The evolution of bacterial genomes under horizontal gene transfer

Franz Baumdicker – Peter Pfaffelhuber

Albert-Ludwigs Universität Freiburg

April 01, 2013

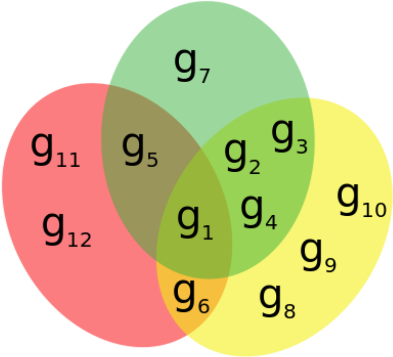
1. introduction
2. The Infinitely Many Genes Model
3. ancestral gene transfer graph
4. main results
5. outlook

Introduction

The distributed genome hypothesis

The set of genes in a population of bacteria is distributed over all individuals that belong to the specific taxon.

- individuals of the same population do not have the same set of genes
- no organism contains the full complement of genes of the species



previous work

Extrapolation:

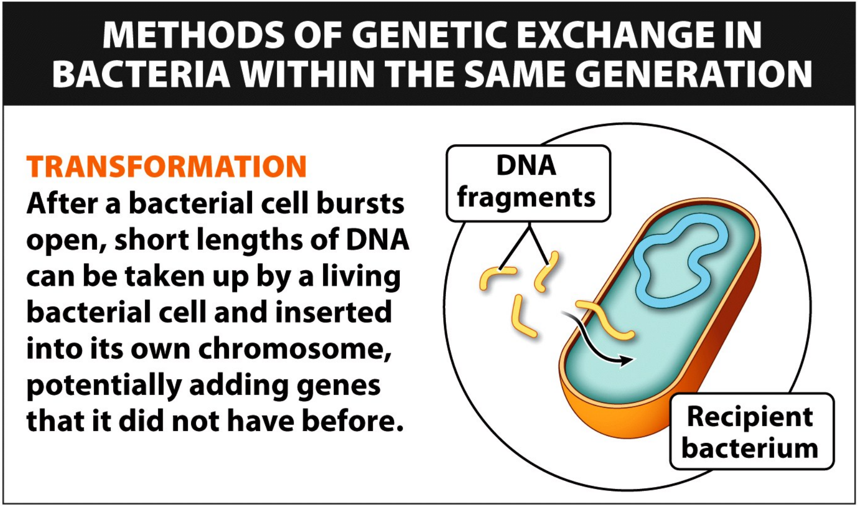
- **coregenome:** a function fitted to the number of genes common to n individuals converges to some number c for $n \to \infty$
- **pangenome:** if a function fitted to the total number of genes in n individuals
  - goes to infinity:
    open pangenome
  - saturates at some finite level:
    closed pangenome

Kittichotirat et al. 2011, Kettler et al. 2007

Modelling genomic diversity

- **Goal:**
  Describe the diversity of distributed genomes in bacterial populations
- base the model on the underlying biological mechanisms
  - random reproduction - genealogy
  - gain of genes
  - loss of genes
  - horizontal gene transfer within the population

Horizontal Gene Transfer in bacteria



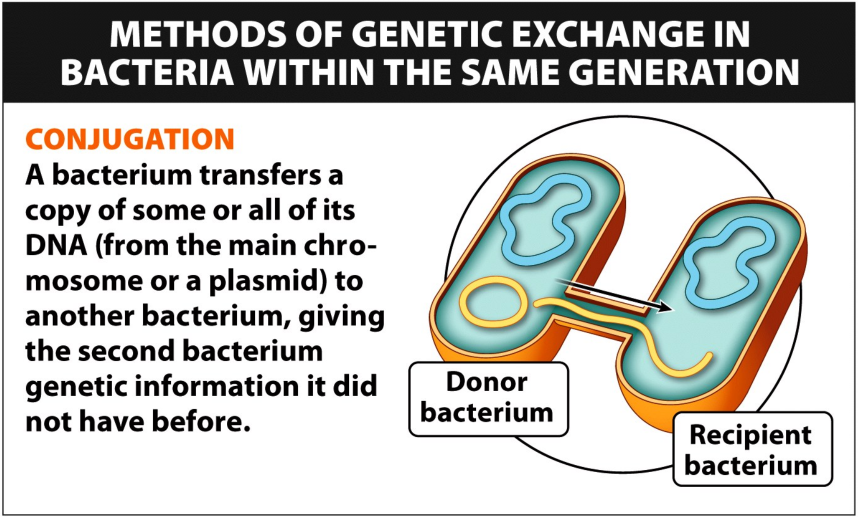
METHODS OF GENETIC EXCHANGE IN BACTERIA WITHIN THE SAME GENERATION

TRANSFORMATION

After a bacterial cell bursts open, short lengths of DNA can be taken up by a living bacterial cell and inserted into its own chromosome, potentially adding genes that it did not have before.

