Hybridization and Genome Rearrangement

Nadia El-Mabrouk[1] and David Sankoff[2]

[1] Département d'informatique et de recherche opérationnelle, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec, H3C 3J7.
mabrouk@iro.umontreal.ca

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

Abstract. We infer post-hybridization rearrangements in a hybrid genome, given the gene orders on its chromosomes and some knowledge of the two parent genomes. We study this in two biologically and computationally different contexts, genome fusion and interspecific fertilization. Exact algorithms are furnished for some cases, and a heuristic based on the Hannenhalli-Pevzner theory for another.

1 Introduction

An important mechanism for the rapid emergence of a new, qualitatively different species is the hybridization of two existing species. These parent species will generally be fairly closely related, but may have very different phenotypic expressions. There are actually several types of biological processes that give rise to hybrids, and these are perhaps most widespread in the plant kingdom. In this paper, we explore two such processes – genome fusion and interspecific fertilization. In the first case we give an exact, linear time algorithm for reconstructing the ancestral hybrid from knowledge of the modern genome and data about which gene came from which parent species. We then introduce additional data, on parental species gene order, and try to reconstruct two stages of hybrid genome evolution, intra- and intergenomic (referring to the haploid components originating from the two parents). We adapt the techniques of Hannenhalli and Pevzner [2, 3] in a heuristic for separating these stages and give upper and lower bounds for the optimal transition point between them.

In the case of interspecific fertility, we hypothesize that a key stage in the stabilization of the hybrid genome can be found by calculating the median of three diploid genomes, the two parents and the hybrid. We refer to a reduction of this problem [5] to the Traveling Salesman Problem.

Definitions.

A **genome** G is a collection of N **chromosomes** $G_1, \cdots, G_N$. A chromosome is a string of **signed** ($+$ or $-$) elements from a set $\mathcal{E}$ of **genes**. Each gene in $\mathcal{E}$ appears exactly once in the set of N chromosomes.

For string $X = x_1, \cdots, x_m$, we write $-X$ for the inverted string $-x_m, \cdots, -x_1$. We define the following **rearrangement operations** as in Figure 1: **Inversion**, (or reversal) where any proper substring of a chromosome is inverted. (Inverting the entire chromosome only invokes an alternate notation for the identical chromosome, and does not constitute a rearrangement operation.) **Translocation**, where two chromosomes (one or both inverted), exchange prefixes of any length. A **fusion** is a translocation where one of the prefixes is the entire chromosome and the other prefix is null. A **fission** is a translocation where one of the starting chromosomes is the null string. Our analyses of translocations implicitly include fusions and fissions.

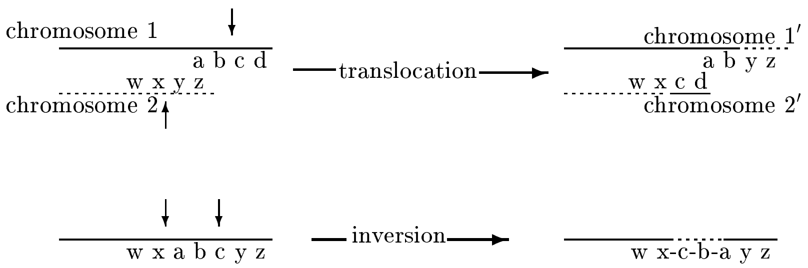

Fig. 1. Schematic view of genome rearrangement processes. Letters represent positions of genes. Vertical arrows at left indicate boundaries of affected substrings. Translocation exchanges prefixes of two chromosomes. Inversion reverses the order and sign of genes in a substring (dotted segment).

2 Resolution of tetraploidy; ancestral synteny unknown.

One form of hybridization of two karyotypically distinct species sees the fusion of two genomes followed by a series of chromosomal rearrangement events until the hybrid genome is finally stabilized as a diploid (e.g. [1]). The two homologous versions of each gene, one from each parent species, may diverge functionally to create a gene family. From the moment of hybridization till the present, the two parent species may also undergo chromosomal rearrangement. Thus we have direct access to neither the ancestral hybrid genome nor the two contributing strains. In this section we provide a method for reconstructing the ancestral hybrid, given the order of the genes on its chromosomes as well as data (obtained, for example, from sequence analysis) on which of these genes originated from each of the parent species.

