Basic Coalescent

Qian Feng

University of Melbourne

fengq2@student.unimelb.edu.au

April 4, 2018

Overview

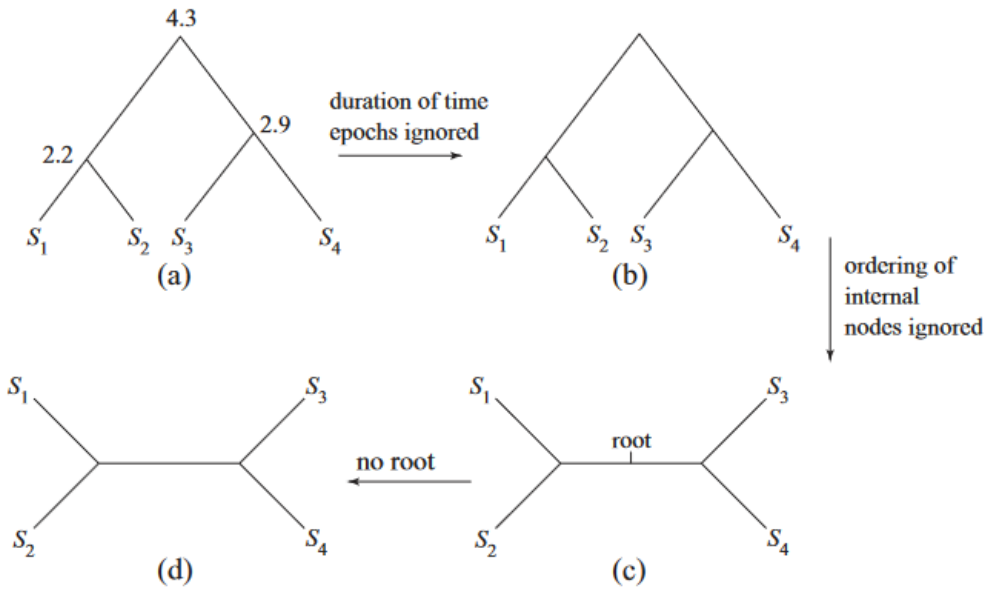
1. Coalescent Tree
   - The Coalescent
   - Height of the tree
   - Total branch length of the tree
2. Gene Tree
   - Mathematical models of alleles
   - Two gene tree building approaches
3. Models
   - Wright-Fisher models
   - Wright-Fisher model with mutation
   - Wright-Fisher model with recombination(ARG)

