Durrett and Interian Supplementary Material S1

Description of the Algorithm

Our algorithm (MEDBYLS) starts at an arbitrary genome and iteratively performs a sequence of rearrangements to examine possible ancestral genomes. After a predefined number of steps the algorithm outputs the best midpoint seen so far. At each step t, the algorithm has a current midpoint M_t and a set of *elementary rearrangements* R_t to get from M_t to M_{t+1}. The algorithm chooses a rearrangement uniformly at random from the set R_t of elementary rearrangements. If applying the rearrangement gives a midpoint with a better total distance $\sum_{i=1}^{3} d(M_{t+1}, G_i)$ than M_t the move is accepted, otherwise, the move is accepted with probability p. We typically use $p = 0.2$ and initialize our algorithm either with one of the three original genomes, or with a randomly chosen genome.

A key element in the success of the algorithm comes from the choice of the set R_t of possible moves or rearrangements at time t. By results of results of Hannenhalli and Pevzner (1995b) given two genomes G and G' we can identify a sequence of rearrangements to get from G to G' in the minimum number of moves. We define $R(G, G')$ to be a special subset of these moves (see section 3 for the exact definition) and define $R_t = \cup_{i=1}^{3} R(G_t, G^i)$. Thus at each step, we are taking a step in an optimal path toward one (or more) of the original genomes.

1 Notation and Basic Definitions

Genomes with one chromosome are represented as *signed permutations* where each integer corresponds to a gene or marker and the sign gives its orientation. A multichromosomal genome can be written as signed permutation divided into pieces called chromosomes. More precisely, given a set of markers $N = \{1, 2, \ldots n\}$, a *chromosome* τ is an ordering of a subset of the markers in which each marker has a sign, i.e., $\tau = (\tau_1 \ldots \tau_m)$ where $|\tau_i| \in N$, $m \leq n$ and we identify $(\tau_1 \ldots \tau_m)$ and $(-\tau_m \cdots - \tau_1)$. A *genome* G on a set of markers N is a collection of chromosomes in which each marker appears exactly once.

We consider four elementary kinds of rearrangements in a genome: *reversals, translocations, fusions and fissions.* A *reversal* $\rho = \rho_{i,j}$ of the interval (i, j), $1 \leq i \leq j \leq m$, applied to a chromosome $\pi = (\pi_1 \ldots \pi_m)$ takes π to

$$\pi\rho_{i,j} = (\pi_1 \ldots \pi_{i-1} \ -\pi_j \ -\pi_{j+1} \ldots \ -\pi_{i+1} \ -\pi_i \ \pi_{j+1} \ldots \pi_m)$$

A *translocation* $\rho = \rho_{i,j}$, $1 \leq i \leq m+1$, $1 \leq j \leq l+1$, applied to the two chromosomes $\pi = (\pi_1 \ldots \pi_m)$ and $\tau = (\tau_1 \ldots \tau_l)$ results in the two new chromosomes. There are two possibilities. The simplest is

$$\{\pi; \tau\}\rho_{i,j} = \{(\pi_1 \ldots \pi_{i-1} \ \tau_j \ldots \tau_l), \quad (\tau_1 \ldots \tau_{j-1} \ \pi_i \ldots \pi_m)\}$$

but we could also flip the first chromosome before translocating ending up with:

$$\{-\pi; \tau\}\rho_{i,j} = \{(-\pi_m \cdots - \pi_i \ \tau_j \ldots \tau_l), \quad (\tau_1 \ldots \tau_{j-1} \ -\pi_{i-1} \cdots - \pi_1)\}$$

A *fusion* is a particular kind of translocation $\rho = \rho_{m+1,1}$ that concatenates two chromosomes π and τ resulting in a chromosome $(\pi_1 \dots \pi_m\ \tau_1, \dots \tau_l)$ and an empty chromosome (we could also flip one of the chromosomes before fusing). A *fission* is the translocation $\rho = \rho_{i,1}$ that takes π and the empty chromosome resulting in two chromosomes $(\pi_1 \dots \pi_i)$ and $(\pi_{i+1} \dots \pi_m)$.

2 Breakpoint graph and genomic distance

In the study of the genomic distance for unichromosomal genomes Kececioglu and Sankoff (1995) and Bourque and Pevzner (1996) introduced the *breakpoint graph* for signed permutations. Caprara (1999b) generalized this notion to study the median problem for unichromosomal genomes. We will now further extend the notion of the breakpoint graph to the case of multichromosomal genomes and to more than two species.

