Genome-scale evolution: multiple genome rearrangement, phylogeny based on whole genome sequence

Material of this lecture taken from the papers
M. Blanchette, T. Kunisawa, D. Sankoff, J Molecular Evolution 49, 193 (1999)
Gene Order Breakpoint Evidence in Animal Mitochondrial Phylogeny
G. Bourque, PA Pevzner, Genome Res 12, 36 (2002)
Genome-scale evolution: reconstructing gene orders in the ancestral species.

In comparative genomics, the quantitative comparison of gene order differences can be used for phylogenetic inference about a set of organisms.
Traditional Phylogeny

Traditional phylogenetic tree reconstruction is based on the analysis (e.g. level of conservation of amino acids) of individual genes/proteins. Genetic distance is defined as \# mismatches / \# matches. Sequence conservation depends on physico-chemical properties of amino acids (and genome context such as G+C content).

Last lecture (mouse:man) we saw that many more genomic elements are conserved between related species than only the genes.

Therefore, genome rearrangement studies that are based on genome-wide analysis of gene orders rather than individual genes, may provide a more general picture on evolution.

In future both approaches should probably be combined.
Reconstruction of phylogenetic trees from WG data

1. **Phylogeny reconstruction as optimization problem?**
   Attempt to reconstruct an evolutionary scenario with a minimum number of permitted evolutionary events (e.g. duplications, insertions, deletions, inversions, transpositions) on a tree → all known approaches are NP-hard. Also, no automated tool exists so far.

2. **Estimate leaf-to-leaf distances**
   (based on some metric) between all genomes. Then use a standard distance-based method such as *neighbour-joining* to construct the tree. Such approaches are quite fast but cannot recover the ancestral gene order.

2a. **Breakpoint phylogeny (Blanchette & Sankoff)**
   for special case in which the genomes all have the same set of genes, and each gene appears once. Use breakpoint distance as distance matrix.
Reversal distance problem

Although the reversal distance for a pair of genomes can be computed in polynomial time (Hannenhalli & Pevzner 1999 and others), its use in studies of multiple genome rearrangements was somewhat limited since it was not clear how to combine pairwise rearrangement scenarios into a multiple rearrangement scenario.

In particular, Capara (1999) demonstrated that even the simplest version of the Multiple Genome Rearrangement Problem, the Median Problem, is NP-hard.

Therefore, this line of research was abandoned for a while in favor of the breakpoint analysis approach (see Blanchette & Sankoff). The existing tools BPAnalysis or GRAPPA use the so-called breakpoint distance to derive rearrangement scenarios.
Breakpoint phylogeny

When each genome has the same set of genes and each gene appears exactly once, a genome can be described by a (circular or linear) ordering = permutation of these genes. Each gene has either positive \(g_i\) or negative \((-g_i)\) orientation.

Given 2 genomes \(G\) and \(G'\) on the same set of genes, a **breakpoint** in \(G\) is defined as an ordered pair of genes \((g_i,g_j)\) such that \(g_i\) and \(g_j\) appear consecutively in that order in \(G\), but neither \((g_i,g_j)\) \((-g_i,-g_j)\) appears consecutively in that order in \(G'\).

The **breakpoint distance** between two genomes is simply the number of breakpoints between that pair of genomes. The **breakpoint score** of a tree in which each node is labelled by a signed ordering of genes is then the sum of the breakpoint distances along the edges of the tree.
Phylogeny of metazoa: 3 competing models

Some aspects of metazoan phylogeny are still controversial (see left side).

→ Analyze mitochondrial gene order for most diverse members of each group.

Common among 3 models: ECH and CHO are grouped together. Also, ANN and MOL should be closely linked with ART related to these at a deeper level.

