Genome Halving

Nadia El-Mabrouk[1] and Joseph H. Nadeau[2] and David Sankoff[3]

[1] Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec, H3C 3J7.
elmabrou@crm.umontreal.ca

[2] Department of Genetics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106-4955.
jhn4@po.cwru.edu

[3] Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec, H3C 3J7.
sankoff@ere.umontreal.ca

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 (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 orders. The study of genome duplication based on block data includes the inference of the synteny or linkage structure of the pre-duplication genome, the nature of the post-duplication rearrangement events, and the statistics of gene loss versus functional divergence. We propose a suite of *Genome halving* problems for algorithmic solution, some of which address the evolution of gene order, and others which deal with relations of synteny only. We present an efficient and accurate heuristic for the latter type of problem, and apply it to the genome duplication which has been described for *Saccharomyces cerevisiae*.

1 Genome duplication

In the course of evolution, there are a number of mechanisms which result in the appearance of two or more identical or nearly identical genes in the same genome – a gene family. Mechanisms such as unequal crossing over give rise to adjacent or closely linked copies on the same chromosome. Others, such as reverse transcription, can produce copies on different chromosomes. Whatever the mechanism, the existence of duplicate genes can lead to important changes in the genetic makeup of an organism. Over a relatively short period of time, evolutionary speaking, all but one of the genes in a family may accumulate random mutations and lose their function, i.e. become pseudogenes. Sometimes, however, one or more of the genes in a family incur mutations that lead them to encode for similar, but not identical, products. This is an evolutionary "opportunity" for the organism; it allows for specialization of gene products to function optimally under differing conditions in the life cycle or in different tissues. It can generally speed up evolution in that while one copy of a gene is free to "explore" mutational space to find some new function, the original function, necessary for survival, is preserved by the other copy.

Perhaps the rarest, but surely 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 mecanisms 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 [6] 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, 8, 15]. 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 [13, 14] and for less intensively mapped vertebrate genomes. More recent genome duplications are known to have occurred in some vertebrate lines, such as the frogs [22], the salmoniform fish [15] and zebrafish [17].

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 [10].

Genome duplication is particularly prevalent in plants. Comparison of the well-studied rice [1], oats (wild and domestic), corn [1, 7] and wheat [12] genomes indicate several occurrences in the cereal lineage. Soybeans [20], rapeseed [19], 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 [16].

Recently, following the complete sequencing of all *Saccharomyces cerevisiae* chromosomes, the prevalence of gene duplication [9, 3] has led to the conclusion that this yeast genome is also the product of an ancient doubling [21].

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 orders. The study of genome duplication based on block data includes the inference of the synteny or linkage structure of the pre-duplication genome, the nature of the post-duplication rearrangement events, and the statistics of gene loss versus functional divergence [14]. In this paper we propose a suite of *Genome halving* problems for algorithmic solution, one of which addresses the evolution of gene order, and another which deals with relations of synteny only. We present an efficient and accurate heuristic for the latter problem, and apply it to the genome duplication which has been postulated for *Saccharomyces cerevisiae*.

2 Genome halving with ordered chromosomes

As part of their detailed study of the post-duplication evolution of the *S.cerevisiae*, Seoighe and Wolfe [18] ask how many translocations are necessary to account for the present configuration of paralogous segments in this 16-chromosome genome, starting from a tetraploidization of an ancestral 8-chromosome genome.

The *Genome halving on ordered chromosomes* problem can be formalized most simply as follows. We are given a genome containing n chromosomes $\{c_1, \cdots, c_n\}$, where each chromosome c can be represented as a string of blocks (or segments) $s_{i_1(c)} s_{i_2(c)} \cdots s_{i_{k_c}(c)}$ and where each block occurs exactly twice in the genome, either on two different chromosomes or twice on the same chromosome. Since there is no biologically meaningful reason for assigning a particular left-to-right orientation to a chromosome, the reverse string $\rho(s_{i_1(c)} \cdots s_{i_{k_c}(c)}) = s_{i_{k_c}(c)} \cdots s_{i_1(c)}$ is considered identical to the original string as a representation of the chromosome.

A translocation consists of dividing the string of blocks on a chromosome a into a prefix a_1 and a suffix a_2, and similarly for $b = b_1 b_2$, and from them creating two strings $a_1 b_2$ and $b_1 a_2$ (or $a_1 \rho(b_1)$ and $\rho(a_2) b_2$) to replace the former two chromosomes. (Only one of a_1, a_2, b_1, b_2 may be null.) How many translocations does it take to replace the n given chromosomes with n reconstructed ones, such that there are $n/2$ pairs of identical chromosomes?

