Algorithms in Genome Research

Pedro Feijao

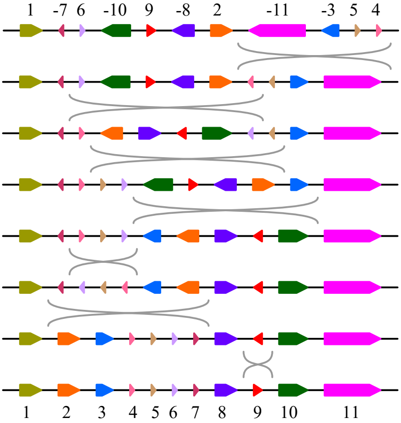
Summer 2015

pfeijao@cebitec.uni-bielefeld.de

Multiple Genome Rearrangement and the Breakpoint Model

Genome Rearrangement Scenarios

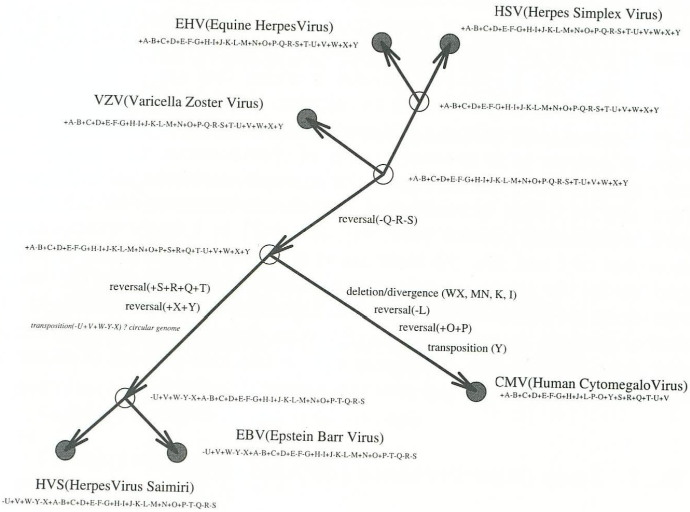
- Finding genome rearrangement scenarios between two genomes is usually easy.



Genome Rearrangement Scenarios

- What if we have more genomes? Can we find an evolutionary scenario?
- Ideally, we want a **rearrangement phylogeny**, explaining ancestral configurations and rearrangement scenarios.
- For instance, something like:

Evolution of Herpes Viruses


EHV(Equine HerpesVirus)
+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P-Q-R-S+T-U+V+W+X+Y
HSV(Herpes Simplex Virus)
+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P-Q-R-S+T-U+V+W+X+Y
VZV(Varicella Zoster Virus)
+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P-Q-R-S+T-U+V+W+X+Y
+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P-Q-R-S+T-U+V+W+X+Y
+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P-Q-R-S+T-U+V+W+X+Y
reversal(-Q-R-S)
+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P+S+R+Q+T-U+V+W+X+Y
reversal(+S+R+Q+T)
reversal(+X+Y)
transposition(-U+V+W-Y-X) ? circular genome
deletion/divergence (WX, MN, K, I)
reversal(-L)
reversal(+O+P)
transposition (Y)
CMV(Human CytomegaloVirus)
+A-B+C+D+E-F-G+H+J+L-P-O+Y+S+R+Q+T-U+V
-U+V+W-Y-X+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P-T-Q-R-S
EBV(Epstein Barr Virus)
-U+V+W-Y-X+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P-T-Q-R-S
HVS(HerpesVirus Saimiri)
-U+V+W-Y-X+A-B+C+D+E-F-G+H-I+J-K-L-M+N+O+P-T-Q-R-S

Pevzner, Computational Molecular Biology: An Algorithmic Approach (2000)

Multiple Genome Rearrangement

- The complexity of many combinatorial problems increases when the number of objects increase from 2 to 3.
- Genome Rearrangement is no exception: when comparing 3 (or more) genomes, most rearrangement models are NP-hard.

Multiple Genome Rearrangement

- We are looking for the *most parsimonious phylogenetic tree*. More formally:

Multiple Genome Rearrangement Problem – MGR

Given n genomes, find a tree T with the n genomes as *leaf nodes* and assign ancestral genomes to internal nodes of T such that the tree is optimal, i.e., the sum of rearrangement distances over all edges of the tree is minimal.

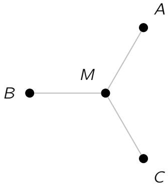
- This problem is also called the **Big Parsimony Problem**.
- In the **Small Parsimony Problem**, a tree T is given, and only the ancestral assignment is needed.
- The simplest form of the MGR is the **median problem**, when three input genomes are considered.

Genome Median Problem

Given three genomes A, B and C, and a genome distance measure d, find a genome M where the **median score**

$$s(M) = d(A, M) + d(B, M) + d(C, M)$$

is minimized.



