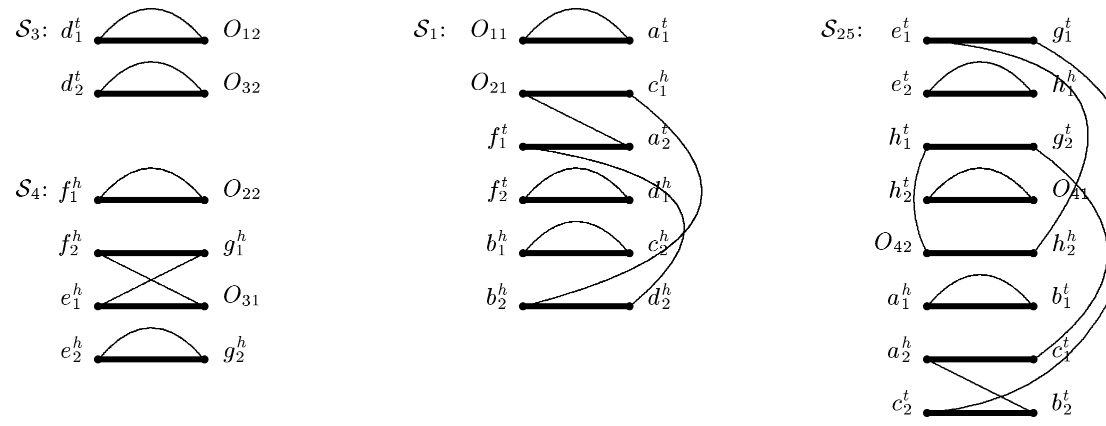

Figure 5: Completed graph $\mathcal{G}(\mathbf{V}, A, D)$ found by algorithm **dedouble**.

As the construction proceeds, whenever a pair of gray edges (x^h, y^t) and $(\overline{x}^h, \overline{y}^t)$ are created, the fragment containing x^h and the one containing y^t are joined together. The final set of fragments contains the $2N$ duplicated chromosomes of the desired genome. A **long fragment** is one that is not unitary. A **terminal fragment** is unitary, consisting of an element of O, or is long, with an extremity in O. **Internal fragments** contain no element of O.

Remark 1 :

- If fragment f is internal, then the only vertices of f not adjacent to gray edges are its two endpoints. If f is terminal, its only vertex not linked by a gray edge is that endpoint in V.
- For all $x \in V$, x and $\widetilde{x}$ are in the same fragment.
- If x and y are two vertices in the same fragment, then $\overline{x}$ and $\overline{y}$ are also in one fragment. In discussing fragment membership, we may speak indifferently of x or $\overline{x}$.

Suppose we have completed the k first supernatural subgraphs of $\mathcal{G}(\mathbf{V}, A)$ and we wish to complete the $(k+1)$-st one, $\mathcal{G}(\mathbf{V}', A')$. Let x, y be two distinct vertices in $\mathbf{V}'$. To be able to construct the gray edge (x, y), we must have $x \neq y$, and conditions I, II below satisfied. These conditions must be satisfied for x or $\overline{x}$ and for y or $\overline{y}$. To simplify notation, we omit $\overline{x}$ and $\overline{y}$.

Condition I. If $x, y \notin O$, then x and y are not in the same fragment. In particular, $x \neq \widetilde{y}$.

Condition II. If x and y are in two different terminal fragments, and if F contains an internal fragment, then F must contain at least two other terminal fragments.

A pair of vertices (x, y) is said to be **possible** if it satisfies these conditions. Otherwise it is **impossible**. If (x, y) is possible, then so are $(x, \overline{y})$, $(\overline{x}, y)$ and $(\overline{x}, \overline{y})$.

We now describe an algorithm for constructing a completed graph $\mathcal{G}(\mathbf{V}, A, D)$. We will not repeat the fact each time a gray edge (x, y) is created, this implies the creation of $(\overline{x}, \overline{y})$.

Algorithm dedouble

We denote by $e(x)$ the black edge incident to vertex x.