There are a number of variants of this problem. First, in the *Genome halving on ordered chromosomes with centromeres* problem each string is punctuated with a special symbol s_{ce} (the centromere). The translocations are constrained so that each new string created must also contain exactly one s_{ce}.

Second, in the *Genome halving with variable chromosome number* problem, the number n' of reconstructed chromosomes is not fixed at n, but is free to vary among the even integers, in the search for a minimizing number of translocations. All that is required is that the reconstructed genome contain $n'/2$ pairs of identical chromosomes.

Third, where the centromere is taken into account, we can define *Genome halving with oriented blocks* problem. Here the data for each block indicates whether it has positive polarity (oriented away from the centromere) or negative polarity (oriented towards the centromere). Since translocation does not change the polarity of blocks, to solve this problem it is necessary to allow reversals of substrings of a chromosome. For strings not containing the centromere (paracentric reversals), the reverse string of $s_1 \cdots s_m$ is $R(s_1 \cdots s_m) = -s_m \cdots - s_1$, where the minus sign indicates a change of polarity. For strings containing the centromere (pericentric reversals), the reverse string is $R(s_1 \cdots s_{ce} \cdots s_m) = s_m \cdots s_{ce} \cdots s_1$. The objective function to be minimized is the number of translocation plus the number of reversals, possibly with different cost coefficients for the two types of operation. Note that in this problem, we still have the property that the reversal of the entire chromosome $R(s_{i_1(c)} \cdots s_{i_{k_c}(c)}) = \rho(s_{i_1(c)} \cdots s_{i_{k_c}(c)}) = s_{i_{k_c}(c)} \cdots s_{i_1(c)}$ since this is necessarily a pericentric reversal.

It is possible to define oriented genome halving problems without taking into account the centromere, but these are not directly interpretable biologically. The reason for this is that the left-to-right labeling of the blocks in a chromosome is no longer innocuous. In contrast to all the previous versions of the problem, reversing a chromosome here changes its nature; every block changes polarity, there being no pericentric reversals. The unfortunate effects of this can be seen in the results of a translocation – newly adjoining blocks from the two contributing chromosomes will have the same or different polarities depending on the arbitrary choice of which ends of these chromosomes were left and which were right. This is not realistic, since whether or not two newly adjoining blocks have the same polarity (i.e. are on the same DNA strand) is predictable from knowledge of their polarity in the original chromosomes.

Thus in genome halving problems incorporating increasing degrees of biological information, centromere location should be included before orientation.

Seoighe and Wolfe devised a heuristic algorithm to solve the ordered genome halving problem, though it is not clear in [18] which version of the problem is addressed. When applied to the yeast data involving 55 segments each in two copies, distributed over 16 chromosomes, the best solution they found was 41 translocations. It is important to note that they did not feel their solution was directly interpretable in terms of evolutionary history, since it severely underestimated the amount of evolution suggested by statistical analysis of ancillary data. Nevertheless, the optimization type of analysis, aside from its intrinsic algorithmic interest, provides an important baseline for evaluating more statistically-oriented approaches.

3 Genome halving with unordered chromosomes

The significance of an analysis based on the order of the blocks on the chromosomes depends on the extent that this order is affected solely by translocation. If in the course of evolution, these blocks were repeatedly shuffled by processes of inversion (reversals) and transposition, a possibility that Seoighe and Wolfe discount (based on meaningful but not overwhelming evidence), then block order might be only indirectly related to translocational history. Then it might be of interest to analyze the data based only on synteny data: which blocks are on which chromosome; a set-theoretical formulation rather than the string theory formulation in the previous section.

The problem we will discuss is *Unordered genome halving with variable chromosome number*, though the algorithm we propose produces a solution with fixed n. It would be both meaningful and feasible to include centromere considerations to the problem with unordered chromosomes, but not questions of block orientation, since reversal operations require taking into account block adjacency and other order information.

Thus we consider a genome G to be a collection of n subsets $S_1, \cdots, S_n$ of a set $B = \{b_1, \cdots, b_k\}$ containing k elements (blocks of genes). We suppose that each block in B appears exactly twice among these n subsets (including the possibility that a block has "multiplicity" 2 within a single subset and appears in no other subset). We refer here to these subsets both as chromosomes and as synteny sets.

The problem is to find the minimum number of translocations necessary to transform G into a genome G' made up of two identical copies of $n'/2$ chromosomes, where n' is even.