Blanchette, Sankoff, J Mol Evol (1999)
Species included in analysis

Table 1. Mitochondrial genomes compared in this investigation, with assumed monophyletic groupings

<table>
<thead>
<tr>
<th>Organism</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>HU Human</td>
<td>CHO</td>
</tr>
<tr>
<td>SS Asterina pectintifera (sea star)</td>
<td>ECH</td>
</tr>
<tr>
<td>SU Strongylocentrotus purpuratus (sea urchin)</td>
<td>ECH</td>
</tr>
<tr>
<td>DR Drosophila yakuba (insect)</td>
<td>ART</td>
</tr>
<tr>
<td>AF Artemia franciscana (crustacean)</td>
<td>ART</td>
</tr>
<tr>
<td>AC Albinaria coerulea (snail)</td>
<td>MOL</td>
</tr>
<tr>
<td>CN Cepaea nemoralis (snail)</td>
<td>MOL</td>
</tr>
<tr>
<td>KT Katharina tunicata (chiton)</td>
<td>ANN</td>
</tr>
<tr>
<td>LU Lumbricus terrestris (earthworm)</td>
<td>ANN</td>
</tr>
<tr>
<td>AS Ascarts suum</td>
<td>NEM</td>
</tr>
<tr>
<td>OV Onchocerca volvulus</td>
<td>NEM</td>
</tr>
</tbody>
</table>

Blanchette, Sankoff, J Mol Evol (1999)

Advantage of breakpoint analysis:
Number of breakpoints can be computed very easily, in linear time.
Distance matrices for 11 species

Number of breakpoints indicates that many of the gene orders seem to be random permutations of each other (random genomes with \( n \) genes would have \( n - 0.5 \) breakpoints with each other, on average).

<table>
<thead>
<tr>
<th></th>
<th>AS-35.1</th>
<th>OV-35.3</th>
<th>HU-30.0</th>
<th>SS-33.4</th>
<th>SU-33.3</th>
<th>DR-30.9</th>
<th>AF-31.3</th>
<th>AC-35.1</th>
<th>CN-33.0</th>
<th>KT-29.1</th>
<th>LU-31.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>21</td>
<td>35</td>
<td>33</td>
<td>33</td>
<td>36</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>OV</td>
<td>21.0</td>
<td>35</td>
<td>33</td>
<td>33</td>
<td>36</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>33</td>
</tr>
<tr>
<td>HU</td>
<td>34.3</td>
<td>35.3</td>
<td>33</td>
<td>33</td>
<td>35</td>
<td>33</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>35</td>
<td>32</td>
</tr>
<tr>
<td>SS</td>
<td>32.5</td>
<td>31.6</td>
<td>26.1</td>
<td>28</td>
<td>24</td>
<td>20</td>
<td>23</td>
<td>35</td>
<td>35</td>
<td>33</td>
<td>28</td>
</tr>
<tr>
<td>SU</td>
<td>34.3</td>
<td>34.2</td>
<td>24.2</td>
<td>3.0</td>
<td>5</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>35</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td>DR</td>
<td>34.2</td>
<td>33.4</td>
<td>19.4</td>
<td>31.2</td>
<td>32.2</td>
<td>6</td>
<td>35</td>
<td>35</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>AF</td>
<td>35.2</td>
<td>33.3</td>
<td>21.2</td>
<td>32.0</td>
<td>31.3</td>
<td>4.0</td>
<td>34</td>
<td>31</td>
<td>21</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td>AC</td>
<td>34.1</td>
<td>34.2</td>
<td>33.3</td>
<td>32.2</td>
<td>33.4</td>
<td>32.5</td>
<td>8</td>
<td>34</td>
<td>34</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>CN</td>
<td>31.3</td>
<td>31.3</td>
<td>33.2</td>
<td>30.5</td>
<td>29.5</td>
<td>30.6</td>
<td>31.2</td>
<td>5.1</td>
<td>31</td>
<td>31</td>
<td>29</td>
</tr>
<tr>
<td>KT</td>
<td>32.2</td>
<td>30.3</td>
<td>28.2</td>
<td>32.3</td>
<td>32.4</td>
<td>19.3</td>
<td>21.2</td>
<td>32.3</td>
<td>31.2</td>
<td>31.2</td>
<td>22</td>
</tr>
<tr>
<td>LU</td>
<td>32.5</td>
<td>32.4</td>
<td>30.4</td>
<td>32.1</td>
<td>32.2</td>
<td>26.3</td>
<td>26.5</td>
<td>33.2</td>
<td>29.4</td>
<td>22.2</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Average breakpoint distance to torgroup members (e.g., excluding MOL corembers CN and KT for AC, excluding ECH corember SU for SS, and no exclusions for HU) on the top row. Note that AS and OV each have only 36 genes and CN has 35, while the other genomes have 37.