This can be used as a subproblem to solve the Small Parsimony, iteratively finding the median in the internal nodes of the tree until convergence is achieved.

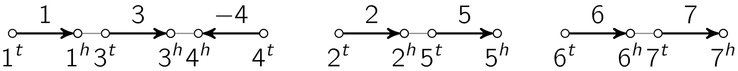
Genome Median Problem

Unfortunately, the median problem is NP-hard for most rearrangement distances, except for *breakpoint distances* in some cases.

- **Unichromosomal BP**: NP-hard
  - Linear Genomes: Pe'er and Shamir, 1998
  - Circular Genomes: Bryant, 1998
- **Reversal**: NP-hard (Caprara, 1997)
- **DCJ**: NP-hard (Caprara, 1997; Tannier et al. 2009)
- **Multichromosomal BP**: $O(n^3)$ (Tannier et al. 2009); $O(n\sqrt{n})$ (Kováč, 2013)
- **Single-Cut-or-Join**: $O(n)$ (Feijão and Meidanis, 2009)

Multichromosomal BP Distance

- Proposed by Tannier et al., in 2009.
- Similarly to the DCJ model, genomes are defined as sets of adjacencies and telomeres, given a gene set $\mathcal{A}$.
- For instance, given $\mathcal{A} = \{1, 2, 3, 4, 5, 6, 7\}$, we can define the genome $A = \{1^t, 1^h3^t, 3^h4^h, 4^t, 2^t, 2^h5^t, 5^h, 6^t, 6^h7^t, 7^h\}$



Multichromosomal BP Distance

Multichromosomal BP Distance – Tannier et al., 2009

Given genomes A and B, the multichromosomal BP distance is defined as

$$d_{\mathrm{BP}}(A, B) = N - A - \frac{T}{2}$$

where N is the number of genes, A is the number of common adjacencies and T the number of common telomeres in A and B.

Alternatively, using the **Adjacency Graph**:

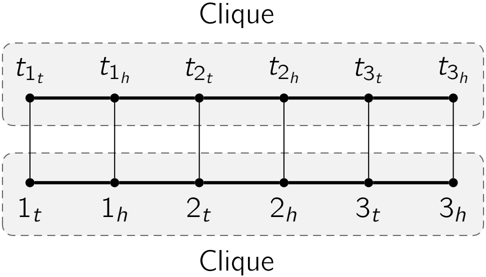
$$d_{\mathrm{BP}}(A, B) = N - C_2 - \frac{P_1}{2}$$

where N is the number of genes, C_2 is the number of cycles of lenght 2 and T the number of paths of lenght 1 in $AG(A, B)$.

Median Problem - BP Distance

- Given a gene set $\mathcal{A}$, consider a graph G whose vertex set has two vertices, x and t_x, for each extremity x of the genes in $\mathcal{A}$.
- There is an edge between x and t_x, for all extremities x, and also and edge between **all** pairs of x vertices and all pairs of t_x vertices.

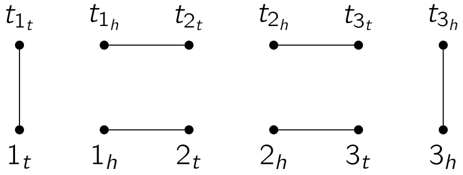
For instance, for $\mathcal{A} = \{1, 2, 3\}$ we have this graph:



Property: **Perfect Matching** in $G \iff$ **Genome** in $\mathcal{A}$.

Example

For gene set $\mathcal{A} = \{1, 2, 3\}$, and genome A = $\{1_t, 1_h 2_t, 2_h 3_t, 3_h\}$ we have the following matching:



- "Horizontal edges" → Adjacencies in the genome.
- "Vertical edges" → Telomeres in the genome.

Median Problem - BP Distance

Now consider the same graph G, in an weighted form: Given genomes A, B and C, assign weights to the edges of G in this form:

- **Adjacency weights**: for each adjacency edge (x, y), the weight is # of genomes that have adjacency xy ($w = 0, 1, 2$ or 3).
- **Telomere weights**: for each telomere edge (x, t_x), weight is # of genomes that have telomere x divided by 2 ($w = 0, 1/2, 1$ or $3/2$).
- Any other edge has weight 0.

Matching Weight and Median Score

Claim

Consider three genomes A, B and C, and the weighted graph G. For any genome M, the corresponding weighted matching in G has total weight

$$w = 3N - (d_{\mathrm{BP}}(A, M) + d_{\mathrm{BP}}(B, M) + d_{\mathrm{BP}}(C, M)) = 3N - s(M)$$

where $s(M)$ is the **median score** of M.

Proof?

Therefore, solving the **maximum weight perfect matching** problem in G (can be done in $O(n^3)$), we find a median with minimum score, solving the median problem.