$\boxed{\text{Subgraphs in } \mathcal{CE}, n = 1}$

For every subgraph $\mathcal{G}(\mathbf{V}', A')$ of $\mathcal{CE}$ of size $2n = 2$ such that $A' = \{(a_1, b_1), (\overline{a_1}, \overline{b_1})\}$, add edges (a_1, b_1) and $(\overline{a_1}, \overline{b_1})$ to D.

$\boxed{\text{Subgraphs in } \mathcal{CE}, n > 1}$

Let $\mathcal{G}(\mathbf{V}', A')$ be a subgraph in $\mathcal{CE}$ of size $2n$.

If (a_1, b_1) and $e(\overline{b_2})$ are possible, create edge (a_1, b_1).
Otherwise, create edge $(\overline{a_1}, b_2)$.

For $2 \leq i \leq n$:
  If $\overline{b_{i-1}}$ is not already incident to a gray edge,
    If $e(\overline{b_{i+1}})$ is possible, create edge $(a_i, \overline{b_{i-1}})$.
    Otherwise, create edge $(\overline{a_i}, b_{i+1})$.
  If $\overline{b_{i-1}}$ is already incident to a gray edge,
    If $e(b_{i+1})$ is possible, create edge $(\overline{a_i}, b_{i+1})$.
    Otherwise, create edge (a_i, b_j), where b_j is the remaining unlinked vertex in the path containing a_i and $\overline{b_{i-1}}$.

Create (a_n, b_j), where b_j is the remaining unlinked vertex on the path containing a_n.

$\boxed{\text{Subgraphs in } \mathcal{CO}_-}$

For $1 \leq i \leq n_1 - 1$ or $n_1 + 1 \leq i \leq n_1 + n_2 - 2$, gray edges are constructed as above.

After step $i = n_1 - 1$, there remain two vertices, counterparts, not yet linked. Denote these vertices b_r and $\overline{b_r}$.

$i = n_1 + n_2 - 1$:
  If $\overline{b_{i-1}}$ is not already incident to a gray edge,
    If $(b_r, \overline{b_{i+1}})$ is possible, create edge $(a_i, \overline{b_{i-1}})$.
    Otherwise, create the edge $(\overline{a_i}, b_{i+1})$.
  If $\overline{b_{i-1}}$ is already incident to a gray edge, let b_j be the remaining unlinked vertex in path containing a_i and b_{i-1}.
    If $(\overline{a_i}, b_{i+1})$ and (b_r, b_j) are possible, create edge $(\overline{a_i}, b_{i+1})$
    Otherwise, create edge (a_i, b_j)

Create the edge $(\overline{b_r}, b_j)$, where b_j is the vertex not already linked remaining in $\mathbf{V}'_2$.

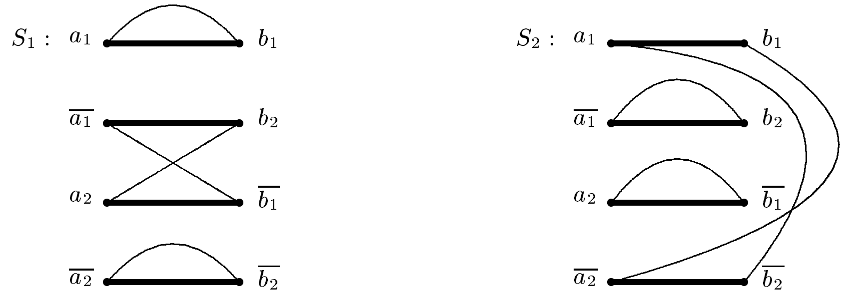

Figure 6: All possible completed subgraphs for $n = 2$.

Lemma 4 *The algorithm is correct, i.e. it produces a completed graph.*

Proof:

$\boxed{\text{Subgraphs in } \mathcal{CE}, n = 1}$

For any supernatural subgraph $\mathcal{G}(\mathbf{V}', A')$ of size $2n = 2$ where $A' = \{(a_1, b_1), (\overline{a_1}, \overline{b_1})\}$, (a_1, b_1) is possible.