Figure 13-8 part 3
What Is Life? A Guide To Biology
© 2010 W. H. Freeman and Company

Horizontal Gene Transfer in bacteria



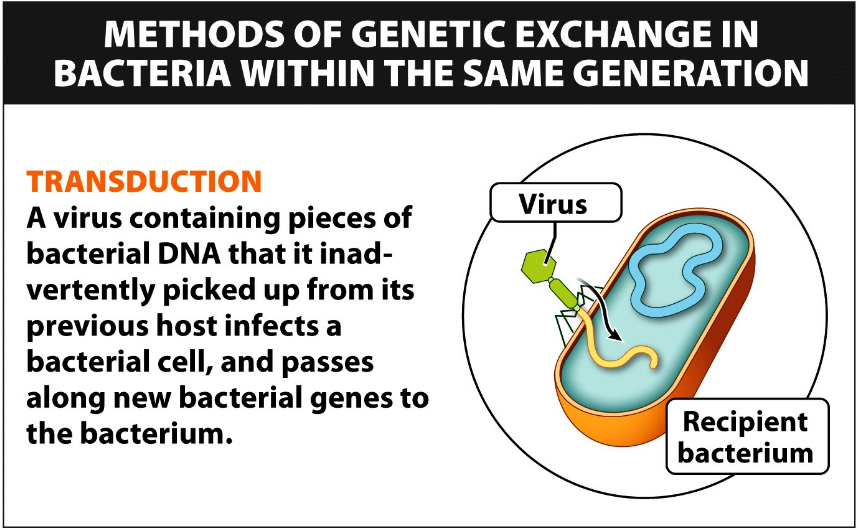
METHODS OF GENETIC EXCHANGE IN BACTERIA WITHIN THE SAME GENERATION

CONJUGATION

A bacterium transfers a copy of some or all of its DNA (from the main chromosome or a plasmid) to another bacterium, giving the second bacterium genetic information it did not have before.

Figure 13-8 part 1
What Is Life? A Guide To Biology
© 2010 W. H. Freeman and Company

Horizontal Gene Transfer in bacteria



METHODS OF GENETIC EXCHANGE IN BACTERIA WITHIN THE SAME GENERATION

TRANSDUCTION

A virus containing pieces of bacterial DNA that it inadvertently picked up from its previous host infects a bacterial cell, and passes along new bacterial genes to the bacterium.

Figure 13-8 part 2
What Is Life? A Guide To Biology
© 2010 W. H. Freeman and Company

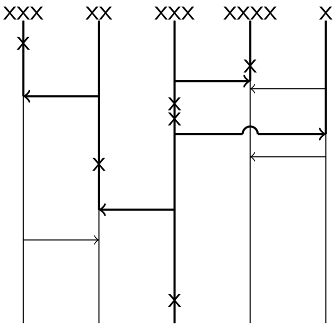
genes and genomes

- the available gene pool is a set of genes $I = [0, 1]$.
- the genome of individual i contains genes $\mathcal{G}_i \subseteq I$
- $\mathcal{G}_i$ is called *dispensable genome* of individual i.

- in addition every individual has a set of c genes absolutely necessary to survive, the *core genome*.
- these genes must be passed from ancestor to offspring.