2.1 Formalization.

Consider two genomes A and B having disjoints sets of genes, $\mathcal{E}(A)$ and $\mathcal{E}(B)$, respectively. Let G be a third genome with N chromosomes and gene set $\mathcal{E} =$

$\mathcal{E}(A) \cup \mathcal{E}(B)$. Given only $\mathcal{E}(A)$, $\mathcal{E}(B)$, and G, including how the genes are distributed and ordered on the N chromosomes of G, the problem is to find $d(G)$, the minimal number of inversions and translocations necessary to transform G into an **ancestral hybrid genome** H (with any number of chromosomes) satisfying the following condition: each chromosome of H contains genes from A only, or from B only. See Figure 2.

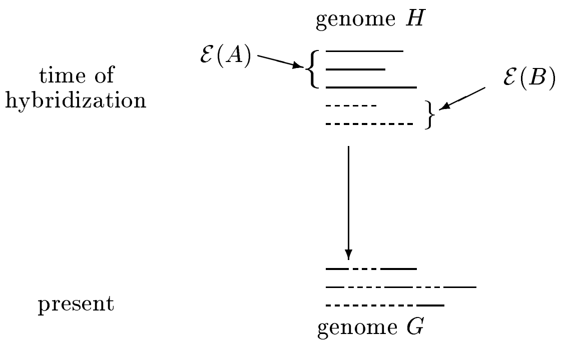

Fig. 2. Evolution of a hybrid genome resulting from genome fusion when gene origins, but not ancestral genome organization, is known. Genome H is to be reconstructed from knowledge of genome G, and ancestral gene sets $\mathcal{E}(A)$ and $\mathcal{E}(B)$ only.

2.2 Algorithm.

The following procedure solves this problem exactly in time linear in the number of genes. The output attains the lower bound of the type found by Watterson *et al.*, except for certain special cases.

- In each chromosome G_i of G, amalgamate each substring of consecutive A-origin genes to form an **A-segment**. Similarly form the **B-segments**. A-segments and B-segments alternate along the length of the chromosome, separated by **breakpoints**.
- Transform each chromosome with an odd number $b_i > 1$ of breakpoints to a chromosome consisting of a single A-segment and a single B-segment by means of $\frac{b_i - 1}{2}$ inversions as follows.
  - While there remain at least 3 breakpoints, invert the fragment between the first and third breakpoints. Two A-segments are thus made adjacent and two B-segments are made adjacent.
  - Erase the breakpoints between the two adjacent A-segments and between the two adjacent B segments, thus reducing the number of breakpoints by two.

- Transform each chromosome with an even number $b_i > 2$ of breakpoints to a chromosome consisting of either two A-segments and a single B-segment, or two B's and one A, by means of $\frac{b_i-2}{2}$ inversions as follows.
  - While there remain at least 4 breakpoints, invert the fragment between the first and third breakpoints. Two A-segments are thus made adjacent and two B-segments are made adjacent.
  - Erase the breakpoints between the two adjacent A-segments and between the two adjacent B segments, thus reducing the number of breakpoints by two.
- Form as many pairs of ABA and BAB chromosomes as possible. Two translocations performed on each pair suffice to produce a homogeneous A chromosome and a homogeneous B chromosome, allowing the erasure of all four breakpoints.
- Suppose some 2-breakpoint chromosomes remain and they are all ABA. They may be amalgamated two by two, each time with a translocation that produces a homogeneous A chromosome and an ABA chromosome, and allows the erasure of two breakpoints, until only one ABA remains.
- Suppose instead of the previous step, the only 2-breakpoint chromosomes remaining are BAB. They may be amalgamated two by two, each time with a translocation that produces a homogeneous B chromosome and an BAB chromosome, and allows the erasure of two breakpoints, until only one BAB remains.
- If there are any 1-breakpoint chromosomes, form as many pairs of them as possible.
  - If there are no 2-breakpoint chromosomes, transform each of the pairs of one-breakpoint chromosomes into one homogeneous A chromosome and one homogeneous B by means of a single translocation, and erase the two breakpoints.
  - If there is a 2-breakpoint chromosome, transform all but one of the pairs of chromosomes into one homogeneous A chromosome and one homogeneous B by means of a single translocation, and erase the two breakpoints. Then two translocations suffice to transform the remaining pair and the 2-breakpoint chromosome into three homogeneous chromosomes, and to erase all four remaining breakpoints.
- If 1- or 2-breakpoint chromosomes remain there are several cases:
  - If all that remains is a single 1-breakpoint chromosome, one translocation (fission) is required to produce two homogeneous chromosomes and remove the breakpoint.
  - If all that remains is a single 1-breakpoint chromosome and a single 2-breakpoint one, two translocations (one a fission) are required to produce two homogeneous chromosomes and to remove all three breakpoints.
  - If all that remains is a single 2-breakpoint chromosome, two translocations (fissions) are required to produce two homogeneous chromosomes and to remove both breakpoints.