Notation :

- Let $S_1,\ S_2,\ T_1,\ T_2$ be four sets such that $S_1 \bigcup S_2 = T_1 \bigcup T_2$ and at most one of these sets is empty. An operation of form $(S_1, S_2) \longrightarrow (T_1, T_2)$ is called a *translocation* of S_1 and S_2, in the sense of [5].
  - If none of the sets is empty, it is a *strict translocation*.
  - If S_1 or S_2 is empty, it is a *fission*.
  - If T_1 or T_2 is empty, it is a *fusion*.
- We define a *duplicated genome* to be made up of two identical copies of $n'/2$ chromosomes.
- For a genome G, we define $D(G)$ to be the *halving cost* of G, the minimal number of translocations necessary to transform G into some duplicated genome.
- Given a genome G, we call a sequence $(\sigma_1, \cdots, \sigma_t)$ of $t = D(G)$ translocations, which transform G into a duplicated genome, an *optimal sequence of translocations* for G.

We define the *intersection graph induced by G* to be $\mathcal{S}_G = (U, E)$ where the vertex set is $U = \{1, \cdots, n\}$ and the edge set E satisfies $(i, j) \in E$ if and only if $S_i \bigcap S_j \neq \emptyset$ (for $i \neq j$), or S_i contains a block of multiplicity 2 (for the case $i = j$).

For the purposes of the genome halving problem, the intersection graph $\mathcal{S}_G$ is equivalent to G. In fact, if two synteny sets contain more than one block in common, then it suffices to designate any one of these blocks as a "representative" for the purposes of analyzing potential translocations.

For all i, $1 \leq i \leq n$, let B_i be the subset of B defined by $b \in B_i$ if and only if for some $j \neq i$, the designated element of $S_i \bigcap S_j$ is b, or else b has multiplicity 2 within S_i. We label each vertex i of $\mathcal{S}_G$ with B_i.

Then, let $\sigma = (S_i, S_j) \longrightarrow (S_i', S_j')$ be any translocation that operates on two chromosomes S_i and S_j of G, and transforms them to chromosomes S_i' and S_j'. This translocation corresponds to moving a number of blocks from B_i to B_j, and from B_j to B_i, and subtracting and adding appropriate edges to the graph. Henceforth, we need consider only the sets B_i, rather than the synteny sets S_i.

3.1 Properties of translocations

For each vertex (chromosome) i of $\mathcal{S}_G$, we define the set $C_i = \{j, (i,j) \in E\}$ to contain those chromosomes j which have at least one block in common with i.

Let e be the number of edges in $\mathcal{S}_G$. Suppose $\sigma = (\{B_{i_1}, B_{i_2}\}, \{B_{j_1}, B_{j_2}\}) \longrightarrow (\{B_{i_1}, B_{j_1}\}, \{B_{i_2}, B_{j_2}\})$ is a translocation of chromosomes i and j. Let $\mathcal{S}_G' = \sigma(\mathcal{S}_G)$ be the graph obtained from $\mathcal{S}_G$ by applying σ, and e' the new number of edges. Then $e' \leq e$; a translocation can only reduce or maintain the number of edges. This is a consequence of our statement of the problem where we considered the blocks of genes in B as the smallest units to be manipulated. In general, in a description of genome evolution through translocation, we cannot define in advance which blocks of genes will be uninterrupted by translocation, so that a translocation may well increase the number of edges in an intersection graph representation, by subdividing a block. In the present context, however, every time a block is transferred from one chromosome to another, the whole block is transferred, and any edge induced by this block, say between a third chromosome l and either chromosome i or j, becomes an edge between l and one of the new chromosomes formed by the translocation.

If the third chromosome l was connected to both i and j, it could be that the number of edges is actually reduced by the translocation. In effect, $e' < e$ if and only if there are two blocks $b_1 \in \{B_{i_1}, B_{i_2}\}$, $b_2 \in \{B_{j_1}, B_{j_2}\}$ satisfying $\{b_1, b_2\} \in \{B_{i_1}, B_{j_1}\}$ or $\{b_1, b_2\} \in \{B_{i_2}, B_{j_2}\}$ and there is some chromosome $l \in C_i \cap C_j$ such that $\{b_1, b_2\} \in B_l$. In other words, a translocation which reduces the number of edges is one which increases the size of intersections between chromosomes.

The translocation σ maximally decreases the number of edges only if $C_i \cap C_j$ is maximal.

3.2 Some bounds

Let G be a genome with n chromosomes. The following is a trivial consequence of the fact that a fusion reduces the number of chromosomes by 1.

