Bacterial population genomics

Peter Pfaffelhuber
joint with Franz Baumdicker, Wolfgang Hess
University of Freiburg
Banff, September 2009

Genomic bacterial data

- **Distributed Genome Hypothesis**
  Bacteria possess a genome that is much larger than the genome of a single bacterium

The supragenome

- **Supragenome**
  Gene frequency spectrum
- **Predicted**
  using a test dataset of 8 individuals

- Data from 13 *Haemophilus influenzae* strains (Hogg et al 2007)

Modelling genomic diversity

- **Goal**: describe diversity of genes in a bacterial population
- **Genealogy**: given by Kingman’s coalescent
**Phylogenetic trees based on gene content**

Daniel H. Huson¹,² and Mike Steel³

- **New genes** taken from the environment at rate $\theta$
- Present **genes lost** at rate $\rho$
- A set of **core genes** cannot be lost

**The infinitely many genes model**

- Gene gain: rate $\theta$
- Gene loss: rate $\rho$

**Single lines**

- Present, $t=T$

**Data, Genealogies and Mutations**

- Individual 1: - - - ✓ -
- Individual 2: ✓ ✓ ✓ ✓ -
- Individual 3: ✓ ✓ ✓ ✓ ✓
- Individual 4: ✓ ✓ ✓ ✓ ✓

Is the genomic pattern compatible with neutral evolution?
Questions

- How many **genes** does a single individual carry?
- How many **different genes** are there in the sample?
- How many **new genes** are there in the nth individual?
- What does the **gene frequency spectrum** look like?

Single individual

- **$|G_i|$**: number of genes in individual i
- $\frac{\theta}{2} dt$: average number of genes gained a time $t$ in the past
- $p(t)$: probability that a gene gained at time $t$ not lost
  \[ p(t) = e^{-\frac{\theta}{2} t} \]

Proposition

\[ |G_i| \sim \text{Po}(\int_0^\infty e^{-\frac{\theta}{2} t} \frac{\theta}{2} dt) = \text{Po}(\frac{\theta}{\rho}) \]

The gene frequency spectrum

- $G_i$: Number of genes present in $i$ individuals
- **Theorem** For the gene frequency spectrum
  \[ E[|G_i|] = \theta \sum_{k=0}^{n-1} \frac{1}{k+\rho} \]

\[ \text{Corollary} \]

\[ E[\text{new genes in } (n+1)st \text{ individual}] = \frac{\theta}{n+\rho}. \]
**Incongruent pairs**

- A pair of genes is **incongruent**, if

![Diagram showing incongruent gene pairs]

- **Theorem**
  
  Let $P$ be the number of pairs of incongruent genes

  $$
  E[P] = \frac{\theta^2 \rho}{4} \left( \frac{18 + 11 \theta^2 + 203 \theta^2 + 108 \theta^2}{(1+\theta^2)(1+2\theta^2)(1+4\theta^2)(3+4\theta^2)(3+5\theta^2)(6+5\theta^2)(7+5\theta^2)} \right).
  $$

**Prochlorococcus**

- **Tiny**: length $\sim 0.6 \mu m$, Genome size 2Mbp
- **Smallest known photosynthetic** bacterium
- **Abundant**: $\sim 10^5$ cells per ml (in the ocean)
- **Structure**: by water depth
- **Recently discovered**: first isolated in 1988

**Fit of model and data**

- **Data**
  - 1888 genes per genome on average
  - 1033 singleton genes

![Bar chart showing distribution of genes per genome]

- **Estimates**
  - $\hat{\theta} = 1135.27$, $\hat{\rho} = 1.94$, number of core genes = 1268.

**Outlook**

- **Biological questions**
  - Are estimates **biologically realistic**?
  - What are rates of gene gain/loss **per generation**?
  - Which **function** do non-core genes have?

- **Mathematical questions**
  - How can we **test** neutral evolution in the infinitely many genes model?
  - How do **selection, population structure, lateral gene transfer** change predicted patterns?