Horizontal gene transfer in bacteria: the ancestral gene transfer graph

Peter Pfaffelhuber

joint work with Franz Baumdicker

University of Freiburg

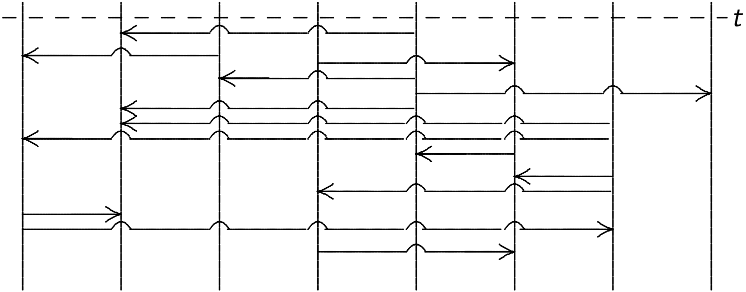
Torun, June 2013

Populations genetics

- **evolution of variation in natural populations**
- **Biology**: part of evolutionary theory
  Keywords: mutation, selection, genetic drift, recombination(=horizontal gene transfer)
- **Mathematics**: dynamical systems, stochastic processes
  Keywords: Branching processes, measure-valued diffusions, coalescent, random genealogy

The Moran model

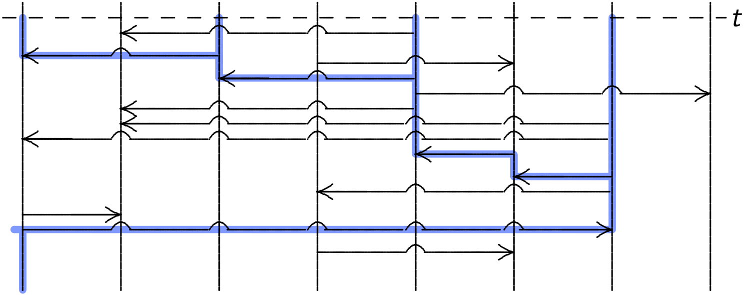
- A population consists of N individuals
- Each pair of individuals **resamples** at rate 1
- Resampling means: one individual **dies**, the other **reproduces**



The Moran model

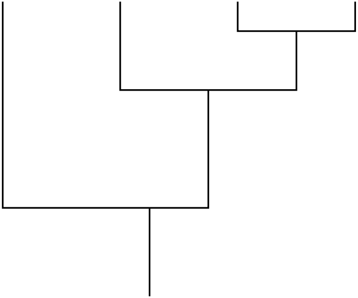
- A population consists of N individuals
- Each pair of individuals **resamples** at rate 1
- Resampling means: one individual **dies**, the other **reproduces**

Ancestral lineages coalesce



Genealogy of samples

- Kingman's **coalescent** describes ancestry of population samples
    - Start with n lines (n=sample size)
    - If k lines left, merge two lines after $\exp\left(\binom{k}{2}\right)$ waiting time



Genomic bacterial data

- **Observation**
  Not all bacteria of one population carry the same genes
- **Difference** in number of genes: up to 25%

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7
Individual 1	–	–	–	–	✓	✓	–
Individual 2	✓	✓	✓	–	–	–	–
Individual 3	✓	✓	✓	✓	–	–	✓
Individual 4	✓	✓	✓	✓	–	–	✓

✓ gene present

– gene absent

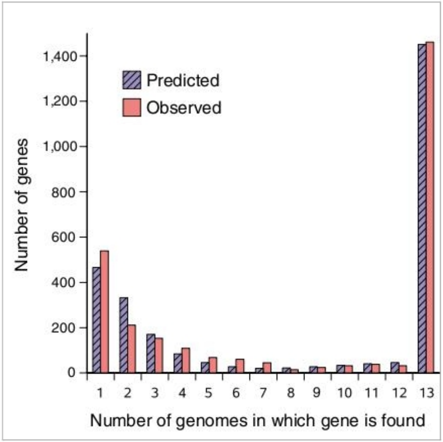
Genomic bacterial data

- **Distributed Genome Hypothesis**

  Bacteria possess a genome that is much larger than the genome of a single bacterium

The gene frequency spectrum

- Predicted
  using a test dataset of 8 individuals
- Data from 13 *Haemophilus influenzae* strains (Hogg et al 2007)



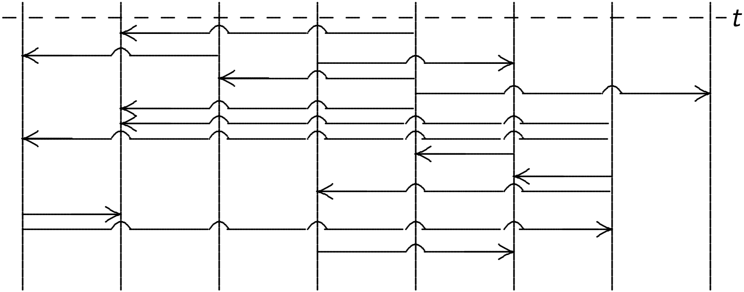
Modelling genomic diversity

- **Goal**: describe diversity of genes in a bacterial population
- **New genes gained** from the environment at rate $\frac{\theta}{2}$
- **Present genes lost** at rate $\frac{\rho}{2}$ but a set of core genes cannot be lost
- A gene present in frequency x but absent in individual j is **horizontally transferred** at rate $\frac{\gamma}{2}x$.