Blanchette, Sankoff, J Mol Evol (1999)
Cleared data

same data, but highly mobile tRNA genes deleted. Good correlation between breakpoint distance and the other two distances.

Table 3. Distance matrices omitting tRNA genes (triangular matrices, top to bottom: breakpoints, minimal inversion, combined inversion/transposition)\(^a\)

<table>
<thead>
<tr>
<th></th>
<th>AS-13.3</th>
<th>OV-13.4</th>
<th>HU-9.1</th>
<th>SS-12.0</th>
<th>SU-11.8</th>
<th>DR-10.1</th>
<th>AF-10.1</th>
<th>AC-14.3</th>
<th>CN-13.6</th>
<th>KT-8.9</th>
<th>LU-11.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HU</td>
<td>13</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>14</td>
<td>14</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SU</td>
<td>13</td>
<td>13</td>
<td>6</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td>13</td>
<td>13</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AF</td>
<td>13</td>
<td>13</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CN</td>
<td>14</td>
<td>14</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KT</td>
<td>13</td>
<td>13</td>
<td>6</td>
<td>11</td>
<td>11</td>
<td>5</td>
<td>5</td>
<td>14</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LU</td>
<td>13</td>
<td>13</td>
<td>9</td>
<td>14</td>
<td>14</td>
<td>8</td>
<td>8</td>
<td>14</td>
<td>14</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AS</th>
<th>OV</th>
<th>HU</th>
<th>SS</th>
<th>SU</th>
<th>DR</th>
<th>AF</th>
<th>AC</th>
<th>CN</th>
<th>KT</th>
<th>LU</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>5</td>
<td>11</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>OV</td>
<td>4.2</td>
<td>13</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>HU</td>
<td>11.4</td>
<td>13.5</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>12</td>
<td>5</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>SS</td>
<td>12.3</td>
<td>12.2</td>
<td>4.1</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>14</td>
<td>12</td>
<td>9</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>SU</td>
<td>11.1</td>
<td>11.5</td>
<td>4.2</td>
<td>1.0</td>
<td>7</td>
<td>7</td>
<td>14</td>
<td>13</td>
<td>9</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>DR</td>
<td>11.3</td>
<td>11.3</td>
<td>3.0</td>
<td>7.1</td>
<td>7.1</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>AF</td>
<td>11.3</td>
<td>11.3</td>
<td>3.0</td>
<td>7.1</td>
<td>7.1</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>AC</td>
<td>12.4</td>
<td>12.1</td>
<td>12.2</td>
<td>14.2</td>
<td>14.2</td>
<td>12.2</td>
<td>12.2</td>
<td>12.2</td>
<td>3</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>CN</td>
<td>13.5</td>
<td>12.1</td>
<td>12.1</td>
<td>12.3</td>
<td>12.3</td>
<td>12.3</td>
<td>12.3</td>
<td>2.1</td>
<td>11</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>KT</td>
<td>12.2</td>
<td>12.2</td>
<td>5.1</td>
<td>9.1</td>
<td>9.2</td>
<td>3.1</td>
<td>3.1</td>
<td>11.2</td>
<td>11.2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>LU</td>
<td>11.3</td>
<td>10.5</td>
<td>7.2</td>
<td>11.2</td>
<td>11.3</td>
<td>7.2</td>
<td>7.2</td>
<td>12.2</td>
<td>10.2</td>
<td>5.1</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) Average breakpoint distance to nongroup members (e.g., excluding MOL comembers CN and KT for AC, excluding ECH comember SU for SS, and no exclusions for HU) on the top row. Note that AS and OV each have only 14 genes, while the other genomes have 15.

Blanchette, Sankoff, J Mol Evol (1999)
Tree inference

Compare 3 criteria for optimum tree topology in the light of theories of metazoan evolution:
(a) Neighbour-joining
(b) Fitch-Margoliash
(b) minimum breakpoint.

(a) and (b) operate on the genome data as reduced to the breakpoint distance matrix.
(c) is based on the gene orders themselves.
Tree from neighbor-joining analysis

Neighbour joining disrupts the deuterostomes by grouping ART with the human genome, and disrupts the molluscs.