The output from this algorithm consists of homogeneous A chromosomes and homogeneous B chromosomes only, and the number of steps is $\lceil \sum_i b_i \rceil /2 + \Psi$, where $\Psi = 0$ except if the last step of the algorithm must be activated, i.e., when there are no chromosomes G_i of forms $A \cdots B$ or $B \cdots A$, and an unequal number of chromosomes of forms $A \cdots A$ and $B \cdots B$. Here, $\Psi = 1$.

Note that there are generally many equally good solutions to this problem. In the next section, we reformulate the problem in order to pin down the structure of the ancestral genome somewhat. This will require additional data on the parent genomes and some assumptions about the amount of evolution in the hybrid compared to the purebred descendants of the parents.

3 Resolution of tetraploidy; ancestral synteny and gene order inferred.

A second version of the hybridization problem uses the modern configurations A, B and G of the two parent genomes and the hybrid genome, respectively, to infer the three ancestral genomes A', B' and G' at the moment of hybridization, as on the left of Figure 3. Note that G' consists of the chromosomes in A' plus the chromosomes in B'.

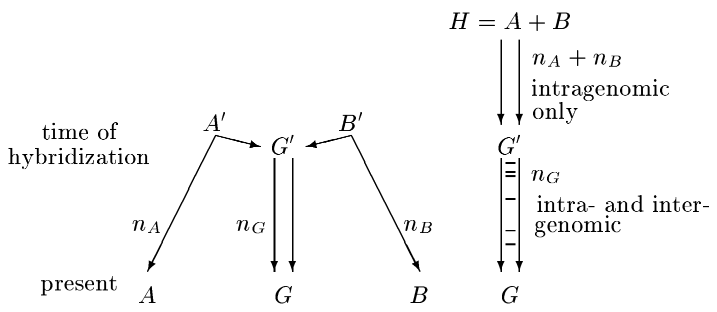

Fig. 3. Localization of ancestral hybrid immediately before intergenomic translocations

As a first step, we infer the total number n of evolutionary steps required to produce G from a construct H consisting of the chromosomes of A and the chromosomes of B, as on the right of Figure 3. We assume that G' is one of the intermediate steps in this evolution so that $n = n_A + n_B + n_G$, where n_X is the number of steps from genome X' to genome X, for $X \in \{A, B, G\}$.

Under the assumption that one of the first translocations to occur in the stabilization of the hybrid will be an **intergenomic** one, involving chromosomes from both A' and B', we could locate G' at the last step on the path from H to G before the first intergenomic translocation, as on the right of the figure. Unfortunately, the optimal path is not unique, and there will generally be one optimum whose *first* step is intergenomic, so that $n_A + n_B = 0$ and $n = n_G$.

This may indeed be biologically meaningful in some contexts where hybrids evolve more rapidly than their parents. In other cases we may prefer to look for the path where n_G is as small as possible, to allow for a maximum of evolution in the parent species.

