The inversion process in bacteria: distance metrics with group-theoretic models

Andrew Francis

Centre for Research in Mathematics
School of Computing, Engineering and Mathematics
University of Western Sydney

Phylomania

7th November, 2013.

Distance

Why think about distance?

- Science wants to quantify difference, to compare, to measure.
- We want to organise information and knowledge about life, relating organisms by phylogeny.

Distance provides the input to several important phylogeny methods.
(UPGMA, Neighbour-joining)

Distance in bacteria

we use *large-scale* rearrangements

- Why large-scale?

  Because standard eukaryotic methods (looking at a particular gene and SNPs on the gene) might be confounded by horizontal gene transfer in bacteria: differences might not be due to vertical heredity.

Distance in bacteria

we use *large-scale* rearrangements

- Why large-scale?

  Because standard eukaryotic methods (looking at a particular gene and SNPs on the gene) might be confounded by horizontal gene transfer in bacteria: differences might not be due to vertical heredity.
- Large scale rearrangements are studied by identifying preserved regions ("locally colinear blocks") in a family of taxa.
- Inversions take a segment — a sequence of regions — and reverse their order.

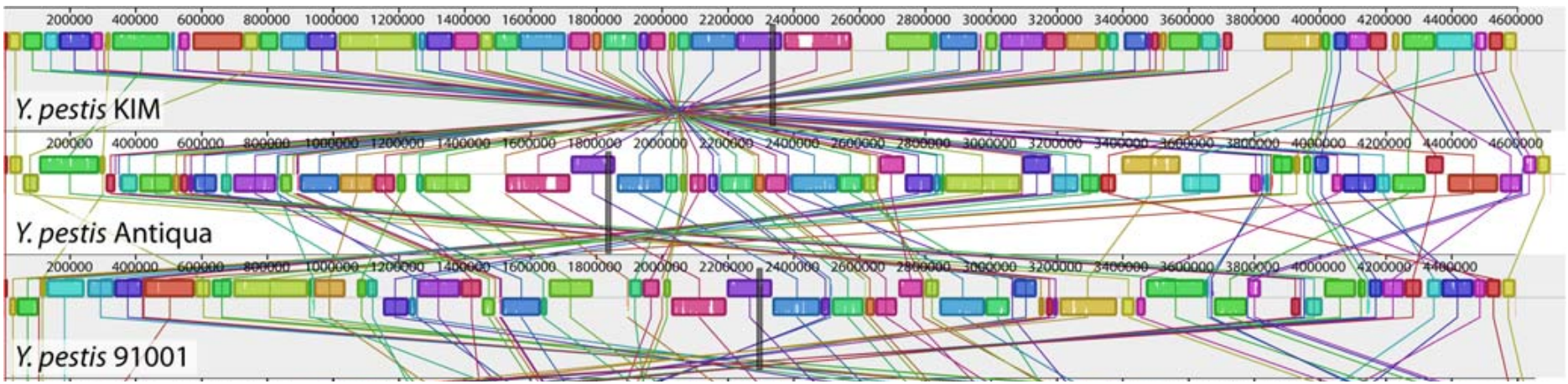


Figure from Darling et al, 2008.

Large-scale rearrangements $\longrightarrow$ genomes as permutations

- If we identify preserved regions we can treat each as a unit and regard all taxa as rearrangements of regions.
- Numbering regions $1, \ldots, n$ makes each genome a permutation.
  - Incorporating orientation of regions gives a *signed* permutation.
- This assumes
  - all regions are the same size, and
  - they are evenly distributed around the genome.

Standard model

no, not physics

- Standard models in the literature assume
  - that all inversions are possible, and
  - that all are equally probable.