Lemma 1. *Any genome G' obtainable from G through $D(G)$ translocations contains $n' \geq n - D(G)$ chromosomes.*

Thus for small values of $D(G)$, n' cannot be much smaller than n. For biological realism, we expect n' to be close to n. This is guaranteed on the upper side by:

Lemma 2. *There exists a duplicated genome G' obtainable from G through a sequence of $D(G)$ translocations, such that $n' \leq n$.*

Proof. We proceed by induction on $D(G)$. If $D(G) = 1$, let σ be the translocation that transforms G into a duplicated genome G'. If σ is a fusion or a strict translocation, then $n' \leq n$ since these do not increase the number of chromosomes. If σ is a fission, then to arrive at a duplicated genome we must have three chromosomes of form $S_1 = (T_1, T_2), S_2 = T_1, S_3 = T_2$, and σ must be the fission $(T_1, T_2) \longrightarrow T_1, T_2$ operating on S_1 and dividing it into two chromosomes containing T_1 and T_2, respectively. We can replace this fission by fusion $T_1, T_2 \longrightarrow (T_1, T_2)$ operating on the two chromosomes S_2 and S_3. This produces a duplicated genome without increasing the number of chromosomes.

Suppose now that the induction hypothesis is true up to $t - 1$, and suppose that $D(G) = t$.

DasGupta *et al.* [4] proved that if there exists an optimal sequence of translocations for transforming one genome into another, then there exists another sequence, containing the same number of fusions, fissions and strict translocations, such that the fissions are ordered after all the translocations and fusions.

Then there exists an optimal sequence of translocations $\sigma = (\sigma_1, \cdots, \sigma_t)$ which transforms G into a duplicated genome G', such that σ_1 is a fusion or a translocation. Let $G_1 = \sigma_1(G)$ and let n_1 be the number of chromosomes de G_1. We have $n_1 \leq n$, and by the induction hypothesis $n' \leq n_1$. Thus $n' \leq n$.

An upper bound for the minimum number of translocation $D(G)$ follows directly from the observation that a trivial duplicated genome can be obtained through $n - 1$ fusions followed by a single fission. Thus

Lemma 3.

$$D(G) \leq n$$

Define the C_i as above. Then a lower bound is given by

Lemma 4. *Let e be the number of edges of $\mathcal{S}_G$ and p the size of the largest intersection $C_i \cap C_j$. Then*

$$D(G) \geq \left\lceil \log_2 \left(\frac{e - \frac{n}{2}}{p} + 1 \right) \right\rceil$$

Proof. The maximum number of edges in the graph of a duplicated genome is $n/2$, so that at least $e - n/2$ edges must be removed.

Suppose there exists optimal sequence of translocations $(\sigma_1, \cdots, \sigma_t)$, where $t = D(G)$, such that at each step i, σ_i is a translocation whch removes a maximum number of edges. For each i, let r_i ne the number of edges removed by σ_i. By definition, there are no translocations removing more than r_i edges. Thus σ_i may add edges so that the next translocation removes at most $2r_i$ edges. Thus $r_{i+1} \leq 2r_i$. Note that σ_1 must remove p edges.

Let $r \geq 0$ the smallest integer $2^r - 1 \geq \frac{e-\frac{n}{2}}{p}$. Then it follows from the above that $D(G) \geq r$. Now, $r = \left\lceil \log_2 \left(\frac{e-\frac{n}{2}}{p} + 1 \right) \right\rceil$, and our result follows immediately.

For the yeast genome with $n = 16$ chromosomes, based on the 55 blocks found by Seoighe et Wolfe [18], the corresponding graph $\mathcal{S}_G$ contains $e = 40$ edges and $p = 5$. Lemmas 3 and 4 assure us that $3 \leq D(G) \leq 16$.

4 An algorithm

In this section, we will deal with the case where n is even, and we wish the duplicated genome to contain n chromosomes also. The data typically contain many more than $n/2$ duplicated blocks, and may be represented by a graph whose vertices represent chromosomes and whose edges connect vertices where the two chromosomes share at least one block. The goal is the efficient transformation of the graph to a matching bipartite graph (with $n/2$ edges), by eliminating (and occasionally adding) appropriate edges through translocation.

4.1 The data

Let $CH = \{c_1, \cdots, c_n\}$ be the set of chromosomes and $BL = \{b_1, \cdots, b_k\}$ be the set of gene blocks, each belonging to exactly two chromosomes (or to a single chromosome, but with multiplicity 2). Let $S_1, \cdots, S_n$ be the synteny sets of chromosomes $c_1, \cdots, c_n$, respectively.