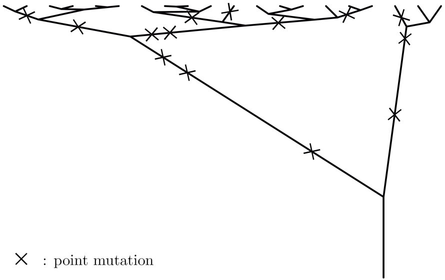
infinitely many sites model

- pairs resample at rate 1
- mutations accumulate at rate θ along the lineages



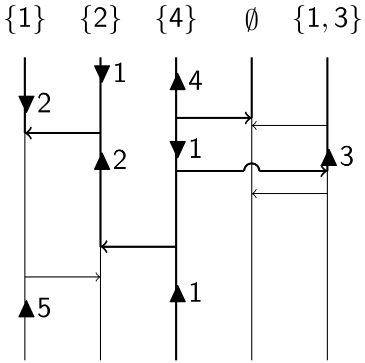
infinitely many sites model

- genealogy is given by Kingman's coalescent
- pairs of lineages coalesce at rate 1
- mutations accumulate at rate θ along the lineages of the Kingman tree



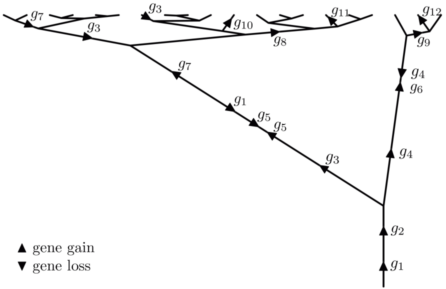
infinitely many genes model

- pairs resample at rate 1
- gene gains occur at rate $\frac{\theta}{2}$ along the lineages
- each gene is lost at rate $\frac{\rho}{2}$



infinitely many genes model

- genealogy is given by Kingman's coalescent
- pairs of lineages coalesce at rate 1
- genes are gained at rate $\frac{\theta}{2}$
- each gene is lost at rate $\frac{\rho}{2}$

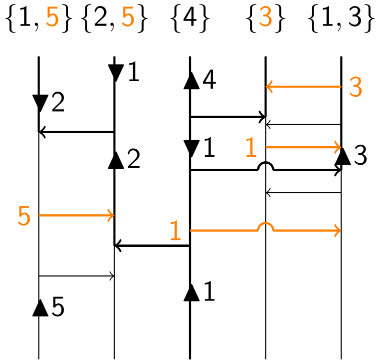


mutation dynamics borrowed from

Phylogenetic Trees based on gene content *Daniel H. Huson,Mike Steel*

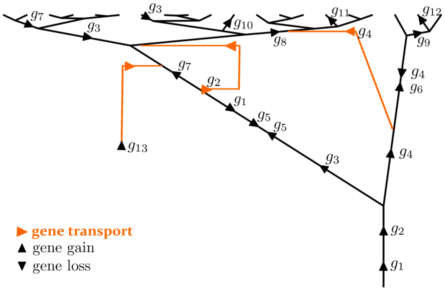
infinitely many genes model with HGT

- pairs resample at rate 1
- gene gains occur at rate $\frac{\theta}{2}$ along the lineages
- each gene is lost at rate $\frac{\rho}{2}$
- a present gene is transfered at rate $\frac{\gamma}{2}$ to a random individual
- a transfered gene is a copy.
- donor and acceptor both carry the gene



infinitely many genes model with HGT

- genealogy is given by Kingman's coalescent
- pairs of lineages coalesce at rate 1
- genes are gained at rate $\frac{\theta}{2}$
- each gene is lost at rate $\frac{\rho}{2}$
- each gene is transfered at rate $\frac{\gamma}{2}$ from a unknown line
- a transfered gene is a copy.
- donor and acceptor both carry the gene

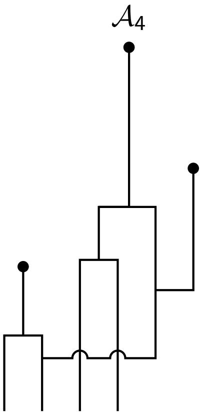


infinitely many genes model with HGT

- genealogy is given by Kingman's coalescent
- pairs of lineages coalesce at rate 1
- genes are gained at rate $\frac{\theta}{2}$
- each gene is lost at rate $\frac{\rho}{2}$
- each gene is transfered at rate $\frac{\gamma}{2}$ from a unknown line
- a transfered gene is a copy.
- donor and acceptor both carry the gene

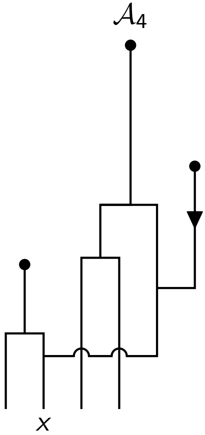
The ancestral gene transfer graph for a single gene