- Frequency path of gene u approximately given by

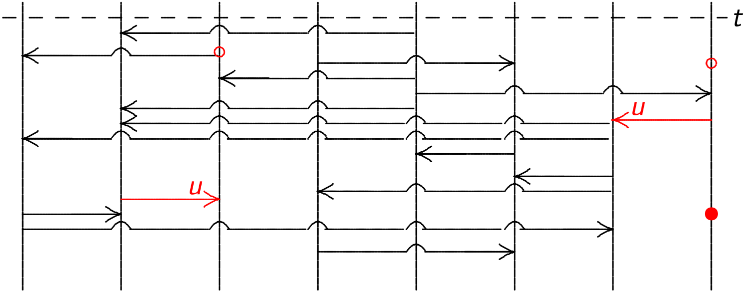
$$dX = \big(-\tfrac{\rho}{2}X + \tfrac{\gamma}{2}X(1-X)\big)dt + \sqrt{X(1-X)}dW \qquad (*)$$

The Moran model with gene gain, loss and HGT



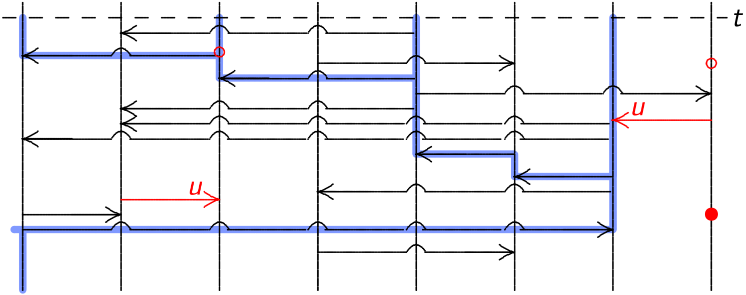
The Moran model with gene gain, loss and HGT

- ●: gene gain, ○: gene loss, $\xrightarrow{u}$: horizontal gene transfer of gene u



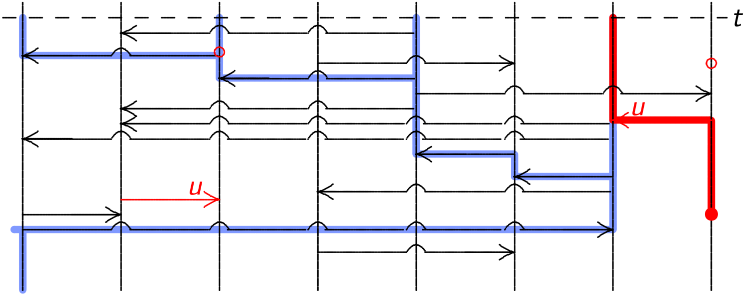
The Moran model with gene gain, loss and HGT

- ●: gene gain, ○: gene loss, $\xrightarrow{u}$: horizontal gene transfer of gene *u*
- **Clonal genealogy**

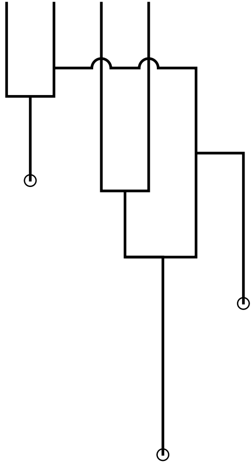


The Moran model with gene gain, loss and HGT

- ●: gene gain, ○: gene loss, $\xrightarrow{\mathbf{u}}$: horizontal gene transfer of gene u
- **Clonal genealogy**
- **Genealogy of gene u in the sample**

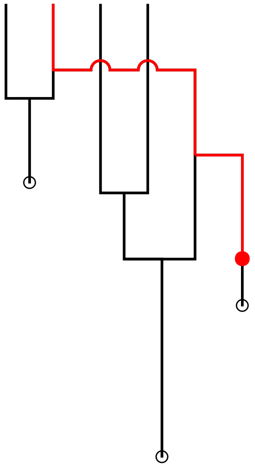


The ancestral gene transfer graph for a single gene



- Start with sample of size $n = 4$
- Pair **coalescence** rate 1
- Every line **killed** at rate $\rho/2$
- Every line produces **incoming line** at rate $\gamma/2$

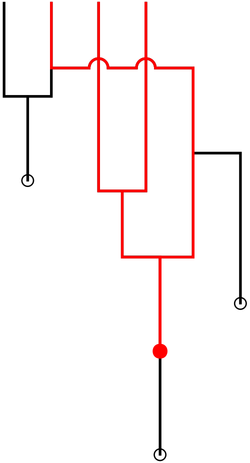
The ancestral gene transfer graph for a single gene



- Start with sample of size $n = 4$
- Pair **coalescence** rate 1
- Every line **killed** at rate $\rho/2$
- Every line produces **incoming line** at rate $\gamma/2$