4.2 Initialization

Construct the set BR of subsets of BL, such that each element of BR is either the intersection of two synteny sets, or a subset of form $\{b\}$ where b is of multiplicity 2 within a single chromosome. A block (any one) of an element of BR is chosen as its representative. In the course of the algorithm, some elements of BR will be amalgamated, and representatives designated anew.

Construct the set E of edges of the intersection graph. Each edge e of E is defined by two chromosomes c_i, c_j and an element of BR represented by some block b. We write $e = (c_i, c_j, b)$

For all $c \in CH$, let $C_c = \{c_j \in CH \mid e = (c, c_j, b) \in E \text{ for some } b \in BR\}$.

The intersection graph is completely determined by E. During the execution of the algorithm, we will denote by $\mathcal{G}(E)$ the current graph.

4.3 Results

When the algorithm stops, the results are contained in the variables F, BR and T, where F is the edge set of the output graph containing $n/2$ independent edges, BR is the final set of intersections and T is the sequence of translocations which have been used. Each translocation is represented by a quadruplet $t = (c, c', \mathcal{B}, \mathcal{B}')$, where $c, c' \in CH$, $\mathcal{B}, \mathcal{B}' \subset BR$, indicating the movement of $\mathcal{B}$ from c to c', and the movement of $\mathcal{B}'$ from c' to c. Only one of the two sets of blocks $\mathcal{B}$ or $\mathcal{B}'$ can be missing. In this case, we write $\mathcal{B} = \emptyset$ or $\mathcal{B}' = \emptyset$.

4.4 Description of the algorithm

After initializing $F = \emptyset$, the first part of the algorithm tries to maximally reduce the number of edges in $\mathcal{G}(E)$. At each step we search for the translocation which removes the most edges. By a **fan** of size r, we denote a variable of form $f = (\{c, c'\}, \{c_1, c_2, \cdots, c_r\})$, where $r > 1$, and $c, c', c_1, \cdots, c_r$ are vertices of $\mathcal{G}(E)$ such that $\{c_1, \cdots, c_r\} \subset C_c \cap C_{c'}$. For all i, $1 \leq i \leq r$, suppose $e_i = (c, c_i, b_i)$ and $e'_i = (c', c_i, b'_i)$ the two edges of E linking c to c_i and c' to c_i, respectively.

A translocation removes a maximum number of edges if and only if it operates on a fan $f = (\{c, c'\}, \{c_1, c_2, \cdots, c_r\})$ satisfying $\{c_1, \cdots c_r\} = C_c \cap C_{c'}$ and $|C_c \cap C_{c'}|$ is maximal, and if it is of form $t = (c, c', \mathcal{B}, \mathcal{B}')$, where $\{b_1, \cdots, b_r\} \subset \mathcal{B} \cup \mathcal{B}'$. Such a translocation removes exactly r edges. We will only consider translocations such that $\mathcal{B} \cup \mathcal{B}' = \{b_1, \cdots, b_r\}$.

In particular, for all i, $1 \leq i \leq r$, one of the two edges e_i, e'_i is removed. In order to keep as many fans as possible, we keep the edge which is involved in the structure of the largest number of fans. Suppose that the edge e'_i is removed (i.e. the edge e_i is maintained). This means that the translocation t moves block b'_i from chromosome c' to chromosome c. In this case, we amalgamate the subsets corresponding to blocks b_i and b'_i in BR and we designate b_i to be the representative of this new subset (see Figure 1).

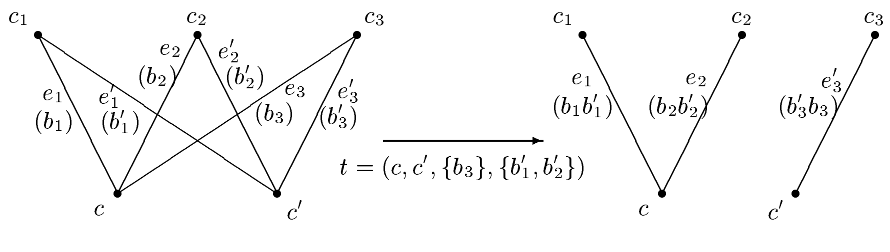

Fig. 1. Translocation of type 1 removing three edges.

As long as $\mathcal{S}(E)$ still contains fans, we choose one of maximal size. If there are two such, two other choice criteria come into play:

Since we are trying to construct a matching bipartite graph of size $n/2$, we choose a translocation which maintains perfect matching (assuming perfect matching already holds). A corollary of Tutte's Theorem is that a necessary condition for a graph to constitute a perfect matching is that it have no connected component with an odd number of vertices. We thus try to find a fan such that the corresponding translocation results in a graph $\mathcal{S}(E)$ containing no connected component of odd size.