- each pair of lines coalesces at rate 1,
- each line disappears at rate $\rho/2$
  - the gene was lost
- each line splits in two lines at rate $\gamma/2$
  - the gene was horizontally transferred from another individual
  - the gene can now have two different origins



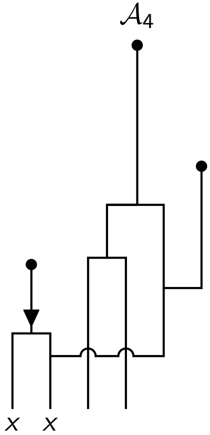
The ancestral gene transfer graph for a single gene

- each pair of lines coalesces at rate 1,
- each line disappears at rate $\rho/2$
  - the gene was lost
- each line splits in two lines at rate $\gamma/2$
  - the gene was horizontally transferred from another individual
  - the gene can now have two different origins



The ancestral gene transfer graph for a single gene

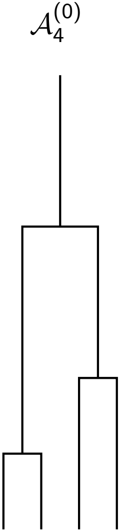
- each pair of lines coalesces at rate 1,
- each line disappears at rate $\rho/2$
  - the gene was lost
- each line splits in two lines at rate $\gamma/2$
  - the gene was horizontally transferred from another individual
  - the gene can now have two different origins



The ancestral gene transfer graph for infinitely many genes

- start with the clonal genealogy of the sample $\mathcal{A}_n^{(0)}$.

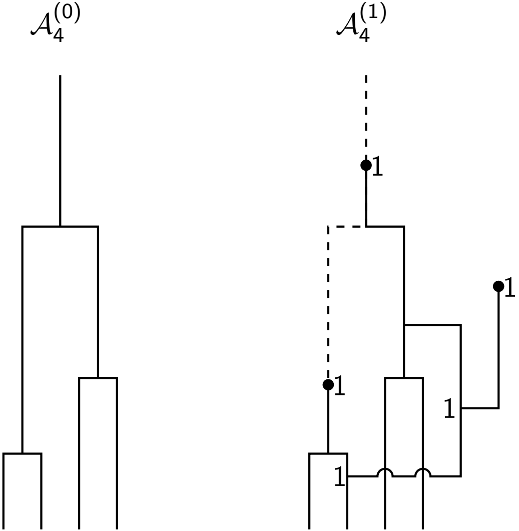
The ancestral gene transfer graph for infinitely many genes



The ancestral gene transfer graph for infinitely many genes

- start with the clonal genealogy of the sample $\mathcal{A}_n^{(0)}$.
- construct the genealogy of the first gene $\mathcal{A}_n^{(1)}$
  - loss events at rate $\rho/2$
  - additional splitting events at rate $\gamma/2$
  - add coalescence events for each new line

The ancestral gene transfer graph for infinitely many genes

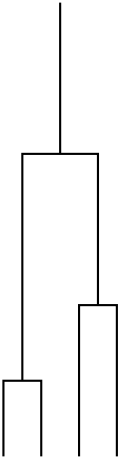


The ancestral gene transfer graph for infinitely many genes

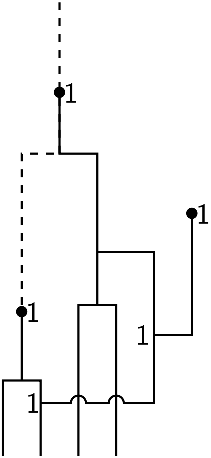
- start with the clonal genealogy of the sample $\mathcal{A}_n^{(0)}$.
- construct the genealogy of the first gene $\mathcal{A}_n^{(1)}$
  - loss events at rate $\rho/2$
  - additional splitting events at rate $\gamma/2$
  - add coalescence events for each new line
- Iteratively, construct $\mathcal{A}_n^{(k+1)}$
  - keep all lines in $\cup_{i=0}^{k}\mathcal{A}_n^{(i)}$
  - add splitting, loss and coalescence events.