$\boxed{\text{Subgraphs in } \mathcal{CE}, n > 1}$

Suppose the current subgraph is $\mathcal{G}(\mathbf{V}', A') \in \mathcal{CE}$, where $n > 1$, and the set of fragments of H constructed to this point is F. Since $A' \in \mathcal{CE}$, it cannot contain any edge of form $(x, \overline{x})$.

1. Suppose first that edge (a_1, b_1) is impossible. We will show that in this case (a_1, b_2) must be possible.

If a_1 and b_1 do not satisfy condition I, i.e., $a_1, b_1 \notin O$ and a_1, b_1 belong to the same fragment. Then a_1 and b_2 cannot contradict Condition I, otherwise a_1, b_1 and b_2 would be in the same fragment, an impossibility because these vertices are not on gray edges. Since $b_1 \notin O$, a_1 is not in a terminal fragment. Condition II is thus satisfied.

Suppose that a_1 and b_1 do not satisfy Condition II. a_1 and b_1 are thus in different terminal fragments. a_1 and b_2 cannot contradict Condition I by the same reasoning as above. Condition II is also satisfied, since if b_2 were in a terminal fragment, there would have to be another one, since the number of terminal fragments is even. In this case, a_1 and b_1 would satisfy Condition II, which is a contradiction.

Similarly, if $e(\overline{b_2})$ is impossible, then (a_1, b_2) is possible.

2. Let $2 < i < n$. Suppose that $\overline{b_{i-1}}$ is not on a gray edge. Note first that because of how we link the vertices, if b_{i-1} is not already linked, it must be that (a_i, b_{i-1}) is possible. Similarly as above, if $e(\overline{b_{i+1}})$ is impossible, then (a_i, b_{i+1}) must be possible.

Suppose now that $\overline{b_{i-1}}$ is already connected by a gray edge and that $e(b_{i+1}) = (\overline{a_i}, b_{i+1})$ is impossible. Let $b_j \neq a_i$ be the vertex on the path containing $\overline{b_{i-1}}$, not yet connected by a gray edge. We must show that (a_i, b_j) is possible.

Suppose that a_i and b_{i+1} do not satisfy Condition I. In other words, $a_i, b_{i+1} \notin O$, and a_i, b_{i+1} are in the same fragment. Then since b_j is not connected by a gray edge, a_i and b_j are not in the same fragment. These two vertices thus satisfy Condition I. On the other hand, since a_i is not in a terminal fragment, a_i and b_j also satisfy Condition II.

Suppose that a_i and b_{i+1} contradict Condition II. Then it is clear that a_i and b_j can contradict neither Condition I nor Condition II.

3. By an analogous argument, (a_n, b_j) must be possible.

$\boxed{\text{Subgraphs in } \mathcal{CO}_-}$

If $\mathcal{G}(\mathbf{V}', A') \in \mathcal{CO}_-$, then the validity of the construction can be proved as in steps 1 – 3 of the preceding case. $\square$

Example 3 *Consider genome G in Example 1, and the decomposition of its graph $\mathcal{G}(\mathbf{V}, A)$ into the supernatural subgraphs of Figure 3. In constructing the completed graph $\mathcal{G}(\mathbf{V}, A, D)$ by our method, we first complete subgraph $\mathcal{S}_3$, then $\mathcal{S}_1$ and $\mathcal{S}_4$, and finally the subgraph $\mathcal{S}_{25}$. Figure 5 depicts the completed graph thus produced.*

The number of cycles in the completed graph is $c_D = 12$. Now, $\alpha_p = 2$ and $|A| = 20$, so that, according to Theorem 2, it is a maximal completed graph.

The corresponding duplicated genome H contains two identical copies of the following two chromosomes:

1: $+a$ $+b$ $-c$ $+h$ $+e$ $-g$;   2: $+d$ $+f$.

