The Private Life of Bacteria

Daniel Lawson (University of Bristol)
Xavier Didelot (University of Warwick)
Peter Green (University of Bristol)
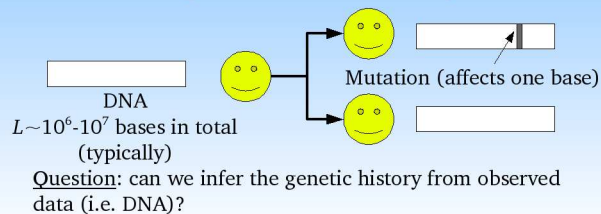Daniel Falush (University College Cork)

Inferring genetic history using the "weak" Ancestral Recombination Graph
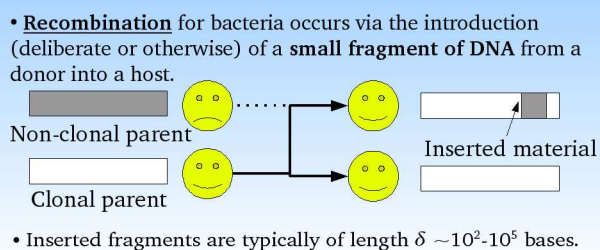
Abstract:

We can model phylogeny and recombination history using the biologically meaningful Ancestral Recombination Graph. This model cannot normally be used for inference but in the case of bacteria a suitable approximation can be found that is both mathematically rigorous and biologically useful.
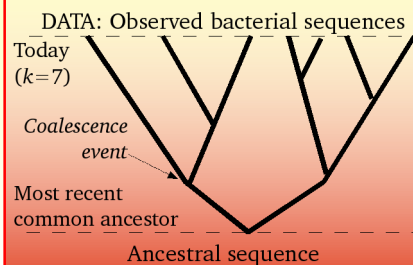
Bacteria Reproduce asexually...

DNA
$L \sim 10^6 \text{-} 10^7$ bases in total (typically)

Mutation (affects one base)

Question: can we infer the genetic history from observed data (i.e. DNA)?

... but they still exchange DNA

- **Recombination** for bacteria occurs via the introduction (deliberate or otherwise) of a **small fragment of DNA** from a donor into a host.

Non-clonal parent

Clonal parent

Inserted material

- Inserted fragments are typically of length $\delta \sim 10^2 \text{-} 10^5$ bases.

"weak" ARG

The same as the Ancestral Recombination Graph except:

- Only recombination of clonal lineages allowed
- Only coalescence events involving at least one clonal lineage is allowed

Notes:

- Reduces graph space.

- Successful inference:
  - MLST data (i.e. data consists of many small sections of DNA)
  - Full genomes

- Good approximation when $\delta < L/\rho$.

weak ARG
Clonal lineages
Recombinant lineages
Ancestral sequence

Implementation

Implemented as a Bayesian inference algorithm using Markov Chain Monte Carlo (MCMC).

Program to be made publicly available in 2009 – contact the authors for details.

Features optional GUI, parallel processing version, optional mixing over trees, and more...

The Coalescent Model

- Start with the observed bacterial sequences;
- Work **backwards in time** to the ancestral sequence.

● Coalescence of lineages occurs at rate $k(k-1)/2$ where k is the number of lineages.

Clonal Tree

DATA: Observed bacterial sequences

Today ($k=7$)

Coalescence event

Most recent common ancestor

Ancestral sequence

Notes:

- Quickly gives probability of a tree.

- Mutation models give the probability of the data given the tree.
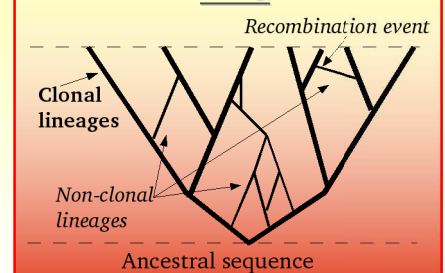
- No recombination.

Ancestral Recombination Graph (ARG)

- Coalescence of lineages occurs at rate $k(k-1)/2$.

● Recombination – **splitting lineages** – occurs at rate $k\rho/2$.

Notes:

- Quickly gives the probability of a graph.

- But: **too many graphs** to search.

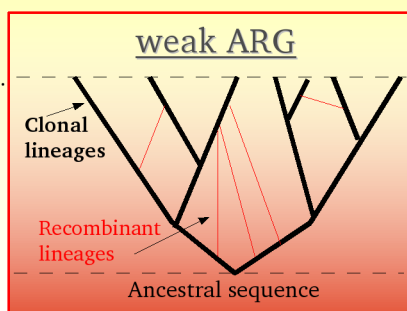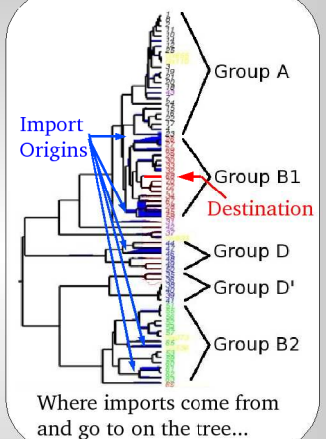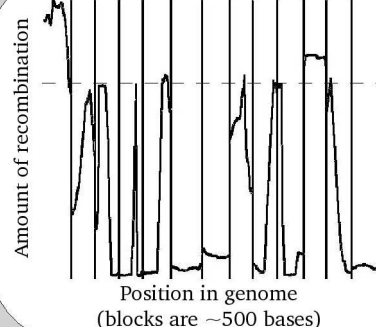- Inference is impossible in most real cases.

ARG

Recombination event

Clonal lineages

Non-clonal lineages

Ancestral sequence

Results for *E. coli*

- Actual recombination events
- Density of recombination on the tree
- Recombination "hotspots"
- Imports within vs. without the tree

Import Origins

Group A

Group B1
Destination

Group D

Group D'

Group B2

Where imports come from and go to on the tree...

Amount of recombination

Position in genome (blocks are ~500 bases)

.. and which segments of which genes, and how certain the import is.