The ancestral gene transfer graph for infinitely many genes

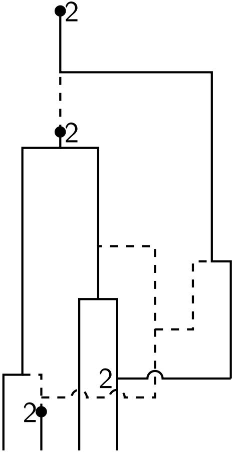
$\mathcal{A}_4^{(0)}$



$\mathcal{A}_4^{(1)}$



$\mathcal{A}_4^{(2)}$



gene gains in the AGTG

Consider the events $(T_m, U_m)_{m=1,2,\dots}$ of a Poisson point process on $[0,\infty) \times [0,1]$ with intensity measure $\frac{1}{2}\theta\, dt\, du$.

If $T_k \leq L(\mathcal{A}_n^{(k)})$, pick a point uniformly at random on $\mathcal{A}_n^{(k)}$, where the gene U_k was gained.

weak convergence

Gene distributions from Moran model and AGTG coincide

Let $(\mathcal{G}_1^N, ..., \mathcal{G}_n^{(N)})$ be the genes of individual $1, \dots, n$ in the previously described moran model of size N.
And let $(\mathcal{G}_1, ..., \mathcal{G}_n)$ be the gene distribution read off from the AGTG.
Then,

$$(\mathcal{G}_1^N, ..., \mathcal{G}_n^{(N)}) \xrightarrow{N \to \infty} (\mathcal{G}_1, ..., \mathcal{G}_n)$$

weak convergence

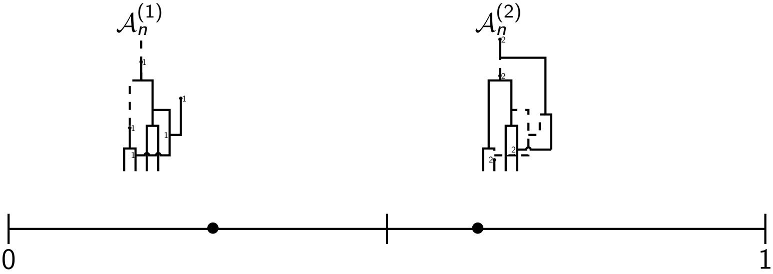
Gene distributions from Moran model and AGTG coincide

Let $(\mathcal{G}_1^N, ..., \mathcal{G}_n^{(N)})$ be the genes of individual $1, \ldots, n$ in the previously described moran model of size N.

And let $(\mathcal{G}_1, ..., \mathcal{G}_n)$ be the gene distribution read off from the AGTG.

Then,

$$(\mathcal{G}_1^N, ..., \mathcal{G}_n^{(N)}) \xrightarrow{N\to\infty} (\mathcal{G}_1, ..., \mathcal{G}_n)$$



weak convergence

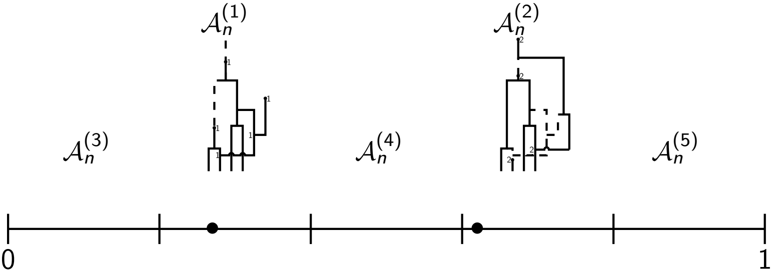
Gene distributions from Moran model and AGTG coincide

Let $(\mathcal{G}_1^N, ..., \mathcal{G}_n^{(N)})$ be the genes of individual $1, \dots, n$ in the previously described moran model of size N.

And let $(\mathcal{G}_1, ..., \mathcal{G}_n)$ be the gene distribution read off from the AGTG.

Then,

$$(\mathcal{G}_1^N, ..., \mathcal{G}_n^{(N)}) \xrightarrow{N\to\infty} (\mathcal{G}_1, ..., \mathcal{G}_n)$$



single individual – average number of genes

- $|\mathcal{G}_i|$: number of genes in individual i

$$\mathbb{E}[|\mathcal{G}_i|] = \frac{\theta}{\rho} + \frac{\theta}{\rho}\sum_{m=1}^{\infty}\frac{\gamma^m}{(1+\rho)_m} + c$$

with $(a)_b := a(a+1)\cdots(a+b-1)$.

