Algorithms in Comparative Genomics

Lecture: Marília D. V. Braga Thursdays, 10:15-11:45

Tutorial:

Leonard Bohnenkämper Thursdays (or Fridays?), 8:30-10:00

Topics:

Intro Arction 1. Genomes as gene orders or list of adjacencies 2. Large-scale rearrangements 3. Types of genomes 4. Breakpoint distance, breakpoint double distance Grome (cafres/ 5. Common intervals and gene clusters 6. Relational diagram of two genomes Double-cut-and-join (DCJ) operation 7.1 DCJ distance Comme reargements 7.2 DC I-indel distance 7.3 Capping (circularizing genomes) 7.4 Duplications and ILP 7.5 Family-free setting 8. Inversion distance

9. Single-cut-or-join (SCJ) distance and median

10. Small parsimony

Topics of today - Introduction:

- $1. \ \mbox{Genomes}$ as gene orders or list of adjacencies
- 2. Large-scale rearrangements
- 3. Types of genomes
- 4. Breakpoint distance, breakpoint double distance

Genomes





Genomes as gene orders or list of adjacencies - linear chromosomes

Each gene is an oriented DNA fragment: it lies on one of the two complementary anti-parallel DNA strands



Genomes as gene orders or list of adjacencies - circular chromosomes













1. Singular genome: each family occurs exactly once

2. Duplicated genome: each family occurs exactly twice

- 3. Perfectly duplicated or doubled genome: each adjacency or telomere occurs exactly twice
- 4. Natural genome: no restriction on the number of occurrences of families

Comparison of genomes



Types of genome pairs







each family occurs the same number of times in each genome

Pair of natural genomes:

no restriction on the number of occurrences of families

Types of genome pairs

Pair of singular genomes: each family occurs at most once in each genome Pair of canonical genomes: singular and balanced



Pair of natural genomes:

no restriction on the number of occurrences of families



Breakpoint distance of canonical genomes



Breakpoint distance of canonical genomes



Breakpoint distance of canonical genomes



Breakpoint distance of balanced genomes

 ${\sf Common \ adjacency} \ \times \ {\sf breakpoint}$



Breakpoint distance of balanced genomes

Common adjacency \times breakpoint



Breakpoint distance of balanced genomes

Common adjacency \times breakpoint



Breakpoint double distance



Given a duplicated genome D, the breakpoint double distance is defined as:

$$d_{\rm BP}^2(D,S)=d_{\rm BP}(D,S\oplus S)$$





Reference

Multichromosomal median and halving problems under different genomic distances (Eric Tannier, Chunfang Zheng and David Sankoff) BMC Bioinformatics volume 10, Article number: 120 (2009)

Quiz

Given genomes $A = (1234) [1\bar{5}\bar{4}5\bar{3}\bar{2}]$, B = [12345] and $C = [\bar{2}\bar{1}] [\bar{4}\bar{3}5]$.

- 1 Which of the following statements are true?
- 2 What is the breakpoint distance of B and C?
- A Genome A is linear.A 1,5B Genome A is multichromosomal.B 2C Genome A is duplicated.C 2,5D Genome A is doubled.D 3
- 2 How many families occur in genome A?

4 What is the breakpoint double distance of A and B?

A	4	A	4
В	5	В	4,2
С	5,5	С	4,5
D	6	D	5

Breakpoint halving

Given a duplicated genome *D*, find a singular genome *S* that minimizes the breakpoint double distance:

 $d^2_{\scriptscriptstyle\mathrm{BP}}(D,S)=d_{\scriptscriptstyle\mathrm{BP}}(D,S\oplus S)$