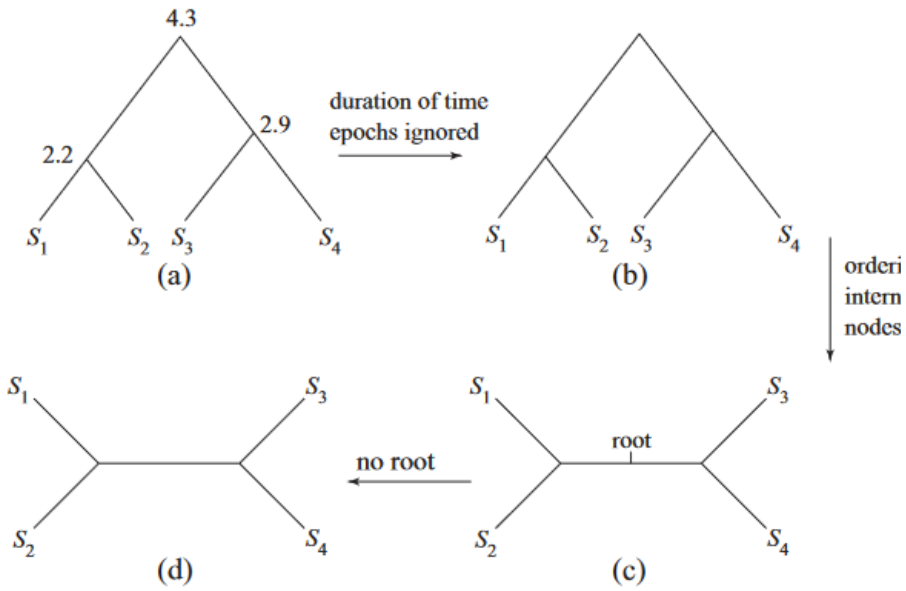
The Coalescent

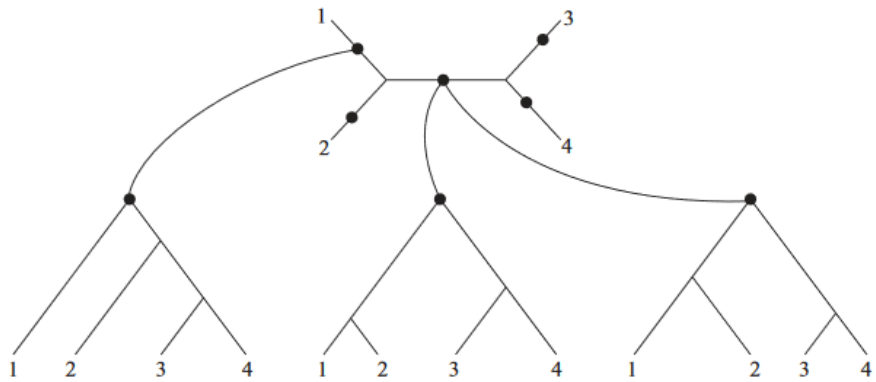
- is a model that describe the relationships within a sample from the present individuals back to the MRCA(pavlos pavlidis,2013).
- Using this coalescent model allows repeatedly to draw samples from a population, and then to estimate the population genetic parameters.
- The coalescent is NOT a tree reconstruction method. It is a sampling method.

Coalescent differs from Phylogenetic tree concept

- In phylogenetics, the goal is to obtain the tree that best describes the data.Related question is what the evolutionary relationships between the given set of sequences are.
- In population genetics,the genealogy is used exclusively to obtain statistical samples, and then infer the properties of these samples over all possible genealogies.


4.3
2.2
2.9
S_1 S_2 S_3 S_4
(a)
duration of time
epochs ignored
S_1 S_2 S_3 S_4
(b)
ordering of
internal
nodes ignored
S_1 S_3
S_2 S_4
(d)
no root
S_1 S_3
root
S_2 S_4
(c)


4.3
2.2
2.9
S_1 S_2 S_3 S_4
(a)
duration of time
epochs ignored
S_1 S_2 S_3 S_4
(b)
orderi
intern
nodes
S_1 S_3
S_2 S_4
(d)
no root
S_1 S_3
root
S_2 S_4
(c)


1 2 3 4
1 2 3 4
1 2 3 4
1 3
2 4

Count tree topologies

- C_k and B_k denote the number of coalescent and binary unrooted tree topologies with k leaves, respectively.
- Then the recursion and initial conditionals are:

Count tree topologies

- C_k and B_k denote the number of coalescent and binary unrooted tree topologies with k leaves, respectively.
- Then the recursion and initial conditionals are:

$$C_k = \binom{k}{2} C_{k-1}, C_2 = 1$$

Count tree topologies

- C_k and B_k denote the number of coalescent and binary unrooted tree topologies with k leaves, respectively.
- Then the recursion and initial conditionals are:

$$C_k = \binom{k}{2} C_{k-1}, C_2 = 1$$

and

$$B_k = (2k-5)B_{k-1}, B_3 = B_2 = 1$$

Count tree topologies

- C_k and B_k denote the number of coalescent and binary unrooted tree topologies with k leaves, respectively.
- Then the recursion and initial conditionals are:

$$C_k = \prod_{j=2}^{k} \frac{j(j-1)}{2} = \frac{k!(k-1)!}{2^{k-1}}$$

Count tree topologies

- C_k and B_k denote the number of coalescent and binary unrooted tree topologies with k leaves, respectively.
- Then the recursion and initial conditionals are:

$$C_k = \prod_{j=2}^{k} \frac{j(j-1)}{2} = \frac{k!(k-1)!}{2^{k-1}}$$

and

$$B_k = (2k-5)!!$$

- Example: $B_5 = 5!! = 1 * 3 * 5 = 15$

Count tree topologies

k	2	3	4	5	6	8	10	15
B_k	1	1	3	15	105	$1.0 * 10^4$	$2.0 * 10^6$	$7.9 * 10^{12}$
C_k	1	3	18	180	2700	$1.6 * 10^6$	$2.6 * 10^9$	$7.0 * 10^{18}$