Standard model

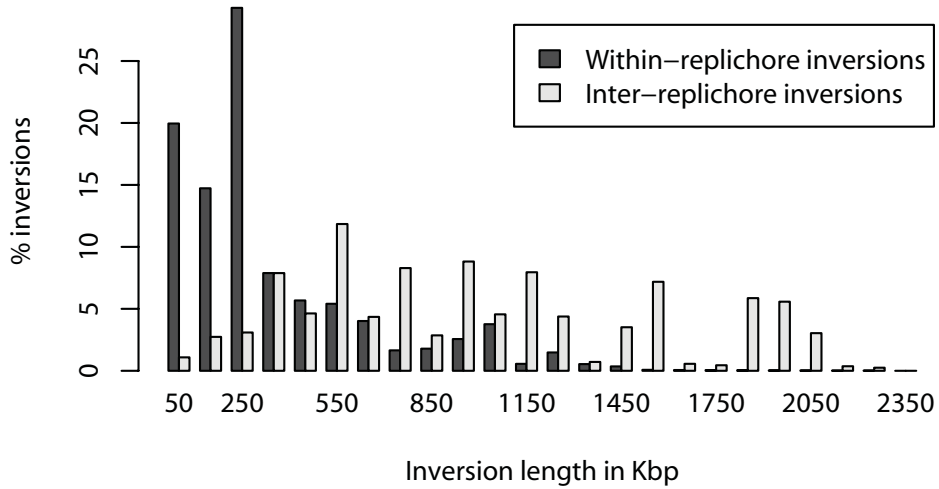
no, not physics

- Standard models in the literature assume
  - that all inversions are possible, and
  - that all are equally probable.
- This means that circular arrangements can be dealt with as linear arrangements
  - because inversions across any given point can be performed on the complementary segment.
- There are fast algorithms for solving the inversion distance problem in this case, using the "breakpoint graph" (Bafna and Pevzner 1993).

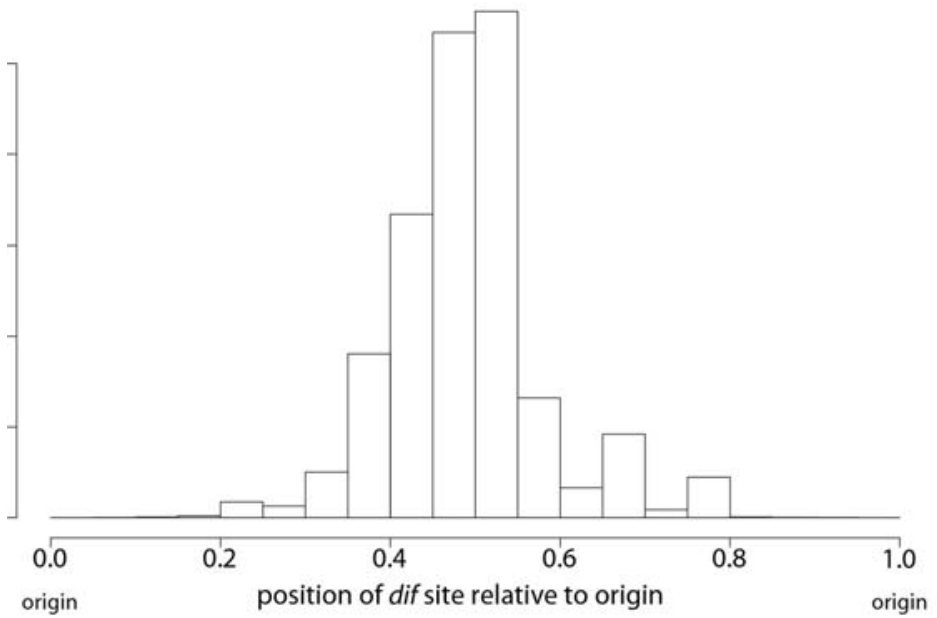
However

Not all inversions are equally likely.

- Length: shorter ones are more likely.



- Location: ones that fix terminus more likely.



[Figures from Darling *et al*, 2008.]

Group-theoretic approach

- Incorporating these constraints makes cutting-linearizing invalid.

  $\Longrightarrow$ We must model permutations on the circle.
- There are two features of permutations on a circle:
  - inversions can occur across any cut, e.g $(n, 1)$.
  - there is circular symmetry — the action of the dihedral group.