It is this latter problem we investigate in this section. First we sketch the method of Hannenhalli and Pevzner [2, 3], hereafter "H-P", for finding the minimum number of translocations and inversions necessary to transform one genome into another, and show how a heuristic for finding a minimal n_G solution to the hybridization problem may be grafted onto their algorithm. We then show how to calculate, relatively quickly, an upper bound for this heuristic based on one step of the algorithm. Finally, we construct a lower bound based on a breakpoint argument.

3.1 The H-P algorithm and a heuristic for $n_\mathbf{G}$.

We will only sketch the H-P procedure, which is rather complex, and give additional details for those aspects which are modified in our heuristic. The first step in the comparison of two multi-chromosomal genomes through translocations and inversions is to reduce it to a problem of comparing two single chromosome genomes through inversion only. These latter genomes are constructed essentially by concatenating the individual chromosomes in the original genomes end-to-end in an arbitrary order. (Additional dummy genes, called **caps** must be appropriately inserted at the ends of the original chromosomes of both genomes). Translocation in an original genome becomes inversion in the new one. In the string representing a chromosome each gene $+x$ is replaced by the pair $x^t x^h$, and $-x$ by $x^h x^t$.

To find the minimum inversions $d(H,G)$ necessary to transform one single-chromosome genome H to another, G, H-P constructs a **cycle graph**, a bi-colored graph $\mathcal{G}(V,E)$ with vertex set V containing x^t and x^h for all genes in $\mathcal{E}$, where black edges connect neighboring vertices in H, and gray edges connect neighboring vertices in G. Each vertex is thus adjacent to exactly one black edge and one gray edge. $\mathcal{G}$ therefore has a unique decomposition into disjoint alternating cycles. We set $b(\mathcal{G}) = |\mathcal{E}| - 1$, the number of black edges of $\mathcal{G}$, and $c(\mathcal{G})$ to be the number of cycles of $\mathcal{G}$. Note that $c(\mathcal{G})$ is maximal when $G = H$. The **size of cycle** C is the number of black (or gray) edges in C. The inversion distance between H and G is then:

$$d(H,G) = b(\mathcal{G}) - c(\mathcal{G}) + h(\mathcal{G}) + f(\mathcal{G}) \qquad (1)$$

where $h(\mathcal{G})$ is the number of **hurdles** in $\mathcal{G}$, and $f(\mathcal{G}) = 1$ if $\mathcal{G}$ is a **fortress** and $f = 0$ if not. (These concepts will be discussed below.)

A key concept in the algorithm is the **oriented component**. A gray edge in a cycle is oriented if the inversion disrupting the two adjacent black edges, i.e.,

a adjacent to b in H, b adjacent to c in G, c adjacent to d in H
becomes
a adjacent to c in H, c adjacent to b in G, b adjacent to d in H,

replaces the cycle by two cycles. An oriented cycle is one containing at least one oriented gray edge. Two cycles whose containing gray edges that "cross", e.g., gene i adjacent to gene j in Cycle 1, gene k adjacent to gene t in Cycle 2 in G, but ordered i, k, j, t in H, are connected. A component of $\mathcal{G}(V, E)$ is a subset of the cycles, built recursively from one cycle, at each step adding all the remaining cycles connected to any of those already in the construction.

An oriented component has at least one oriented cycle. Hurdles are a particular class of unoriented components. The entire graph $\mathcal{G}(V, E)$ is a fortress if a certain configuration of hurdles obtains.

The H-P algorithm proceeds by decreasing $h - c$, the number of hurdles minus the number of cycles at each step. It handles each oriented component independently. If component C has γ_C black edges, and κ_C cycles, the algorithm proceeds to find a series of $\gamma_C - \kappa_C$ inversions that reduces the component to a set of γ_C 1-cycles.

Hurdles are treated somewhat differently. There is no inversion which will immediately increase the number of cycles in such a component. Instead, certain hurdles undergo an inversion which changes them into oriented components, decreasing the number of hurdles by one and leaving the number of cycles unchanged – hurdle "cutting". Other pairs of hurdles are merged by means of an inversion that *decreases* the cycle count by one, but also decreases the number of hurdles by two.