Lemma 5 *The algorithm produces a duplicated genome where $\mathcal{G}(\mathbf{V}, A, D)$ has $c_D = \alpha_p + N_A$ cycles, α_p being the number of supernatural subgraphs of $\mathcal{CE}$, and $N_A = |A|/2$.*

Proof: Let $\mathcal{G}(\mathbf{V}', A')$ be a subgraph of $\mathcal{CE}$ of size $2n$ where $n > 1$, and $\mathcal{G}(\mathbf{V}', A', D')$ the completed subgraph obtained by the construction described above. We will show, by induction on n, that $\mathcal{G}(\mathbf{V}', A', D')$ contains $n + 1$ cycles.

For $n = 2$, the only two completed subgraphs of $\mathcal{G}(\mathbf{V}', A')$ that can be obtained are depicted in Figure 6. In both cases, there are 3 cycles in the completed subgraph.

Suppose the induction hypothesis is true for p. The different configurations possible for the cycles containing the last four black edges of the subgraph are depicted in Figure 7.

Let $\mathcal{G}(\mathbf{V}', A')$ be a subgraph of size $2(p + 1)$, such that $\mathbf{V}' = \bigcup_{1 \le i \le p+1} \{a_i, \overline{a_i}, b_i, \overline{b_i}\}$.

Let $\mathbf{V}'' = \mathbf{V}' \setminus \{a_{p+1}, \overline{a_{p+1}}, b_{p+1}, \overline{b_{p+1}}\}$. The subgraph $\mathcal{G}(\mathbf{V}'', A'')$ is of size $2p$. By the induction hypothesis, the completed subgraph $\mathcal{G}(\mathbf{V}'', A'', D'')$ produced by the algorithm contains $n + 1$ cycles, and these cycles must have one of the configurations in Figure 7.

For each configuration in Figure 7, by replacing the black edge $(\overline{a_p}, \overline{b_p})$ by the three black edges $(\overline{a_p}, b_{p+1})$, $(a_{p+1}, \overline{b_p})$ and $(\overline{a_{p+1}}, \overline{b_{p+1}})$, the various subgraphs that may be obtained always contain one more cycle than the initial subgraph $\mathcal{G}(\mathbf{V}'', A'', D'')$.

If $\mathcal{G}(\mathbf{V}', A')$ is a subgraph of $\mathcal{CO}_-$ of size $2n$, we can show, also by induction on n, that the algorithm produces a completed subgraph $\mathcal{G}(\mathbf{V}', A', D')$ containing n cycles $\square$

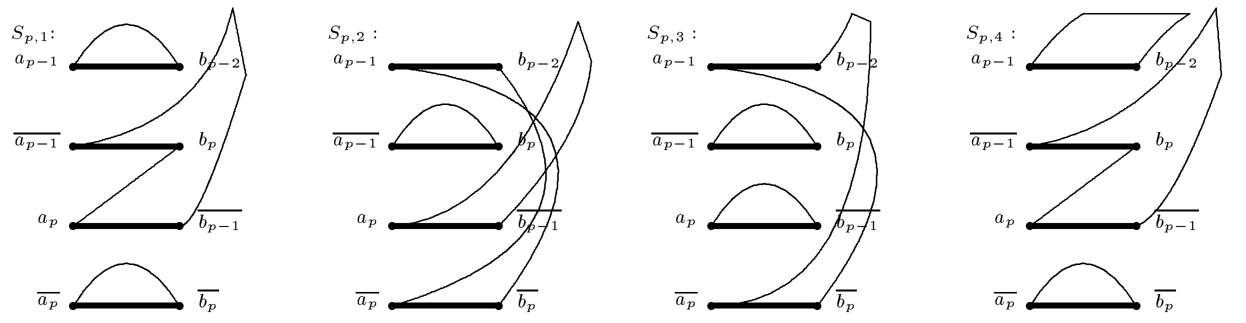

Figure 7: Possible configurations of cycles containing the last four edges of a completed subgraph of $\mathcal{G}(\mathbf{V}', A')$.