If several fans of maximal size satisfy the previous condition (or none do), we choose one such that the corresponding translocation maintains as many fans as possible. To implement this, for each fan $f = (\{c, c'\}, \{c_1, c_2, \cdots, c_r\})$ and each $i \le r$, we calculate the scores s_i and s'_i counting the number of fans other than p, containing the edges e_i and e'_i, respectively. A score s_p for the whole fan p is derived by summing the scores of all edges it would maintain and subtracting the scores of all it would remove. We choose one with the highest score.

A translocation on a fan of size r will be termed a type 1 translocation removing r edges.

Up to now in the algorithm, we have been primarily concerned with reducing the number of edges as quickly as possible. In addition, our reduction of fans is designed so that the remaining edges are potentially (though not necessarily) elements in the ultimate solution to the problem. When there are no more fans, we are left only with translocations which decrease the number of edges of the graph by at most 1. Indeed, it is always possible to find a translocation which will remove an edge and not create any more. However, since our final goal is a graph with exactly $n/2$ independent edges, translocations which create new edges between chromosomes may be necessary to set up pairings for the final duplicated genome.

The second part of the algorithm consists of trying to identify as many edges as possible in $\mathcal{S}(E)$ destined to be contained in the final graph, in order that a minimum of further translocations will be required. This is essentially the classical maximum matching problem of graph theory. Edmonds [11] gave the first polynomial algorithm for finding a maximum matching in a non-bipartite graph, based on the technique of "shrinking" certain odd cycles.

We use this method to find the largest possible number of edges, which are then added to F, the set of edges in the final graph. Each edge $e = (c, c', b)$ so added to F automatically establishes the pairing of chromosomes c and c'.

For the remainder of the algorithm, we will need a set CC consisting of all currently paired chromosomes in CH. At the outset of the second part of the algorithm, $CC = \emptyset$. At the end of the algorithm, we require $CC = CH$. For all $c \in CC$, we define $\overline{c}$ to be the chromosome paired with c. We have $\overline{\overline{c}} = c$

After this step, the edges $e = (c, c', b)$ still in E/F are of three types:

1. $c = c'$ and $c \notin CC$;
2. $c \neq c'$ and only one of the two chromosomes c and c' is already paired, i.e. belongs to CC;
3. The two chromosomes c and c' are already paired but not with each other, i.e. $c, c' \in CC$ and $\overline{c} \neq c'$.

The first two types of edge allow us to pair the as yet unpaired chromosomes.

We group (arbitrarily) all the edges of type (1) and (2), e.g. $e_1 = (c_{11}, c_{21}, b_1)$ and $e_2 = (c_{12}, c_{22}, b_2)$, where $c_{11}, c_{12} \notin CC$. Because n is even, there are either zero, or an even number, of such edges. We carry out the translocation $t = (c_{12}, c_{21}, b_2, b_1)$. This translocation removes edges e_1 and e_2 and creates edges $e = (c_{11}, c_{12}, b_1)$ and $e_0 = (c_{21}, c_{22}, b_2)$. At the same time, we pair chromosomes c_{11} and c_{12} (Figure 2). We call this a translocation of type 2.

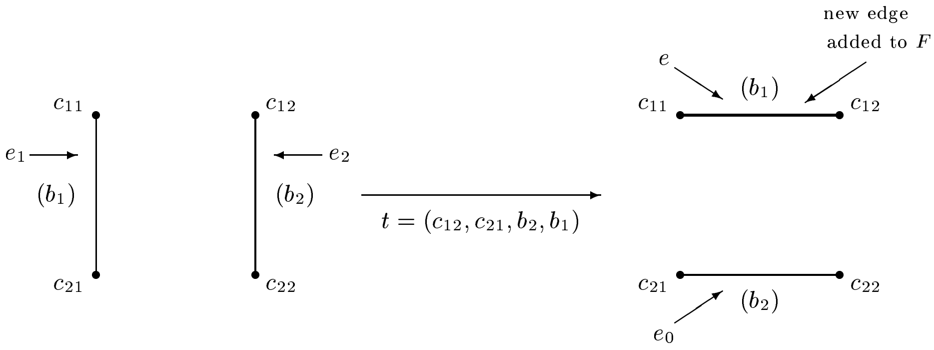

Fig. 2. Translocation of type 2. Two edges are removed, two edges are created, and one of these new edges is added to the final edge set F.