The Fitch-Margoliash routine minimizes the sum of squared differences between distance matrix entries and total path length in the tree between two species, divided by the square of the matrix entry.

Worse grouping than in (a): the rapidly evolving lineages, NEM, snails, and ECH are grouped together, thus completely disrupting both the CHO+ECH grouping and the MOL grouping.

Blanchette, Sankoff, J Mol Evol (1999)
A minimum breakpoint tree is one in which
(a) a genome is reconstructed for each ancestral node,
(b) the number of breakpoints is calculated for each pair of nodes directly connected by a branch of the tree,
(c) the sum is taken over all branches, where the sum is minimal over all possible trees.

Problematic: all possible trees on the set of given data genomes need to be evaluated. For median problem analogy to travelling salesman problem.

Blanchette et al. didn’t want to question the original 3 models solely on basis of this data. All trees not consistent with either of the 3 models was disrupted, leaving 105 trees!

Blanchette, Sankoff, J Mol Evol (1999)
Scores for 105 trees evaluated

Two trees in (c) are optimal, but neither is biologically plausible.

These two „optimal phylogenies“ have 199 breakpoints.

The original 3 trees have suboptimal scores: CAL 203, TOL 205, LAKE 206.

The study of genomic rearrangements cannot provide unique solutions. There are often many distinct solutions, all optimal, and many ways of arriving at these results.

Blanchette, Sankoff, J Mol Evol (1999)
Unambiguously reconstructed segments

20 optimal solutions for CAL case: examine where these solutions are invariant.

When 22 tRNA genes are excluded, a number of long segments are found (see right).

Therefore, only the questions about the correct ordering of these segments and of their orientations remains.

Ancestral nematode: unique reconstruction
nad6 cob cox3 nad4L rns nad1 atp8 atp6 nad2 nad4 cox1 cox2 rnl nad3 nad5

Protostome/deuterostome ancestor: 4 segments
(nad2 cox1 cox2 atp8 atp6 cox3 nad3) (nad4L nad4 nad5) (nad6 cob) (rns rnl nad1)

Ancestral protostome: 4 segments
(nad2 cox1 cox2 atp8 atp6 cox3 nad3) (nad4L nad4 nad5) (nad6 cob) (rns rnl nad1)

Ancestral deuterostome: unique reconstruction
cox1 cox2 atp8 atp6 cox3 nad3 nad4L nad4 nad5 -nad6 cob rns rnl nad1 nad2

Ancestral echinoderm: 2 segments
(cox1 nad4L cox2 atp8 atp6 cox3 nad3 nad4 nad5 -nad6 cob rns) (nad1 nad2 rnl)

Ancestral arthropod: unique reconstruction
nad6 cob -nad1 -rnl -rns nad2 cox1 cox2 atp8 atp6 cox3 nad3 -nad5 -nad4 -nad4L

Annelid/mollusc ancestor: 2 segments
(nad6 cob) (-nad1 -rnl -rns cox3 nad3 nad2 cox1 cox2 atp8 atp6 -nad5 -nad4 -nad4L)

Ancestral mollusc: 2 segments
(nad6 cob) (-nad1 -rnl -rns cox3 nad3 nad2 cox1 cox2 atp8 atp6 -nad5 -nad4 -nad4L)

Ancestral snail: unique reconstruction
nad6 nad5 nad1 nad4L cob cox2 -atp8 -atp6 -rns -nad3 -cox3 nad4 nad2 cox1 rnl

Blanchette, Sankoff, J Mol Evol (1999)
Drawbacks of breakpoint analysis: costly + ambiguous

Let us consider a simple example:
Suppose that the genomes $G_1$, $G_2$, and $G_3$, evolved from the ancestral genome $A = 1 2 3 4 5 6$ by one reversal each such that

\[
\begin{align*}
G_1 &= 1\ 2\ -4\ -3\ 5\ 6 \\
G_2 &= 1\ -4\ -3\ -2\ 5\ 6 \\
G_3 &= 1\ 2\ 3\ 4\ -5\ 6
\end{align*}
\]

Searching for the breakpoint median will produce 4 optimal solutions. $A$, but also $G_1$, $G_2$, and $G_3$. If the median is $A$, then we have two breakpoints on each edge of the tree for a total of six.
But if the median is any of the three genomes, we also get a total of $6 = 0 + 3 + 3$ breakpoints.
Therefore, the breakpoint median fails to unambiguously identify the ancestor.
Multiple Genome Rearrangement Problem

Find a phylogenetic tree describing the most „plausible“ rearrangement scenario for multiple species.