In each case, after the first inversion in a hurdle or pair of hurdles, the resulting configuration is an oriented component which may be reduced as above.

Unoriented components which are not hurdles will eventually become oriented through inversions operating on other components, and may then be reduced accordingly.

Thus the execution of the H-P algorithm involves repeatedly choosing an oriented cycle and performing an inversion around an oriented gray edge, thus increasing the number of cycles by one, except for the first inversion whenever hurdles must be cut or merged. The strategy for our heuristic focuses on the successive choices of cycles and edges within cycles. The idea is to stop the reduction of an oriented component when there is no choice of cycle and edge within the cycle which corresponds to an intragenomic translocation or inversion (i.e. involving genes from A only or genes from B only). Similarly, if either the conversion of a hurdle to an oriented component, or the pairing of two hurdles, corresponds to an intergenomic transfer, it is postponed.

This procedure is validated by the fact that each oriented component may be reduced independent of whatever inversions apply outside the component. Eventually, when no more intragenomic translocations are possible, we have reached a locally maximum value of $n_A + n_B$ (local minimum for n_G), the postponed reductions are re-started and the algorithm proceeds to an optimum solution.

3.2 An upper bound for the heuristic.

Suppose the decomposition of $\mathcal{G}(V, E)$ contains monogenomic oriented components $C_1, \cdots, C_r$ (each involving genes from a single genome only, A or B).

The decomposition may also contain other components. If component C_i has γ_{C_i} black edges and κ_{C_i} cycles, the r components will be reduced by $y = \sum_{i=1}^{r} \gamma_{C_i} - \kappa_{C_i}$ inversions. Then

$$d(H,G) - y \geq n_G,$$

where n_G is the value found by the heuristic.

This bound can be improved in three stages:

- By including, in the calculation of y, at least one monogenomic oriented cycle (if one exists) contained in each bi-genomic oriented component.
- By including, for each bi-genomic oriented component not satisfying the previous criterion, an intragenomic inversion (if one exists) around an oriented gray edge in a bi-genomic oriented cycle.
- By repeating the above steps on certain hurdles whose treatment does not depend on the previous analysis of other hurdles.

3.3 A lower bound for n_G.

Label the genes in G according to whether they come from A or B as in Section 2, and form segments of contiguous A's and B's. Suppose there are b breakpoints in all. Then at least $\lceil \frac{b}{2} \rceil$ translocations and inversions are required to remove these breakpoints, and these are necessarily intergenomic. I.e., $\lceil \frac{b}{2} \rceil \leq n_G$.

4 Hybridization through interspecific fertility.

Hybrids may be formed by the fertilization event of two distinct though related species, an accident in nature but often feasible in the laboratory, e.g. [4, 7]. The parent species A and B may differ from each other by numerous genome rearrangements. The hybrid G' is able to survive and propagate despite the difference between the two haploid components of its diploid genome. Genome rearrangement of the hybrid rapidly ensues, however, first until a normal symmetric diploid configuration G^* is attained, and then while further stabilization of the new genome occurs. This scenario is illustrated in Figure 4. The rapid evolution of the hybrid means that we can often assume the relative stability of the parent genomes A and B if the evolutionary scale is not too lengthy.

Suppose that the rearrangements of the hybrid between G' and G^* are intragenomic. I.e., the two hybrid genomes "conspire" to reorganize internally to a common form, before fixing any intergenomic translocations. Then the inference problem which arises is to find G^*, and the amount of rearrangement which occurred between G' and G^*, and between G^* and the modern genome G.

This is essentially the "median problem" for genomes [6]: Given three genomes A, B and G, find the "median" G^* which minimizes the sum of a genomic distance between G^* and A, G^* and B, and G^* and G. In general, this is a difficult problem, but in one case, namely when the distance is just the sum of the breakpoints between G^* and each of the other three genomes, an algorithm is available

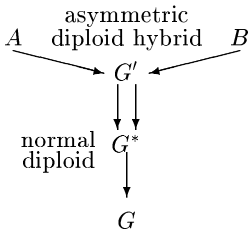

Fig. 4. Rearrangement before and after development of symmetric diploid.