The following theorem is a direct consequence of Theorem 2 and Lemma 5.

Theorem 3 *The number of cycles of a maximal completed graph based on $\mathcal{G}(\mathbf{V}, A)$ is*

$$c_D = \alpha_p + N_A$$

5 The centromere

5.1 A constraint on translocations

Normally all chromosomes contain one functional centromere. This constraint is not necessarily satisfied within the framework of the preceding sections, so that a translocation may well result in one chromosome with two centromeres and the other with zero. In this section, we will consider formal ways avoiding such violations of the centromere constraint in the process of finding the ancestral duplicated genome.

Can we reduce the problem with the centromere constraint to a version of the unconstrained formulation? To do this, we define a new representation of genome G and its N chromosomes $(C_i)_{1\leq i\leq N}$. For each i, we write $C_i = C_{i,1}\sigma_i C_{i,2}$, where $C_{i,1}$ represents the part of the chromosome C_i situated to the left of the centromere, $C_{i,2}$ represents the part to the right, and σ_i represents the centromere itself. We term the **divided genome** of G, the set $GP = \{C_{i,1}, -C_{i,2} \text{ for } 1 \leq i \leq N\}$, a set formed of the left arms and the inverted right arms of the chromosomes de G. We will seek the minimum number of translocations necessary to transform GP into a set of duplicated arms HP. Note that this is not a complete solution to our problem, since additional translocations may be necessary to make sure that the duplicates of the right arm and the left arm of a chromosome are also found on a single chromosome. Seoighe and Wolfe [16] also considered this partial solution to reconstructing the pre-doubling genome satisfying the centromere constraint.

As in the case of genomes without considering centromeres, we differentiate between the two occurrences of a block in $\mathcal{B}$, and replace each block x by the pair x^t and x^h. Furthermore, for each arm B_i in GP, we add O_i to its left end, and X_i, representing the centromere, to its right end.

In this formulation, we prohibit prefix-suffix translocations in order to satisfy the centromere constraint. Note that it is biologically coherent to permit translocations which act on the two arms of the same chromosome, so-called pericentric inversions.

In the rest of this section, we will call the arms in a divided genome "chromosomes", and "translocation" will signify a prefix-prefix translocation or a pericentric inversion. We will designate by G a divided genome with $2N$ chromosomes (N being the number of "true" chromosomes in the undivided genome) made up of chromosomes $(C_i)_{1\leq i\leq 2N}$. We define as before the partial graph $\mathcal{G}(\mathbf{V}, A)$ associated with G.

Since we are confined to prefix-prefix translocations, Hannenhalli's result does not necessarily pertain. Nevertheless, it does hold under certain general conditions.

Theorem 4 *Suppose that $\mathcal{G}(\mathbf{V}, A, D)$ is a maximal completed graph of $\mathcal{G}(\mathbf{V}, A)$, and that H, the perfect duplicated genome induced by $\mathcal{G}(\mathbf{V}, A, D)$ has no subpermutations. If for all x in G, blocks x and $\overline{x}$ have the same orientation, i.e., x^t precedes or follows x^h in a chromosome according to whether $\overline{x}^t$ precedes or follows $\overline{x}^h$, then Hannenhalli's algorithm uses only prefix-prefix translocations.*

In the present context, Theorem 4 requires that in the undivided genome (with N chromosomes) any two corresponding blocks have the same orientation with respect to the centromere. We will assume this in adapting the theory of Section 4 to the case of the centromere constraint.

5.2 Subdividing a graph into supernatural subgraphs

We must first distinguish the two ends of our chromosome arms, i.e. the centromere from the telomere. Thus, in addition to O and V we introduce the set $X = \{X_i^t \text{ for } 1 \leq i \leq 2N\}$.The set of edges of $\mathcal{G}(\mathbf{V}, A)$ is now $\mathbf{V} = V \cup O \cup X$.