The genomic distance in the case of genome rearrangement is defined in terms of (1) reversals, (2) translocations, (3) fusions, and (4) fissions which are the most common rearrangement events in multichromosomal genomes.

The special case of three genomes \((m = 3)\) is called the Median Problem. Given the gene order of three unichromosomal genomes \(G_1, G_2,\) and \(G_3\), find the ancestral genome \(A\) which minimizes the total reversal distance

\[
d(A, G_1) + d(A, G_2) + d(A, G_3)
\]
Multiple Genome Rearrangement Problem

The breakpoint analysis attempts to solve the Median Problem by minimizing the breakpoint distance instead of the reversal distance. However, the breakpoint distance, in contrast to the reversal distance, does not correspond to a minimum number of rearrangement events! As a result, the breakpoint, recovered by breakpoint analysis, rarely corresponds to the ancestral median, the genome that minimizes the overall number of rearrangements in the evolutionary scenario.

New approach:
Given a set of \( m \) permutations (existing genomes) or order \( n \), find a tree \( T \) with the \( m \) permutations as leaf nodes and assign permutations (ancestral genomes) to internal nodes such that \( D(T) \) is minimized, where

\[
D(T) = \sum_{(\pi, \gamma) \in T} d(\pi, \gamma)
\]

is the sum of reversal distances over all edges of the tree.
New algorithm

Aim: Among all possible reversals for each of the three genomes identify good reversals.

A good reversal $\rho$ in a genome $G_1$ is a reversal that brings a genome closer to the ancestral genome. But since this is unknown, it is unclear to find good reversals, oops!

Instead: assume that reversals that reduce the reversal distance between $G_1$ and $G_2$ and the reversal distance between $G_1$ and $G_3$ are likely to be good reversals.

With $\Delta(\rho)$ as the overall reduction in the reversal distances:

$$\Delta(\rho) = (d(G_1, G_2) + d(G_1, G_3)) - (d(G_1 \cdot \rho, G_2) + d(G_1 \cdot \rho, G_3))$$

the reversal $\Delta(\rho)$ is good if $\Delta(\rho) = 2$. 
New algorithm

Iteratively carry on these good rearrangements until the genomes $G_1$, $G_2$, and $G_3$ are transformed into an identical genome, hoping that this is the most likely “ancestral median“.

When we are dealing with multichromosomal genomes and with four different types of rearrangements, ambiguous situations may occur too.
Ambiguities again possible

E.g. \[ G_1 = \begin{array}{c} 1 \ 2 \ 3 \ 4 \ 5 \end{array} \]
\[ G_2 = \begin{array}{c} 1 \ 2 \ -5 \ -4 \ -3 \end{array} \]
\[ G_3 = \begin{array}{c} 1 \ 2 \ 3 \ 4 \ 5 \end{array} \]

The parsimony principle does not allow to unambiguously reconstruct the evolutionary scenario.

If the ancestor coincides with \( G_1 \), then a reversal occurred on the way to \( G_2 \), and a fission occurred on the way to \( G_3 \).

One can as well start with \( G_2 \) or \( G_3 \) as the ancestors. In this case

\[ d(G_1, G_2) = d(G_1, G_3) = d(G_2, G_3) \]

This kind of ambiguity does not exist for unichromosomal genomes because, there, it is impossible to find 3 genomes that would all be within one reversal of each other.
Strategy for choosing reversals

Therefore one has to select carefully among the good rearrangements. Observe that in most genomes of interest reversals and translocations are more common than fusions and fissions.

Therefore use as a rule always to select reversals/translocations before fusions/fissions.

Often, the list of good reversals contains nonoverlapping reversals, and the order in which these reversals are performed is often irrelevant. Compute for each good reversal $r$ the number of good reversals $n_r$ that will be available if $r$ is carried out. Then choose the good reversal with the maximal $n_r$ to be carried out.