Group-theoretic approach

- Incorporating these constraints makes cutting-linearizing invalid.

  $\Longrightarrow$ We must model permutations on the circle.
- There are two features of permutations on a circle:
  - inversions can occur across any cut, e.g $(n, 1)$.
  - there is circular symmetry — the action of the dihedral group.
- We can consider *the group generated by the inversions*, acting on the set of all possible genomes.
- The distance problem becomes a question of a length function in the group.
  - Or the distance between vertices on the *Cayley graph* of the group.

Group-theoretic approach

- Incorporating these constraints makes cutting-linearizing invalid.

  $\Longrightarrow$ We must model permutations on the circle.
- There are two features of permutations on a circle:
  - inversions can occur across any cut, e.g $(n, 1)$.
  - there is circular symmetry — the action of the dihedral group.
- We can consider *the group generated by the inversions*, acting on the set of all possible genomes.
- The distance problem becomes a question of a length function in the group.
  - Or the distance between vertices on the *Cayley graph* of the group.
- We also need to consider equivalence under the action of the dihedral group — not a normal subgroup so simply a (co)set of vertices on the Cayley graph.

There are a range of models

all colours and sizes to suit every household

- **Orientation**:
  1. If we ignore it, we work in the symmetric group
  2. If we include it, we work in the hyperoctahedral group.

There are a range of models

all colours and sizes to suit every household

- **Orientation**:
  1. If we ignore it, we work in the symmetric group
  2. If we include it, we work in the hyperoctahedral group.

- **Terminus fixing**: we work in a stabilizer subgroup.
  - [see talk by Stuart Serdoz after lunch]

There are a range of models

all colours and sizes to suit every household

- **Orientation**:
  1. If we ignore it, we work in the symmetric group
  2. If we include it, we work in the hyperoctahedral group.

- **Terminus fixing**: we work in a stabilizer subgroup.
  - [see talk by Stuart Serdoz after lunch]

- **Restrict inversions by length**:
  1. Change generating set: choose subset of inversions that are allowed. (example to follow)
  2. Give longer inversions higher weight.
     [ongoing work with Praeger and Niemeyer, UWA]

There are a range of models

all colours and sizes to suit every household

- **Orientation**:
  1. If we ignore it, we work in the symmetric group
  2. If we include it, we work in the hyperoctahedral group.

- **Terminus fixing**: we work in a stabilizer subgroup.
  - [see talk by Stuart Serdoz after lunch]

- **Restrict inversions by length**:
  1. Change generating set: choose subset of inversions that are allowed. (example to follow)
  2. Give longer inversions higher weight.
     [ongoing work with Praeger and Niemeyer, UWA]

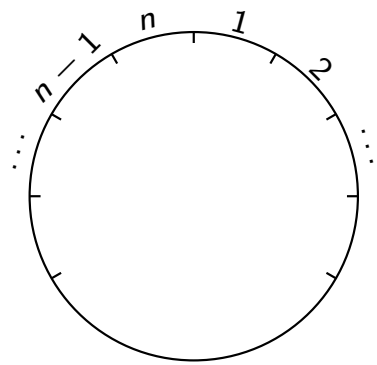
- The approach allows generalizations such as "Double-Cut-and-Join" (Bergeron-Mixtacke-Stoye, 2006).
  - [See talk by Sangeeta Bhatia after lunch]

Example

Two region inversion model

- The 2-region inversions that generate the group are the simple transpositions of adjacent regions.

- . . . noting that they now include $s_n = (n\ 1)$, because we are on the circle.
- We need to use the affine symmetric group.

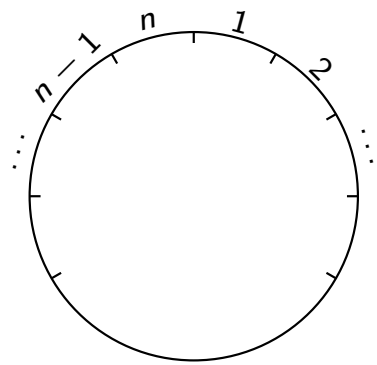


Example

Two region inversion model

- The 2-region inversions that generate the group are the simple transpositions of adjacent regions.