Lemma 6 *The set of gray edges of a completed graph $\mathcal{G}(\mathbf{V}, A, D)$ must satisfy Conditions 1 and 2a of Lemma 1. For (u, v) an edge of D, Condition 2b becomes:*

- *If $u \in O$ and $v \in V$, then $\overline{v}$ is also linked to an element of O in D.*
- *If $v \in X$ and $u \in V$, then $\overline{u}$ is also linked to an element of X in D.*
- *If $u \in O$ and $v \in X$, then there is another element of X linked to another elements of O.*

Lemma 7 *Let $\mathcal{G}(\mathbf{V}', A')$ be a subgraph of $\mathcal{G}(\mathbf{V}, A)$. To be able to complete this graph with gray edges so as to satisfy the Conditions 1, 2a and 2b cited in Lemma 6, the conditions of Lemma 2 must be replaced by:*

1. *If $u \in \mathbf{V}' \cap V$, then $\overline{u} \in \mathbf{V}'$.*
2. *$\mathbf{V}'$ contains an even number of elements of O, or none. It contains an even number of elements of X, or none.*

3. *Let $\mathbf{V}''$ be the subset of $\mathbf{V}'$ containing the vertices u satisfying one of the following properties:*
   - $u \in O \in X$.
   - *If $u \in V$, then $\widetilde{u} \notin \mathbf{V}'$.*

   The number of elements in $\mathbf{V}''$ is a multiple of four.

A **completable subgraph** becomes one that meets Conditions (1), (2) and (3) of Lemma 7.

We must also extend the definition of a **natural subgraph** $\mathcal{G}(\mathbf{V}_e, A_e)$ generated by $e = (u, v)$ as follows:

- $(u, v) \in A_e$;
- For any edge $(x, y) \in A_e$, if $x \notin O \cup X$, then the edge linking $\overline{x}$ is also in A_e.

Let $\mathcal{G}(\mathbf{V}_e, A_e)$ be a natural subgraph of $\mathcal{G}(\mathbf{V}, A)$ of size n (i.e. $|V_e| = 2n$). Analogously to the case without centromeres, $\mathcal{G}(\mathbf{V}_e, A_e)$ satisfies Condition 1 of Lemma 7, and $\mathcal{G}(\mathbf{V}_e, A_e)$ satisfies Condition 3 of Lemma 7 iff n is even. Moreover, a natural subgraph contains exactly two elements of $O \cup X$, or none.

Lemma 8 *Consider a natural subgraph $\mathcal{G}(\mathbf{V}_e, A_e)$ of size n. Under the orientation hypothesis of Theorem 4, $\mathbf{V}_e$ contains either two elements of O, two elements of X, or no elements at all from $O \cup X$, if n is even. Then $\mathcal{G}(\mathbf{V}_e, A_e)$ meets Condition 2 of Lemma 7. For n odd, $\mathbf{V}_e$ contains one element of O and one of X.*

We conclude that a natural subgraph is completable iff it is of even size.

Let GC be the set of coherent subgraphs of $\mathcal{G}(\mathbf{V}, A)$. We divide GC into the following subsets:

- GCE consists of the completable natural subgraphs (i.e. those of even size).
- GCO contains coherent subgraphs of odd size.

A decomposition of $\mathcal{G}(\mathbf{V}, A)$ into completable subgraphs can be found by pairing off the subgraphs of GCO in an arbitrary way. We denote by $\mathcal{CO}$ the subset thus obtained. The **supernatural subgraphs** are then the subgraphs in $GCE \cup \mathcal{CO}$. Completion of each of these supernatural subgraphs can then be carried out in the same way as in the case without centromeres.

6 Analysing the yeast genome

Wolfe and Shields [18] proposed that yeast is a degenerate tetraploid resulting from a genome duplication 10^8 years ago. They identified 55 duplicated regions, representing 50% of the genome.