If we run out of good reversals before reaching a solution, the best reversal to be taken will be the result of a depth $k$ search minimizing the total pairwise rearrangement distances.
How good measure is reversal distance?

Authors claim that the reversal distance is a good approximation of the true distance for many biologically relevant cases.

Let $\gamma$ be a genome that evolved from a genome $\pi$ by $k$ reversals. I.e. the true distance between $\pi$ and $\gamma$ is $k$.

We say that $\pi$ and $\gamma$ form a valid pair if $d(\pi, \gamma) = k$. Otherwise we say that $d(\pi, \gamma)$ underestimates the true distance.

Typically two genomes form a valid pair if the number of rearrangements between them is relatively small – exactly the case in a number of genome rearrangement studies.
Reversal distance vs. True distance

Reversal distance, $d(\pi, \gamma)$, versus the actual number of reversals performed to transform $\pi$ into $\gamma$, where $\gamma$ is a genome/permutation that evolved from the identity permutation $\pi = 1, 2, \ldots, 100$ by $k$ random reversals. The simulations were repeated 10 times for every $k$. Shown is the average difference between the reversal distance and the actual number of reversals performed ($k$).

For a genome with $n=100$ markers, the reversal distance approximates the true distance very well as long as the number of reversals remains below $0.4 \ n$. This is the case in many biological relevant cases.

Bourque, Pevzner, Genome Res (2002)
Test on simulated data

Starting from the identity permutation $A$ with $n$ genes/markers. $n = 30$ or 100. $k$ reversals were performed to get genome $G_1$, $k$ to get genome $G_2$, and $k$ to get genome $G_3$.

Use these as input to MGR-MEDIAN and GRAPPA. Check whether programs reconstruct the ancestral identity permutation.

The simulations were repeated 10 times for every ratio $\#\text{reversals}/\#\text{markers} = 3k/n$. 
Comparison of MGR-MEDIAN and GRAPPA

(a) and (b) show the average difference between the number of reversals on the tree recovered by the algorithm and the number of reversals on the actual tree (equal to $3k$).

(c) and (d) show the average reversal distance between the solution recovered and the actual ancestor.

GRAPPA and MGR-MEDIAN produce very similar solutions for $r < 0.20$. As ratio $r$ increases, GRAPPA starts making errors. MGR-MEDIAN sometimes finds solutions that even have fewer reversals than the actual ancestor. **Reason**: for increasing $r$, assumption that the ancestor corresponds to the most parsimonious scenario sometimes fails.

Bourque, Pevzner, Genome Res (2002)
Tests on simulated data: non-equidistant genomes

The genomes $G_1$, $G_2$, and $G_3$ are obtained by $k$, $k$, and $2k$ reversals, each from the ancestral identity permutation $1 \ 2 \ ... \ n$ ($n = 30$ and $n = 100$). The simulations were repeated 10 times for every ratio $\#\text{reversals}/\#\text{markers} = 4k/n$.

Figs (a) - (d) have same meaning as on previous figure. Same behavior is found.

Also test for 4 – 10 genomes. GRAPPA can’t do more than 10 genomes because the tree space is too large.

Bourque, Pevzner, Genome Res (2002)
Herpes simplex virus (\textit{HSV}), Epstein-Barr virus (\textit{EBV}), and Cytomegalovirus (\textit{CMV}) gene orders (Hannenhalli et al. 1995) as well as the ancestral gene order (A) and optimal evolutionary scenario recovered by MGR-MEDIAN.

MGR finds solution with 7 reversals, GRAPPA finds 8 reversals.

Here, the ratio \( r \) of \#reversals / \#markers is \( \frac{7}{25} = 0.28 \).
mtDNA of human, fruit fly, and sea urchin

Human, sea urchin, and fruit fly mitochondrial gene order taken from Sankoff et al. (1996). A is the ancestral gene order suggested by MGR-MEDIAN.

| Human:     | 26 13 17 12 -24 15 18 32 -2 -16 -3 -33 4 -28 7 5 1 10 19 25 22 11 29 14 20 -21 -8 6 30 -23 9 27 31 |
| Fruit fly: | -26 -31 -27 12 -24 15 18 32 -3 -33 4 13 5 7 1 10 19 2 25 16 29 8 -9 -20 -11 -22 30 -23 21 6 28 -17 -14 |
| A:         | 26 13 17 12 -24 15 18 32 -28 7 -6 21 -20 -29 -11 -22 -25 -16 8 -3 -33 4 14 -2 -19 -10 -1 -5 30 -23 9 27 31 |

Solution found is different from Sankoff et al. but the total reversal distance (39) is the same.
Here, the ratio of #reversals / #markers is $39/33 = 1.18$, marking this as a difficult problem.
Running GRAPPA on these genomes gives a solution with a total reversal distance of 43.