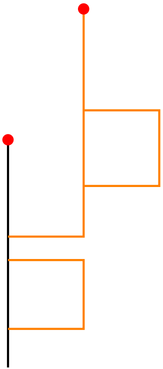
single individual – average number of genes

- without HGT ($\gamma = 0$)
- following one line backwards in time
  - losses occur at rate $\frac{\rho}{2}$

- expected length of unlost line is $\frac{2}{\rho}$
- $\mathbb{E}[|\mathcal{G}_i|] = \frac{\theta}{2}\frac{2}{\rho} = \frac{\theta}{\rho}$

single individual – average number of genes

- with HGT ($\gamma > 0$)
- following one line backwards in time
  - each line dies at rate $\frac{\rho}{2}$
  - each line produces a new line at rate $\frac{\gamma}{2}$
  - each pair of lines coalesces at rate 1
- L_m: length of the ancestral gene transfer graph started with m lines
- $\mathbb{E}[|\mathcal{G}_i|] = \frac{\theta}{2}\mathbb{E}[L_1]$



average number of genes – birth-death processes

The length L_m of an AGTG started with m lines
equals
the time to absorption for a birth-death process started in m with
birth-rate $\lambda_i = \frac{1}{i}\frac{i\gamma}{2} = \frac{\gamma}{2}$
death-rate $\mu_i = \frac{1}{i}\left(\frac{i\rho}{2} + \frac{i(i-1)}{2}\right) = \frac{\rho+i-1}{2}$

Thus,

$$\mathbb{E}[|\mathcal{G}_i|] = \frac{\theta}{2}\mathbb{E}[L_1] = \frac{\theta}{2}\sum_{i=1}^{\infty} p_i = \frac{\theta}{2}\sum_{i=1}^{\infty}\frac{\lambda_1\lambda_2\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_i}$$
$$= \frac{\theta}{\rho}\left(1 + \sum_{i=1}^{\infty}\frac{\gamma^i}{(\rho+1)_i}\right).$$

expected pangenome size – birth-death processes

Use same idea to compute the expected number of genes in n individuals (pangenome size)

$$\mathbb{E}\left[\left|\bigcup_{i=1}^{n}\mathcal{G}_i\right|\right] = \frac{\theta}{2}\mathbb{E}[L_n] = \frac{\theta}{2}\left(\sum_{i=1}^{\infty} p_i + \sum_{r=1}^{n-1}\left(\prod_{k=1}^{r}\frac{\mu_k}{\lambda_k}\right)\sum_{j=r+1}^{\infty} p_j\right)$$

$$= \theta\sum_{k=0}^{n-1}\frac{1}{k+\rho}\left(1+\sum_{m=1}^{\infty}\frac{\gamma^m}{(i+\rho)_m}\right)$$

the gene frequency spectrum

- The *gene frequency spectrum* is given by $G_1^{(n)}, \ldots, G_n^{(n)}$, where

$$G_k^{(n)} := |\{u \in I : u \in \mathcal{G}_i \text{ for exactly } k \text{ different } i\}|.$$

$$\mathbb{E}[G_k^{(n)}] = \frac{\theta}{k} \frac{n \cdots (n-k+1)}{(n-1+\rho)\cdots(n-k+\rho)} \left(1 + \sum_{m=1}^{\infty} \frac{(k)_m \gamma^m}{(n+\rho)_m m!}\right)$$

with $(a)_b := a(a+1)\cdots(a+b-1)$.

diffusion theory and the gene frequency spectrum

Let (X_t) be the frequency of a gene at time t.
Then, $(X_t)_{t\geq 0}$ is a diffusion process, which follows the SDE

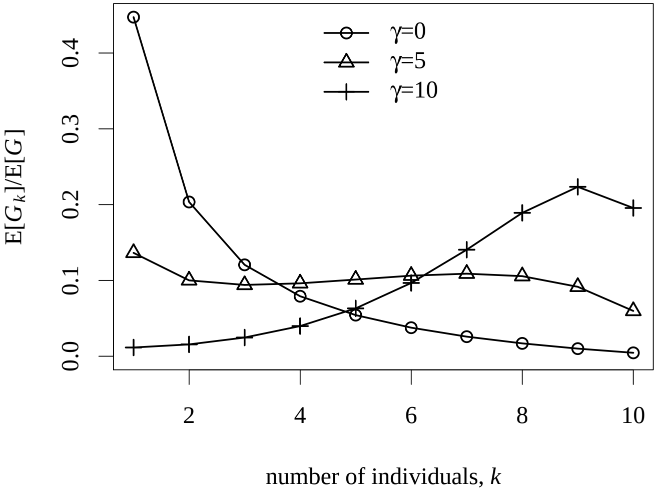
$$dX = -\frac{\rho}{2}Xdt + \frac{\gamma}{2}X(1-X)dt + \sqrt{X(1-X)}dW.$$