6.1 Without centromeres

Applying our algorithm to the yeast genome data [18] in Table 1, we obtain the perfect duplicated genome H in Table 2. The number of cycles of the corresponding completed graph $\mathcal{G}(V, A, D)$ is $c = 81$. Since G (yeast) and G_d do not give rise to subpermutations (in the sense of Hannenhalli [6]), we can deduce that the minimal number of translocations required to transform G into H is

$$t = 2|\mathcal{B}| + |O| - 2N - c = 142 - 16 - 81 = 45.$$

I	:	+2 • −1
II	:	+4 • −3 −7 **+8** −5 +6
III	:	+9 • −10 −11
IV	:	+20 +12 +12 +54 +15 +21 • −3 −13 −16 +17 −24 −22 −14 −23 −19 +18 −9
V	:	+28 • −25 −27 −4 −26 −13
VI	:	+55 • −36
VII	:	+36 +25 +26 +32 **+6** **−33** +5 • −30 −34 −31 −29
VIII	:	+35 • −14 −37 −29 −1
IX	:	+38 +39 +27 •
X	:	+10 +40 +41 • −28 −42
XI	:	+42 +40 +43 +35 • −41 −52 −38
XII	:	+53 • −53 −31 −55 −16 **−18** **−17** −45 −30 −15 −44
XIII	:	+46 +44 +19 • −43 −54 −48 −47 −46
XIV	:	+49 +20 +37 +50 +39 • −11
XV	:	+49 +21 • −22 −52 −50 −23 −45 −51 −47 −2
XVI	:	+48 +32 +33 +51 +8 +24 • −7 −34

Table 1: Order of blocks on each of the 16 chromosomes of the yeast genome. Signs indicate transcriptional orientation. In each chromosome, the • indicates centromere position.

Moreover, since c maximizes the number of cycles of any completed graph and as the number of subpermutations obtained is minimal (equal to zero), t is also the minimal number of translocations that transforms G into any perfectly duplicated genome.

1	:	+2 −1
2	:	+46 +47 +48 +54 +43 +35 −41 −40 −42
3	:	+9 −10 −11
4	:	+44 +15 +21 −22 −14 −23 −19 +18 +16 +13 +26 +32 +33 +51 +45 +17 −24 −8 +7 +3 −4
5	:	+55 −36
6	:	+38 +39 +27 +25 −28
7	:	+29 +37 +50 +52 −53
8	:	+49 +20 +12 +31 +34 +30 −5 +6

Table 2: A solution for the ancestral genome. The present-day yeast genome can be obtained from this one by genome doubling followed by 45 translocations.

6.2 With centromeres

Applying the methodology of Section 5 to the yeast genome, we produce pairs of identical chromosome arms where orientations are conserved with respect to the centromere for all except blocks 6, 8, 17, 18 and 33 (see boldface in Table 1). Three inversions are required to correct these orientations.

Seoighe and Wolfe [16] also considered the problem of producing pairs of duplicated chromosome arms. After the three initial inversions needed to correct orientation, the best solution obtained by their heuristic algorithm is 40 translocations. When applying our method, we find that the minimal number of prefix-prefix translocations that produce 16 pairs of identical chromosome arms is only 38 translocations.

Of course, after producing pairs of identical chromosome arms, there remains the task of ensuring that arms are correctly paired to form duplicated chromosomes.

Discussion

The construction presented in this paper is essentially linear-time in the number of blocks. This gives the ancestral genome, and the number of translocations necessary to derive the modern one from it. Of course, if the actual translocations are needed explicitly, Hannenhalli's cubic algorithm must be utilized.

In maximizing the number of cycles, the minimization of translocations is valid only if the given genome G and the solution genome G_d determine no subpermutations. As mentioned in Section 3, however, the existence of subpermutations is suggestive of the inadequacy of a pure translocational analysis of genomic differences.

Thus, rather than extend our method to take subpermutations into account, which is not only unmotivated but also seems quite difficult analytically, it would be more important to study a combined inversion and translocation version of our problem. This also seems difficult, however.

Another important open problem would see a correct pairing constraint imposed on the duplicate arms constructed in the analysis with centromeres.

References

[1] Ahn, S., Tanksley, S.D.: Comparative linkage maps of rice and maize genomes. Proc. Natl. Acad. Sci. USA **90** (1993) 7980-7984.