Finally, only edges of type (3) remain to be removed. To do this, we apply translocations which remove one edge of $\mathcal{G}(E)$ at a time (Figure 3). Let $e_1 = (c_{11}, c_{21}, b_1) \in E$ be an edge of type (3) and let $e = (c_{11}, c_{12}, b)$ be the edge such that $\overline{c_{11}} = c_{12}$. Then, the translocation $t = (c_{12}, c_{21}, \emptyset, b_1)$ removes the edge e_1. Here, subsets corresponding to blocks b and b_1 have to be amalgamated. We call this a translocation of type 3.

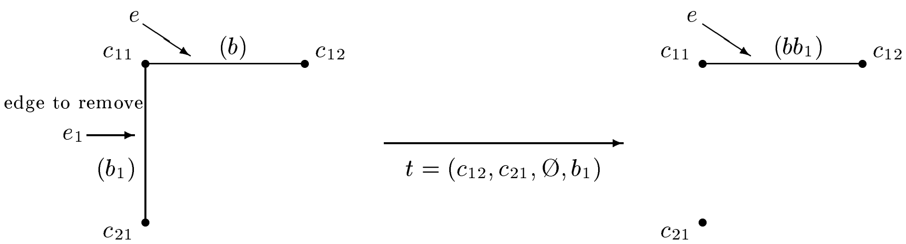

Fig. 3. Translocation of type 3. One edge is removed.

4.5 Procedures

In our implementation, the following routines carry out the steps described above.

Decomposition into fans : The procedure **fan_max** decomposes the graph $\mathcal{G}(E)$ into fans and returns, if possible, an appropriate fan of maximal length.

Processing a fan : The fan $p = (\{c, c'\}, \{c_1, \cdots, c_r\})$ output by the procedure **fan_max** is processed by the procedure **analyze_fan** as in Figure 1.

Choosing a subgraph : The procedure **edmonds** chooses a maximal matching of the graph $\mathcal{G}(E)$.

Pair remaining chromosomes : The procedure **pair_rem_chro** completes the pairing of the as yet unpaired chromosomes as in Figure 2.

Remove remaining edges : The procedure **remove_rem_edges** completes the removal of edges belonging to E/F as in Figure 3.

4.6 The main program

The procedures described in the previous section are used in the main program below. The procedure **fan_max** returns the value T if $\mathcal{G}(E)$ contains a fan.

```
fan_exist = T;
while fan_exist do
        fan_exist = fan_max(fan);
        if fan_exist = T do
               analyze_fan(fan);
end while
edmonds;
pair_rem_chro;
remove_rem_edges;
```

5 Analyzing the yeast data

The data we analyzed, drawn from [21], are listed in Table 1. The Roman numerals are standard notation for the 16 *S. cerevisiae* chromosomes, and the lists of blocks present in each chromosome are numbered according to the Wolfe and Shields notation. It should be noted that the fact that 55 blocks were detected is a function of the criteria and procedures used for assessing similarities between genes and defining blocks (i.e. BLASTP scores ≥ 200 for each paralogous gene pair, at least three genes per block, no two more than 50 kilobases apart, and conservation of gene order and orientation, aside from some short reversals). Conceivably a good number of additional blocks could be added by relaxing some of these.

Our heuristic found a number of solutions to unordered genome halving using only 13 translocations. One of these solutions is described in Tables 2 and 3.

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

Table 1. Lists of blocks corresponding to each of the 16 chromosomes of the yeast genome.

Table 2 contains a possible ancestral genome (note that the numbering of the chromosomes is only very indirectly related, through the algorithm, to the labels used for modern *S.cerevisiae*), and Table 3 traces the algorithm in reconstructing the actual yeast genome from the duplicated genome. The different solutions arose through different choices in the algorithm when there were several possible, such as when several fans had equivalent scores.

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

Table 2. One possible duplicated genome. The actual yeast genome is obtained from this one using 13 translocations.

To be as certain as possible that our solution is optimal, we carried out the following tests: at each step, from among all the translocations which remove the maximum possible number of edges, choose one *randomly* instead of using our choice criteria. Table 4 gives the results in terms of the number of translocations required in 1000 trials on the yeast data set.