- . . . noting that they now include $s_n = (n\ 1)$, because we are on the circle.
- We need to use the affine symmetric group.



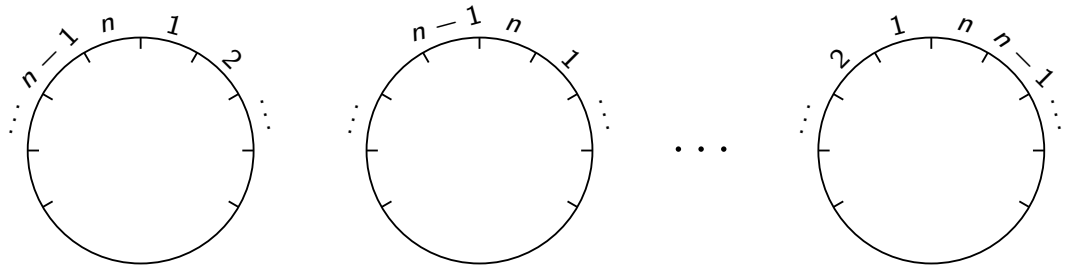
Theorem

If σ *is a minimal length affine permutation representing a circular permutation,* **then** σ *takes the shortest distance between each* i *and* $\sigma(i)$ mod n.

Group-theoretic models of the inversion process in bacterial genomes, Egri-Nagy, Gebhardt, Tanaka & Francis, *J Mathematical Biology*, Online June 2013.

The resulting algorithm

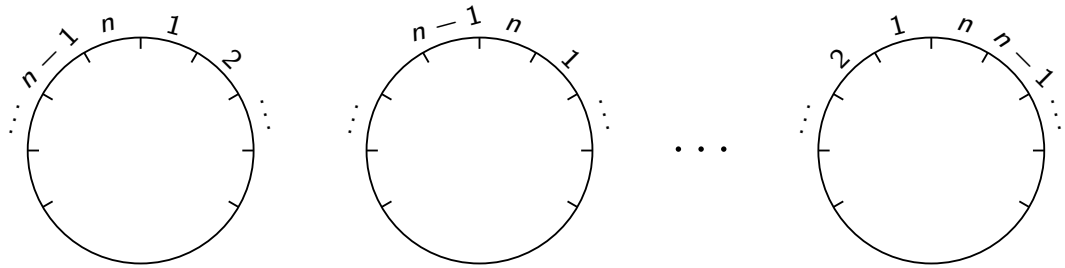
1. For each frame of reference,



   draw an affine permutation with minimal distances for each i.
2. The minimal length of these $2n$ choices is the inversion distance.