[2] Atkin, N. B., Ohno, S.: DNA values of four primitive chordates. Chromosoma **23** (1967) 10–13

[3] El-Mabrouk, N., Nadeau, J.H., Sankoff, D.: Genome halving. Combinatorial Pattern Matching. Ninth Annual Symposium (M. Farach-Colton, ed.) Lecture Notes in Computer Science **1448** (1998) Springer Verlag, 235–250.

[4] Fryxell, K.J.: The coevolution of gene family trees. Trends in Genetics **12** (1996) 364–369.

[5] Gaut, B.S., Doebley,J.F.: DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci., U.S.A. **94** (1997) 6809– 6814.

[6] Hannenhalli, S.: Polynomial-time algorithm for computing translocation distance between genomes. In Combinatorial Pattern Matching. Sixth Annual Symposium (Z. Galil and E. Ukkonen, ed.) Lecture Notes in Computer Science **937** (1995) Springer-Verlag, 162–176.

[7] Hinegardner, R.: Evolution of cellular DNS content in teleost fishes. American Naturalist **102** (1968) 517–523

[8] Muller, F., Bernard, V., Tobler, H.: Chromatin diminution in nematodes. Bioessays **18** (1996) 133–138

[9] Moore, G., Devos, K. M., Wang, Z., Gale, M. D.: 1995. Grasses, line up and form a circle. Current Biology **5** (1995) 737–739.

[10] Nadeau, J. H.: Genome duplication and comparative mapping. In Advanced Techniques in Chromosome Research (ed. Adolph, K.T.) (1991) (Marcel Dekker, New York) 269–296

[11] Nadeau, J.H., Sankoff, D.: Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics **147** (1997) 1259–1266

[12] Ohno, S., Wolf, U., Atkin, N. B.: Evolution from fish to mammals by gene duplication. Hereditas **59** (1968) 169–187

[13] Paterson, A.H., Lan, T.-H., Reischmann, K.P., Chang, C., Lin, Y.-R., Liu, S.-C., Burow, M.D., Kowalski, S.P., Katsar, C.S., DelMonte, T.A., Feldmann, K.A., Schertz, K.F., Wendel, J.F.: Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nature Genetics **14** (1996) 380–382

[14] Postlethwait, J.H, Yan, Y.-L., Gates, M.A., Horne, S., Amores,A., Brownlie, A., Donovan, A., Egan, E.S., Force, A., Gong, Z., Goutel, C., Fritz, A., Kelsh, R., Knapik, E., Liao, E., Paw, B., Ransom, D., Singer, A., Thomson,T., Abduljabbar, T.S., Yelick, P., Beier, D., Joly, J.-S., Larhammar, D., Rosa, F., Westerfield, M., Zon, L.I., and Talbot, W.S.: Vertebrate genome evolution and the zebrafish gene map. Nature Genetics 18 (1998) 345–349.

[15] Scheffler, J. A., Sharpe, A.G., Schmidt, H., Sperling, P., Parkin, I.A.P., Lühs, W., Lydiate, D.J., Heinz, E.: Desaturase multigene families of Brassica napus arose through genome duplication. Theoretical and Applied Genetics **94** (1997) 583–591

[16] Seoighe, C., Wolfe, K.H.: Extent of genomic rearrangement after genome duplication in yeast. Proceedings of the National Academy of Sciences USA **95** (1998) 4447–4452.

[17] Shoemaker, R.C.,Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., Kochert, G. Boerma, H.R.: Genome duplication in soybean (Glycine subgenus soja). Genetics **144** (1996) 329–228

[18] Wolfe, K.H., Shields, D.C.: Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387** (1997) 708–713

[19] Xu, R-H., Kim, J.,Taira, M., Lin, J.J., Zhang, C.-H., Sredni, D., Evans, T., Kung, H.-F.: Differential regulation of neurogenesis by the two Xenopus GATA-1 genes. Molecular and Cellular Biology **17** (1997) 436–443