The number of genes in frequency x is Poisson with mean

$$g(x)dx := \theta\frac{e^{\gamma x}}{x(1-x)^{1-\rho}}dx.$$

and

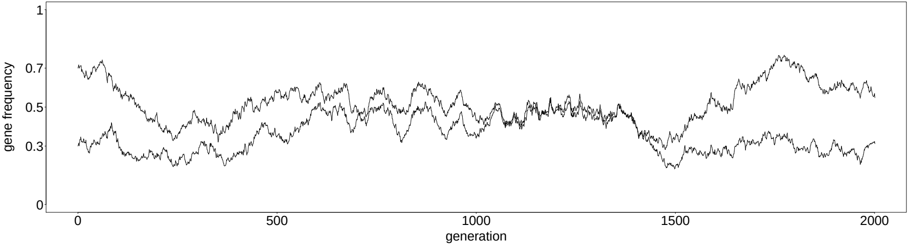
$$\begin{aligned}\mathbb{E}[G_k^{(n)}] &= \binom{n}{k}\int_0^1 g(x)x^k(1-x)^{n-k}dx \\ &= \frac{\theta}{k}\frac{n\cdots(n-k+1)}{(n-1+\rho)\cdots(n-k+\rho)}\left(1+\sum_{m=1}^{\infty}\frac{(k)_m\gamma^m}{(n+\rho)_m m!}\right)\end{aligned}$$



The expected gene frequency spectrum is highly dependent of γ. For high values of γ, most genes are in high frequency, leading to a closed pangenome.

higher moments

- The frequencies of two genes depend on each other.
- can not apply 1-dim diffusion methods to get higher moments



variance – approximations in the AGTG

$$\mathbb{V}ar[|\mathcal{G}_i|] = \frac{\theta}{\rho}\left(1 + \frac{\gamma}{1+\rho}\right) + \mathcal{O}(\gamma^2)$$

$$\mathbb{V}[|\mathcal{G}_1|] = \int_0^1 \mathbb{V}[\mathcal{G}_1(dx)] + \int_0^1\int_0^1 1_{x\neq y}\mathbb{COV}[\mathcal{G}_1(dx), \mathcal{G}_1(dy)]$$

$$\mathbb{V}[|\mathcal{G}_1(dx)|] = \frac{\theta}{2}\mathbb{E}[L(\mathcal{A}^1)]dx + \mathcal{O}(dx^2) = \frac{\theta}{2}\mathbb{E}[L_1]dx + \mathcal{O}(dx^2)$$

$$\mathbb{COV}[|\mathcal{G}_1(dx)|, |\mathcal{G}_1(dy)|] = \frac{\theta^2}{4}\mathbb{COV}[L(\mathcal{A}^1), L(\mathcal{A}^2)]dx\,dy$$

variance – approximations in the AGTG

$$\mathbb{V}ar[|\mathcal{G}_i|] = \frac{\theta}{\rho}\left(1 + \frac{\gamma}{1+\rho} + \frac{\gamma^2}{(1+\rho)(2+\rho)} + \frac{\gamma^2\theta}{(1+\rho)^2(3+2\rho)(2+7\rho+6\rho^2)}\right) + \mathcal{O}(\gamma^3)$$

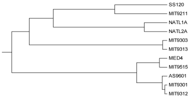
$$\mathbb{V}[|\mathcal{G}_1|] = \int_0^1 \mathbb{V}[\mathcal{G}_1(dx)] + \int_0^1 \int_0^1 1_{x \neq y}\mathbb{COV}[\mathcal{G}_1(dx), \mathcal{G}_1(dy)]$$

$$\mathbb{V}[|\mathcal{G}_1(dx)|] = \frac{\theta}{2}\mathbb{E}[L(\mathcal{A}^1)]dx + \mathcal{O}(dx^2) = \frac{\theta}{2}\mathbb{E}[L_1]dx + \mathcal{O}(dx^2)$$