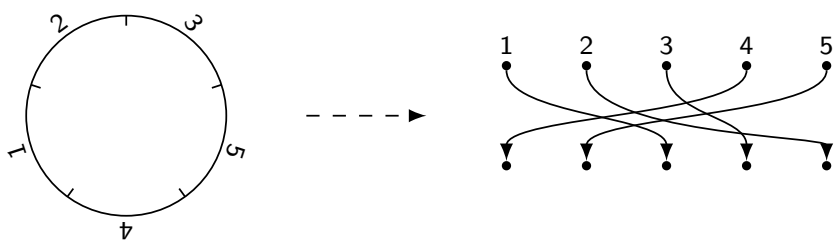
The resulting algorithm

1. For each frame of reference,



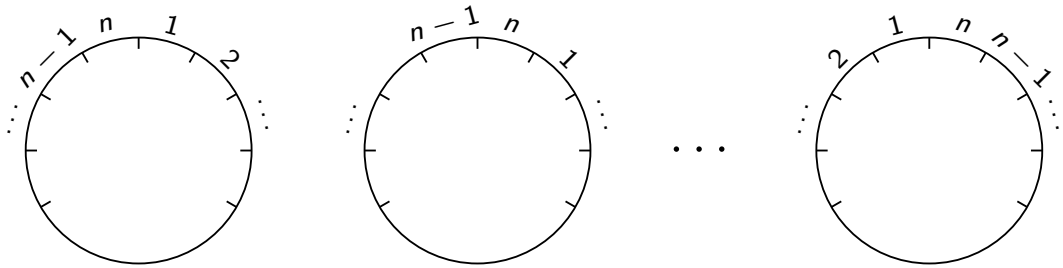
   draw an affine permutation with minimal distances for each i.
2. The minimal length of these $2n$ choices is the inversion distance.

Example: $\sigma = [3, 5, 4, 1, 2]$:



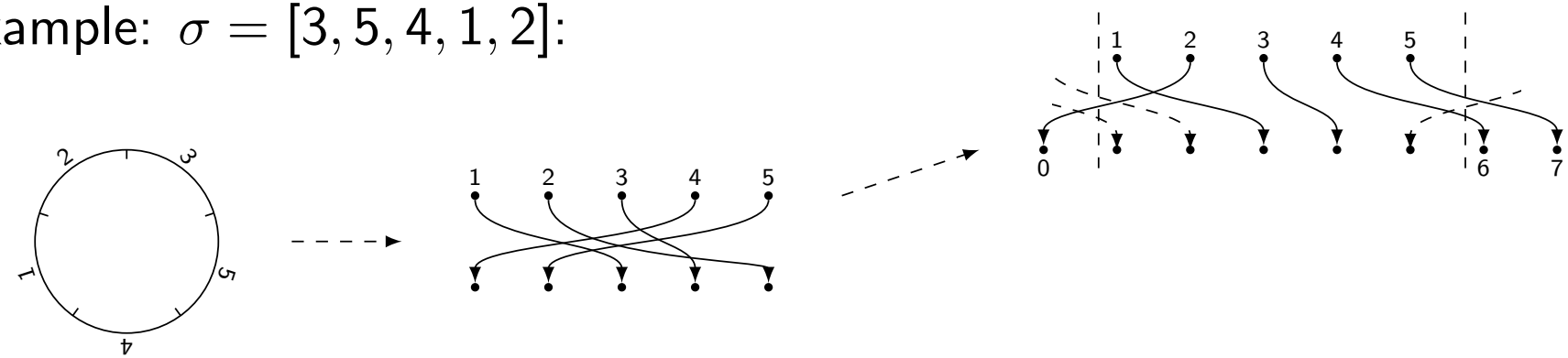
The resulting algorithm

1. For each frame of reference,



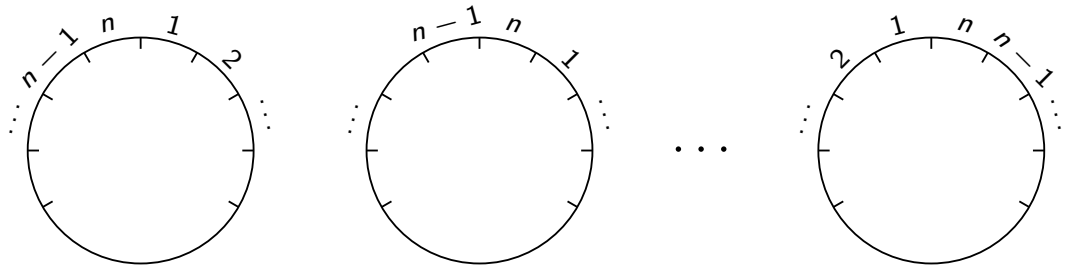
   draw an affine permutation with minimal distances for each i.
2. The minimal length of these $2n$ choices is the inversion distance.