[5], based on a reduction of the median problem to the Traveling Salesman Problem, which functions well for genomes containing a fairly large number of genes.

In the case where the rearrangements between G' and G^* are intergenomic from the start, it is difficult to propose a general model; unrestricted rearrangements in this context allow, for example, two versions of the same chromosomal segment in one haploid component, and zero in the other, meaning that the models of meiosis and mitosis underlying genome rearrangement theory no longer apply.

In the context of hybridization by interspecific fertilization, an additional type of data may be available. Genome typing informs us which chromosomal segments originate in which parental species (cf [7]). This pattern derives from the normal recombination events in the production of gametes. It may or may not be the case that the genomic position of a segment is correlated with that of its homologue in the parental species from which it derives. This illustrates the difference between this mechanism of hybridization and that of Sections 2 and 3, where genome fusion permits the retention of both of a pair of homologous genes, one from each parent.

Discussion

At least four aspects of the study of hybridization and rearrangement play a role in determining the nature of inference problem involved:

– The biological mechanism – genome fusion, interspecific fertilization, or other.

– The kinds of data available – present-day genomes, "ancestral" (i.e. stable or slowly-evolving) genomes, identification of genes in parental species (fusion model), or of segments (fertilization model).

– Assumptions about relative rates of evolution and about types of rearrangement event permitted.

– The entities to be inferred – events, syntenies, gene orders, beginning of intergenomic translocations.

The more detailed the kinds of data, the more detailed the kind of reconstruction that is possible, and the less ambiguity (non-uniqueness) in the results.

For example, the analysis in Section 2 generally reconstructs a large number of optimal solutions, while the one in Section 3 will be less ambiguous.

Each type of problem may require different tools from the inventory of methods developed in recent years for the study of genome rearrangement.

The most obvious domain of application of these methods is in the plant kingdom. The genomes of the cereals are particularly well-mapped and some of these show evidence of hybridization of the genome fusion type. The work of Rieseberg [4, 7] illustrates the possibilities of the analysis in Section 4. As more genomic data becomes available, our methods should be more widely applicable.

Acknowledgments

Research supported by grants to DS from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. DS is a Fellow, and NEM a Scholar, in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

References

1. Gaut, B.S., Doebley,J.F.: DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Science (U.S.A.) **94** (1997) 6809–6814.
2. Hannenhalli, S., Pevzner, P.A.: Transforming cabbage into turnip. (polynomial algorithm for sorting signed permutations by reversals). In: Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing (1995) 178–189.
3. Hannenhalli, S., Pevzner, P.A.: Transforming men into mice (polynomial algorithm for genomic distance problem). In: Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science (1995) 581–592.
4. Rieseberg, L.H, Van Fossen, C., Desrochers, S.M.: Hybrid speciation accompanied by genomic reorganization in wild sunflowers. Nature **375** (1995) 313–316.
5. Sankoff, D., Blanchette, M.: The median problem for breakpoints in comparative genomics. In: Computing and Combinatorics, Proceeedings of COCOON '97. (T. Jiang and D.T. Lee, eds) Lecture Notes in Computer Science **1276** Springer Verlag (1997) 251–263.
6. Sankoff, D., Sundaram, G., Kececioglu, J.: Steiner points in the space of genome rearrangements. International Journal of the Foundations of Computer Science **7** (1996) 1–9.
7. Ungerer, M.C., Baird, S.J.E., Pan, J., Rieseberg, L.H.: Rapid hybrid speciation in wild sunflowers. Proceedings of the National Academy of Science (U.S.A.) **95** (1998) 11757–11762.
8. Watterson, G.A., Ewens, W.J., Hall, T.E., Morgan, A.: The chromosome inversion problem. Journal of Theoretical Biology **99** (1982) 1–7.