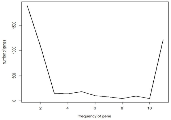
$$\mathbb{COV}[|\mathcal{G}_1(dx)|, |\mathcal{G}_1(dy)|] = \frac{\theta^2}{4}\mathbb{COV}[L(\mathcal{A}^1), L(\mathcal{A}^2)]dx\,dy$$

software IMaGe

estimated genealogical tree



gene frequency spectrum



IMaGe

statistical test for hypotheses of neutral evolution

parameter estimates

estimated pangenome size

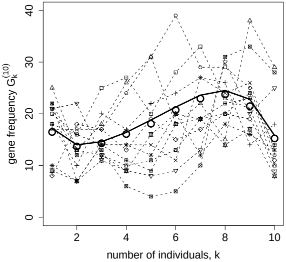
expected no. of new genes in the next individual

. . .

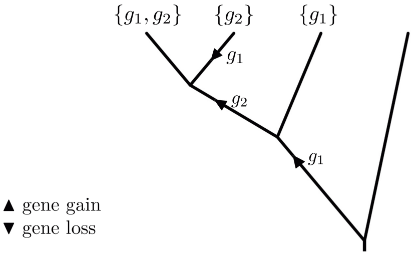
outlook – estimating γ

given the observed gene frequency spectrum it is difficult to estimate θ, ρ, γ and c solely based on the mean gene frequency spectrum

- for $\gamma = 0$ IMaGe uses an a priori tree
- for $\gamma > 0$ each gene has its own genealogy
- need a new statistic besides the gfs which is sensible to γ



pairs of incongruent genes



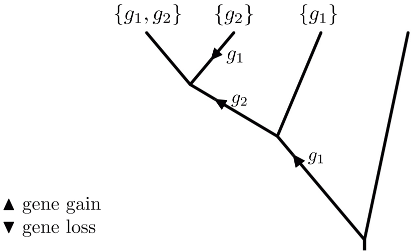
The *number of incongruent pairs of genes* is given by

$$P := \frac{1}{n(n-1)(n-2)(n-3)} \sum_{i,j,k,l=1}^{n} A_{ij,kl} \cdot A_{ik,jl}$$

where

$$A_{ij,kl} := |(\mathcal{G}_i \cap \mathcal{G}_j) \setminus (\mathcal{G}_k \cup \mathcal{G}_l)|, \qquad 1 \leq i, j, k, l \leq n.$$

pairs of incongruent genes



The *average number of incongruent pairs of genes* without HGT is given by

$$\mathbb{E}[P] = \frac{\theta^2\rho}{4} \frac{18 + 117\frac{\rho}{2} + 203\frac{\rho^2}{4} + 105\frac{\rho^3}{8}}{(1+\frac{\rho}{2})^2(1+2\frac{\rho}{2})(1+4\frac{\rho}{2})(3+4\frac{\rho}{2})(3+5\frac{\rho}{2})(6+5\frac{\rho}{2})(6+7\frac{\rho}{2})}$$

$$\mathbb{E}[A_{ij,kl}] = \frac{1}{\binom{4}{2}}\mathbb{E}[G_2^{(4)}] = \frac{\theta}{(3+\rho)(2+\rho)}$$

outlook

- test for HGT ($\gamma > 0$) in the infinitely many genes model, based on the number of incongruent pairs.
- joint distribution of gene frequency and mutations in the corresponding gene sequence
- other possible extensions of the IMG model:
  - selection, structured populations, changing population size
- apply the model to other bacteria:
  - E. Coli, green sulfer bacteria, epidemic strains, gut bacteria, soil bacteria
- ...

Thank you for your attention!

The infinitely many genes model
Baumdicker, F., W. R. Hess, and P. Pfaffelhuber (2010).
The diversity of a distributed genome in bacterial populations.
Ann. Appl. Probab. 20 (5).

model applied to cyanobacterial pangenome, estimates, IMaGe
Baumdicker, F., W. R. Hess, and P. Pfaffelhuber (2012).
The infinitely many genes model for the distributed genome of bacteria.
Genome Biol Evol Vol. 4, 443-456.

ancestral gene transfer graph
Baumdicker, F. and P. Pfaffelhuber
The infinitely many genes model with horizontal gene transfer
arXiv:1301.6547 [math.PR], in review