Example: $\sigma = [3, 5, 4, 1, 2]$:



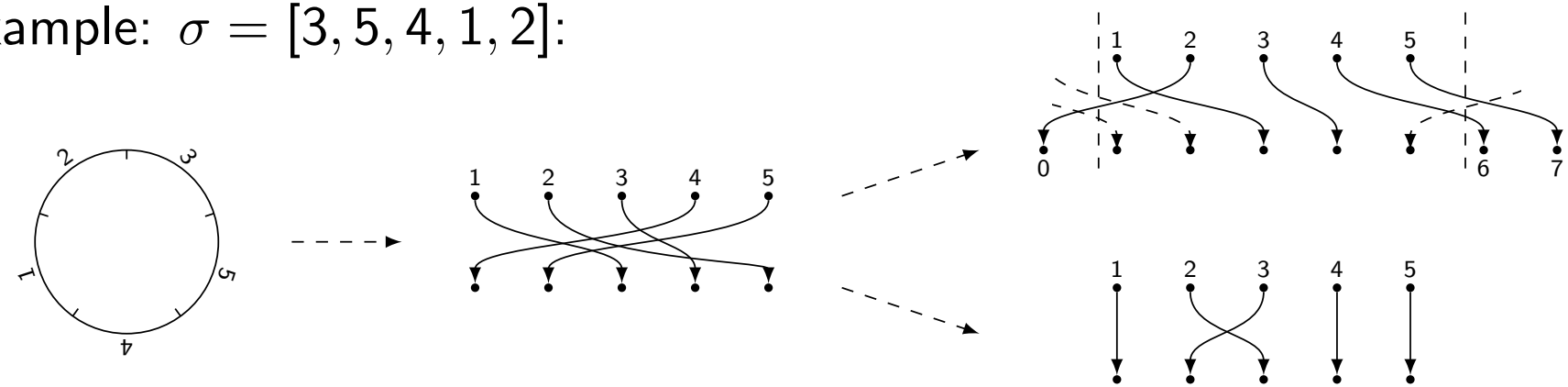
The resulting algorithm

1. For each frame of reference,



   draw an affine permutation with minimal distances for each i.
2. The minimal length of these $2n$ choices is the inversion distance.

Example: $\sigma = [3, 5, 4, 1, 2]$:



Further questions

Phylogeny

We can regard the phylogeny problem as the problem of finding a minimal spanning tree of a set of vertices in the Cayley graph where the taxa we wish to relate are vertices on the graph and we want to minimise the total path length.

Further questions

Phylogeny

We can regard the phylogeny problem as the problem of finding a minimal spanning tree of a set of vertices in the Cayley graph where the taxa we wish to relate are vertices on the graph and we want to minimise the total path length.

Is "distance" answering the right question?

1. Maybe we want the "expected distance". The minimal distance can only underestimate the true distance; when the rate of inversion is high it may *badly* underestimate it. [Stuart Serdoz again, after lunch]

Further questions

Phylogeny

We can regard the phylogeny problem as the problem of finding a minimal spanning tree of a set of vertices in the Cayley graph where the taxa we wish to relate are vertices on the graph and we want to minimise the total path length.

Is "distance" answering the right question?

1. Maybe we want the "expected distance". The minimal distance can only underestimate the true distance; when the rate of inversion is high it may *badly* underestimate it. [Stuart Serdoz again, after lunch]
2. In a random walk on the Cayley graph of a given length some arrangements are more probable than others. You can wake up now: Attila will discuss.

Further questions

Phylogeny

We can regard the phylogeny problem as the problem of finding a minimal spanning tree of a set of vertices in the Cayley graph where the taxa we wish to relate are vertices on the graph and we want to minimise the total path length.

Is "distance" answering the right question?

1. Maybe we want the "expected distance". The minimal distance can only underestimate the true distance; when the rate of inversion is high it may *badly* underestimate it. [Stuart Serdoz again, after lunch]
2. In a random walk on the Cayley graph of a given length some arrangements are more probable than others. You can wake up now: Attila will discuss.

Thank you for listening, thanks to the organisers for organising, and thanks to the ARC for funding.