Project Description: Search Methods for the Genomic Distance Problem

Background

Different genomes may have many genes in common, but the genes may be arranged in a different order or be located on different chromosomes. Such rearrangements are common in molecular evolution. In unichromosomal genomes, the most common rearrangement events are *reversals*, in which a contiguous interval of genes is put into the reverse order. The minimal number of reversals needed to transform one unichromosomal genome into another is called the *reversal distance* and can be used to measure the evolutionary distance between different unichromosomal organisms. The problem of finding the reversal distance between two different unichromosomal genomes is called the *reversal distance problem.*

When considering multichromosomal genomes the most common rearrangement events are not only reversals but also *translocations*, *fissions*, and *fusions*. The minimal number of such rearrangement events between two multichromosomal genomes is called the *genomic distance* thus the problem of finding them the *genomic distance problem.*

Genome rearrangements

A unichromosomal genome is represented as a sequence of genes, where each gene $i \in \{1,...,n\}$. Genes can be arranged in either of the two possible directions which is represented by the sign of the gene, i.e. gene $-i$ is in opposite direction to i. A chromosome (a genome in the unichromosomal case) is represented as a signed permutation of n genes. In the multichromosomal case the n genes are distributed on m chromosomes. For example, a genome with $n = 12$ genes spread over $m = 3$ chromosomes is

{7 -2 8 3}
{5 9 -6 -1 12}
{11 4 10}

The actual order of chromosomes does not matter, i.e. we could call {7 -2 8 3} chromosome 1, 2, or 3. Moreover, the direction of the chromosomes does not matter, e.g. {5 9 -6 -1 12} and {-12 1 6 -9 -5} represent the same chromosome.

Reversals

Reversals are rearrangements that occur within chromosomes. A reversal of a subsequence of genes A on a chromosome reverses and negates that subsequence, denoted $-A$, within the chromosome. For example consider the reversal of genes -6, -1, and 12 on the chromosome {5 9 -6 -1 12}

{7 -2 8 3}
{5 9 -6 -1 12}
{11 4 10}

$\Longrightarrow$
{7 -2 8 3}
{5 9 -12 1 6}
{11 4 10}

i.e. –{-6 -1 12} = {-12 1 6}

Translocations

Translocations are rearrangements where two chromosomes AB and CD are rearranged into AD and CB. As an example consider the following translocation

{7 -2 8 3}
{5 9 -6 -1 12}
{11 4 10}
$\Longrightarrow$
{7 -2 10}
{5 9 -6 -1 12}
{11 4 8 3}

Alternative translocations events could be to allow simultaneous reversals. For example AB and CD could result in A-C and –BD.

Fusions

Two chromosomes may be fused into one. Reversals are allowed in fusions, so the fusion of the chromosome sequences *A* and *B* could yield 4 different results:

AB (= -*B*-*A*)
BA (= -*A*-*B*)
-*AB* (= -*BA*)
B-*A* (= *A*-*B*)

For example

{7 -2 8 3}
{5 9 -6 -1 12}
{11 4 10}
$\Longrightarrow$
{7 -2 8 3 -10 -4 -11}
{5 9 -6 -1 12}

Fissions

Fissions are rearrangements events where one chromosome is split into two chromosomes between any pair of genes. For example

{7 -2 8 3}
{5 9 -6 -1 12}
{11 4 10}
$\Longrightarrow$
{7 -2 8 3}
{5 9}

{-6 -1 12}
{11 4 10}

Assignment

Your task is to:

1. Find a good representation for multichromosomal genomes.

2. Develop methods for genome rearrangements, i.e. reversal, translocation, fusion, and fission.

3. Formulate the genomic distance problem as a search problem.

4. Implement a search algorithm for solving the problem, i.e. finding the minimal number of rearrangements. The actual path could be interesting, at least for debugging, so primitive procedures for printing rearrangements events is needed.

5. Think of a heuristic to be used by the search algorithm. Incorporate the heuristic in the algorithm.

6. Test your algorithm on some examples.

As an example problem consider the two genomes

G1	**G2**
{1 2 3 4 5 6}	{1 -5 2 6}
{7 8}	{21 -22 -20 8}
{9 10}	{-4 14 11 -15 3 9}
{11 12 13 14 15 16 17 18}	{7 16 -18 17}
{19 20 21 22}	{-19 24 -26 27 25}
{23 24 25 26 27}	{-12 23 13 10}

The genomic distance between G1 and G2, i.e. the number of rearrangements needed to transform G1 into G2 is 20 (see Rick Durrett, Genome Rearrangement: Open Progress and Open Problems, link on course home page).

For hints and more details see for example Rick Durretts survey on genome rearrangements and Anne Bergeron, A Very Elementary Presentation of the Hannenhalli-Pevzner Theory, linked from the course home page. It is of course ok to fetch inspiration from other literature sources for solving the problem.

```
(define programming-language ‘scheme)
```