Chro. A	*Chro. B*	*Blocks A*	*Blocks B*	*Type of translocation*
7	12	55		3
11	12	53		3
1	15	45 44 15 16 17 18 30 31		3
8	12	45 44 15 16 17 18 30 31	35 52 43	2
10	4	13 25 26 4 27 38	10	1, removes 2 edges
9	4	20 49 50 37 12 21 22 23 19 54 14 29 3 5 6	27 38	1, removes 2 edges
5	11	38 40 41 42		1, removes 2 edges
14	4	12 21 22 23 19 54 14 29 3 5 6	11	1, removes 2 edges
4	2	3 5 6 4	24 32 33 34 51 48	1, removes 3 edges
15	4	15 16 17 18 30 31	21 22 23 19 54 14 29 49 50 37 51 48	1, removes 4 edges
15	8	14 29 37	2 52 43	1, removes 4 edges
15	13	19 54 43 44 46 48		1, removes 5 edges
4	7	5 6 25 26 29 30 31 32 33 34		1, removes 5 edges

Table 3. Translocations applied in reconstructing the current yeast genome from the duplicated genome of table 2. The first line should be read as Block 55 transferred from current chromosome number 7 (initialized as *S.cerevisiae* chromosome VII), to chromosome number 12, with nothing transferred in the other direction, according to a translocation of type 3.

Translocations required	*Frequency*
13	35
14	194
15	268
16	418
17	85

Table 4. Number of translocations required in 1000 trials, using randomized choice of maximal translocation at each step.

6 Discussion and Conclusions

How are we to interpret our solution of 13 translocations, compared to that of [18] with 41? There is little danger in assuming that both results are optimal for their respective problems, which though not mathematically guaranteed, is likely given that they are each based on many runs of a locally optimal procedure. The major reason for the different scores is of course the difference between the unordered and ordered versions of the problem.

This again brings up the question of which problem is more appropriate. Seoighe and Wolfe argue that there is little evidence that reversals have played a major role in scrambling the yeast chromosome. However, if there were as many as 70 or 80 translocations, plus a few reversals, we could expect the blocks on any given chromosome to be paired with blocks from a random sequence of the 16 chromosomes. If this were the case, it would be vanishingly unlikely to find a pattern such as the four blocks out of eleven on Chromosome XII paired with blocks on a single other chromosome (IV), and three occurrences of three blocks in common between pairs of chromosomes. Another way of looking at this is that under the random model, we could expect about six cases where two blocks occur on the same pair of chromosomes. In fact, there are 21 such pairs.

These results suggest the need for unbiased methods to evaluate the relative rate of translocations versus reversals. Perhaps the *Ordered genome halving with oriented blocks* problem would be useful here, with some way of optimizing the relative costs of the two rearrangement operations. In the same way as unordered genome halving is related to synteny distance [5, 4], the ordered genome halving problems have obvious relationships with minimal rearrangement distance problems, and techniques for solving the latter may have some implications for the problems enunciated here.

7 Acknowledgements

Research supported by grants to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Genome Analysis and Technology program, and a CRM postoctoral fellowship to N E-M. DS is a Fellow of the Canadian Institute for Advanced Research.

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. Coissac, E., Maillier, E., Netter, P.:A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Molecular Biology and Evolution **14** (1997) 1062–1074
4. DasGupta,B., Jiang, T., Kannan, S., Li, M., Sweedyk, Z.: On the complexity and approximation of syntenic distance. RECOMB 97. Proceedings of the First Annual International Conference on Computational Molecular Biology (1997) ACM Press, 99-108
5. Ferretti, V., Nadeau, J.H., Sankoff, D.: Original synteny. In Combinatorial Pattern Matching. Seventh Annual Symposium (D. Hirschberg and G .Myers, ed.) Lecture Notes in Computer Science **1075** (1996) Springer Verlag, 159–167
6. Fryxell, K.J.: The coevolution of gene family trees. Trends in Genetics **12** (1996) 364–369.

7. 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.
8. Hinegardner, R.: Evolution of cellular DNS content in teleost fishes. American Naturalist **102** (1968) 517–523
9. Mewes, H.W., Albermann, K., Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., Pfeiffer, F., Zollner, A.: Overview of the yeast genome. Nature **387**(suppl.) (1997) 7–65
10. Muller, F., Bernard, V., Tobler, H.: Chromatin diminution in nematodes. Bioessays **18** (1996) 133–138
11. Lovász, L., Plummer, M.D.: Matching Theory. Annals of discrete mathematics **121** (1986) 357–369
12. Moore, G., Devos, K. M., Wang, Z., Gale, M. D.: 1995. Grasses, line up and form a circle. Current Biology **5** (1995) 737–739.
13. 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
14. 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
15. Ohno, S., Wolf, U., Atkin, N. B.: Evolution from fish to mammals by gene duplication. Hereditas **59** (1968) 169–187
16. 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
17. 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.
18. 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.
19. 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
20. 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
21. Wolfe, K.H., Shields, D.C.: Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387** (1997) 708–713
22. 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