- **Red leaves** carry the gene

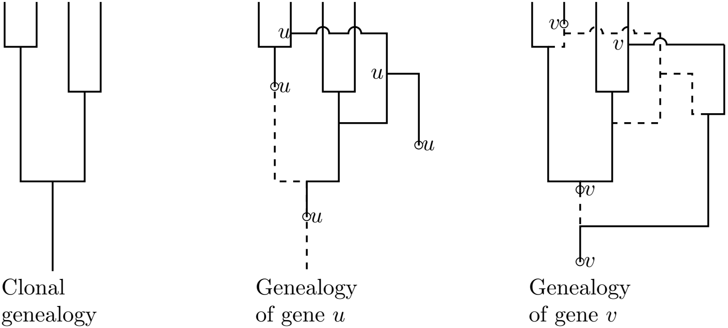
The ancestral gene transfer graph for a single gene



- Start with sample of size $n = 4$
- Pair **coalescence** rate 1
- Every line **killed** at rate $\rho/2$
- Every line produces **incoming line** at rate $\gamma/2$
- **Red leaves** carry the gene

The ancestral gene transfer graph for many genes

- When constructing the ancestral gene transfer graph for the second gene, all lines in the genealogy of the first gene have to be taken into account.



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 *n*th individual?
- What does the **gene frequency spectrum** look like?

Questions

Let $\mathcal{G}_k$ be the set of genes carried by individual k

- How many **genes** does a single individual carry?

  What is $|\mathcal{G}_k|$?
- How many **different genes** are there in the sample?

  What is $|\mathcal{G}|$ for $\mathcal{G} = \bigcup_{k=1}^{n} \mathcal{G}_k$?
- How many **new genes** are there in the nth individual?

  What is $\left|\mathcal{G}_n \setminus \left(\bigcup_{k=1}^{n-1} \mathcal{G}_k\right)\right|$
- What does the **gene frequency spectrum** look like?

  What is $G_k := |\{u \in \mathcal{G} : u \in \mathcal{G}_i \text{ for exactly } k \text{ different } \mathcal{G}_i\}|$?

Results

Based on the SDE (∗), one can show:

- Theorem: With the Pochhammer symbol $(k)_{\bar{m}}$,

$$\mathbb{E}[G_k] = \frac{\theta}{k}\frac{(n)_{\underline{k}}}{(n-1+\rho)_{\underline{k}}}\Big(1+\sum_{m=1}^{\infty}\frac{(k)_{\bar{m}}\gamma^m}{(n+\rho)_{\bar{m}}m!}\Big)$$

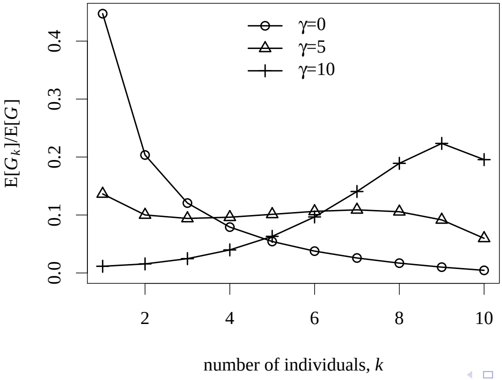
- Corollaries:

$$\mathbb{E}[|\mathcal{G}_k|] = \frac{\theta}{\rho}\left(1+\sum_{m=1}^{\infty}\frac{\gamma^m}{(1+\rho)_{\bar{m}}}\right),$$

$$\mathbb{E}[|\mathcal{G}|] = \theta\sum_{k=0}^{n-1}\frac{1}{k+\rho} + \theta\sum_{m=1}^{\infty}\frac{\gamma^m}{m}\Big(\frac{1}{(\rho)_{\bar{m}}} - \frac{1}{(n+\rho)_{\bar{m}}}\Big)$$

Results

- The gene frequency spectrum is **highly sensitive** to γ
- $n = 10$, $\rho = 2$



Results

Based on the Ancestral Gene Transfer Graph, one can show:

- Theorem: In the limit $\gamma \to 0$,

$$\mathbb{V}[|\mathcal{G}_k|] = \frac{\theta}{\rho}\Big(1 + \frac{1}{1+\rho}\gamma + \Big(\frac{1}{(1+\rho)(2+\rho)} + \frac{\theta}{(1+\rho)^2(3+2\rho)(2+7\rho+6\rho^2)}\Big)\gamma^2\Big) + \mathcal{O}(\gamma^3).$$

- Proof: Take into account at most two gene transfer events in the gene transfer graph of at most two genes.

Remarks and Outlook

Remarks

- Baumdicker, Hess, Pfaffelhuber (2010); Haegeman, Weitz (2012); Lobkovsky, Wolf, Koonin (2013) introduced similar models without horizontal gene transfer

→ **Useful models for data analysis**

- Model with horizontal gene transfer graph not used on data yet

Outlook

- How can we **detect horizontal gene transfer** from data? Which statistics are sensitive to gene transfer events?
- How do **selection**, **population structure** change predicted patterns?