Bourque, Pevzner, Genome Res (2002)
Metazoan mtDNA data

Data (36 common genes) of 11 metazoan genomes that was studied before by BPA.

Shown here: Phylogeny reconstructed by MGR. The genomes come from 6 major metazoan groupings: nematodes (NEM), annelids (ANN), mollusks (MOL), arthropods (ART), echinoderms (ECH), and chordates (CHO). Numbers show the number of reversals (150 in total).

Tree is very similar to that of Blanchette et al. that was constructed in a semiautomated fashion. GRAPPA finds after 48 CPUhours three optimal trees with 175 reversals and 200 breakpoints.

Bourque, Pevzner, Genome Res (2002)
Campanulaceae cpDNA data

*Campanulaceae* chloroplast with 13 cpDNAs and 105 markers. The tree space for 13 genomes cannot be searched exhaustively by GRAPPA. Therefore, trees were constrained by Moret et al. (2001). They found 216 trees with a total of 67 reversals. MGR (without using constraints) gives a tree with 65 reversals. Tree topology corresponds to GRAPPA tree but labelling of internal nodes differs.

Bourque, Pevzner, Genome Res (2002)
Ancestral median for human, mouse, and cat

Most existing comparative maps of multichromosomal species are pairwise maps representing genome organisation of two species.

Number of established universal markers is relatively small.

First sufficiently detailed triple comparative map from rat and cat comparative mapping projects. Here: integrate pairwise human-mouse, human-cat, and mouse-cat comparative maps into a triple human-mouse-cat map. Murphy et al. (2000) identified 193 markers shared by all 3 species. This number of markers is still too small to derive a detailed rearrangement scenario.
Ancestral median for human, mouse, and cat

Problem: comparative maps usually correspond to unsigned permutations = no direction is available on the orientation of the genes. Note that the algorithmic complexity of this problem is NP-hard.

Therefore assign orientation to markers. Use strips in unsigned permutations to infer the signed permutations from the original unsigned permutations. Using the human genome as a reference, all strips in both mouse and cat genomes were identified and assigned an orientation. Any marker where no orientation could be assigned (79) was removed.

This is an application to a multichromosomal problem.
**Ancestral median for human, mouse, and cat**

**Numbers** above chromosomes correspond to 114 markers. Numbering is such that human genome corresponds to the identity permutation broken into 20 pieces. **Names** below chromosomes correspond to the name of the markers. For comparison, each human chromosome is shown in a different **color**. Each marker **segment** is also traversed by a **diagonal line**. These diagonals are such that the human chromosomes are traversed from top left to bottom right and are designed to provide visual help to see where rearrangements occurred. E.g., for chromosome X, the gene order of the ancestor coincides with the cat gene order and only differs by one segment consisting of genes 108 and 109 (break in the diagonal line) from the human gene order. The mouse X chromosome is broken into 7 segments compared to the ancestor.

---

Bourque, Pevzner, Genome Res (2002)
Summary

Breakpoint analysis (BPA) is a robust technique for small rearrangement problems. Problem of ambiguity between different optimal solutions. Although complexity could be dramatically reduced by algorithmic improvements (e.g. GRAPPA), method is still too expensive for more than 10 genomes.

Heuristic algorithm by Bourque & Pevzner minimizes reversal distance instead of breakpoint distance. (Recall from lecture 5 that (number of breakpoints) \times 2 was not the optimal lower bound for the reversal distance.)
Runs more efficient + can be applied to much larger problems + provides only one or a few solutions.

Analogy to conformational search in some energy landscape ...

The problem remains what is the correct way to identify the biologically correct = true evolutionary trees: by minimizing the breakpoint distance or the reversal distance or something else?