The first step is to double the markers. For a signed marker $+x$, $u(+x) = 2x-1, 2x$ and $u(-x) = 2x, 2x-1$. Given a signed chromosome $(\tau_1, \dots, \tau_m)$ define an unsigned chromosome

$$u(\tau) = (u(\tau_1), \dots, u(\tau_m)) = (x_1, x_2, \dots, x_{2m-1}x_{2m})$$

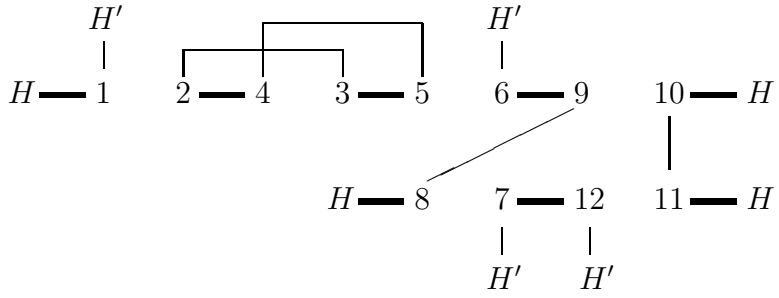
In this case the adjacency graph for the chromosome will have edges

$$\{(x_{2i}, x_{2i+1}) : i \in 1, \dots, n\} \cup \{(H, x_1), (H, x_{2m})\}.$$

For a genome we apply this procedure to each chromosome and take the union to form the adjacency graph $\Gamma(G)$. The H's in this graph denote the adjacencies to the chromosome "ends". The H's are all different points in the graph, but we denote them by the same symbol to simplify the notation.

Given two genomes G and G', with $k \geq \ell$ chromosomes, the breakpoint graph $\Gamma(G, G')$ is defined by combining the adjacency graphs $\Gamma(G)$ and $\Gamma(G')$ using different labels H and H' for chromosome ends corresponding to different genomes, and different colors for edges, say *black* for G and *gray* for G'. If $k > l$ we add $k - l$ empty chromosomes to G'.

The next picture gives the example $\Gamma(G, G')$, for $G = \{1 - 2\ 3\ 5;\ -4 - 6\}$ and $G' = \{1\ 2\ 3;\ 4\ 5\ 6\}$. Doubling the markers in the first genome gives edges $H - 1$, $2 - 4$, $3 - 5$, $6 - 9$, $10 - H$, $H - 8$, $7 - 12$, $11 - H$ which are drawn as thick lines (black edges). We then add the adjacencies in the other genome as thin lines (gray edges).



Note that, except for the special nodes H and H', which always have degree one, all the other nodes in $\Gamma(G, G')$ have degree 2, and are incident with one gray and one black edge.

We have two types of connected components in the breakpoint graph $\Gamma(G, G')$, one without special vertices $\mathcal{C} = \{x_1 x_2 \dots x_{2r}\}$ that we call *cycles*, and a second type that begins and ends with special vertices: $\mathcal{C} = \{H' x_1 x_2 \dots x_{2r+1} H\}$, $\mathcal{C} = \{H x_1 x_2 \dots x_{2r} H\}$, $\mathcal{C} = \{H' x_1 x_2 \dots x_{2r} H'\}$ that we call *paths*. We can write $\Gamma(G, G')$ in a unique way as a union of paths that start and end at the special nodes (with no special node in the middle of each path), and cycles of non-special nodes. Let $c(G, G')$ be the number of paths and cycles, including empty chromosomes. Let $\#(H, H')$ be the number of cycles that start with H and end with H'.

In the example drawn above the breakpoint graph has five components $\{H\ 1\ H'\}$, $\{2\ 4\ 5\ 3\ 2\}$, $\{H'\ 6\ 9\ 8\ H\}$, $\{H\ 10\ 11\ H\}$ and $\{H'\ 7\ 12\ H'\}$ so $c(G, G') = 5$, and the number of same genome cycles $\#(H, H) = \#(H', H') = 1$.

The *graph distance* between two genomes $G = \{\tau^1, \dots, \tau^k\}$, $G' = \{\pi^1, \dots, \pi^l\}$ is defined as

$$d(G, G') = n + k - c(G, G') + \#(H', H').$$

The graph distance for the example in Figure 4 is $6 + 2 - 5 + 1 = 4$. Since any move can only decrease $d(G, G')$ by at most 1, the graph distance is a lower bound

on the genomic distance. It is easy to check in the example we are considering that we can transform G into G' in four rearrangements.

3 Elementary rearrangements

The final step in the description of the algorithm is to define the elementary rearrangments $R(G, G')$. We say that a component of the breakpoint graph is *elementary* if it has the form $\{H\ x\ H'\}$ or $\{x\ y\}$. The first kind corresponds to a common end, the second one corresponds to x and y being adjacent in both genomes. We say a rearrangement ρ acting on a cycle (path) of the breakpoint graph $\Gamma(G, G')$ is *elementary* if $\Gamma(G\rho, G')$ is obtained from $\Gamma(G, G')$ by one of the following operations:

- splitting one of the cycles, (with or without special vertices), $\mathcal{C}$ into a 2-cycle and a smaller cycle. This applies to all paths and cycles
- splitting one of the cycles (paths) $\mathcal{C}$ with special vertices into a 3-vertex path and a smaller path.