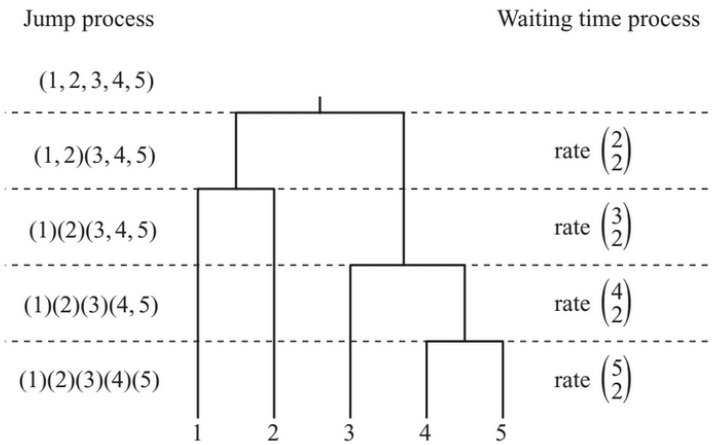
Table: The number of topologies. R_k of rooted topologies is equal to B_{k+1}

Basic coalescent process

We refer to the continuous time coalescent as the basic coalescent process, it's divided into two independent processes: **waiting time process** and **jump process**(describes which genes coalescent).

Basic coalescent process

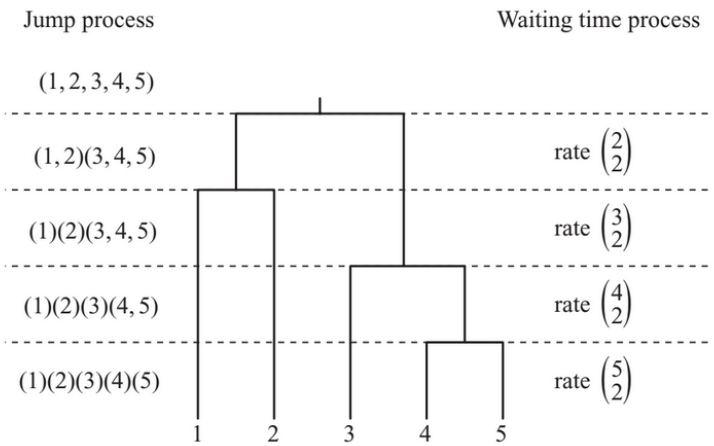
We refer to the continuous time coalescent as the basic coalescent process, it's divided into two independent processes: **waiting time process** and **jump process**(describes which genes coalescent).



Basic coalescent process

We refer to the continuous time coalescent as the basic coalescent process, it's divided into two independent processes: **waiting time process** and **jump process**(describes which genes coalescent).

Algorithm 1



1. Start with $k = n$ genes. Simulating the waiting time T_k to the next event, $T_k \sim Exp(\binom{k}{2})$
2. Choose a random pair (i, j) uniformly among the $\binom{k}{2}$ possible pairs.
3. Merge i and j into one gene and decrease the sample size by one, $k \to k-1$,repeat until $k = 1$

how to build a coalescent tree

Let's look at one animation showing the process of building a coalescent tree[Hudson animator:http://www.coalescent.dk/].

Height of thep tree

The height ,H_n of the tree with n sample is the sum of epochs, T_j,$j = n, n-1...2$. Here we focus on the mean and variance of H_n even though the cdf of H_n is also given by authors.

- The mean is:

$$E(H_n) = \sum_{j=2}^{n} E(T_j) = 2\sum_{j=2}^{n} \frac{1}{j(j-1)} = 2(1 - \frac{1}{n})$$

Height of thep tree

The height ,H_n of the tree with n sample is the sum of epochs, T_j,$j = n, n-1...2$. Here we focus on the mean and variance of H_n even though the cdf of H_n is also given by authors.

- The mean is:

$$E(H_n) = \sum_{j=2}^{n} E(T_j) = 2\sum_{j=2}^{n} \frac{1}{j(j-1)} = 2(1 - \frac{1}{n})$$

and

$$\lim_{n \to +\infty} E(H_n) = 2$$

Height of the tree

The height, H_n of the tree with n samples is the sum of epochs, $T_j, j = n, n-1 \ldots 2$. Here we focus on the mean and variance of H_n even though the cdf of H_n is also given by authors.

- The variance is

$$Var(H_n) = \sum_{j=2}^{n} Var(T_j) = 4 \sum_{j=2}^{n} \frac{1}{j^2(j-1)^2}$$

Height of the tree

The height, H_n of the tree with n samples is the sum of epochs, $T_j, j = n, n-1 \ldots 2$. Here we focus on the mean and variance of H_n even though the cdf of H_n is also given by authors.

- The variance is

$$Var(H_n) = \sum_{j=2}^{n} Var(T_j) = 4 \sum_{j=2}^{n} \frac{1}{j^2(j-1)^2}$$

and

$$\lim_{n \to +\infty} Var(H_n) = \lim_{n \to +\infty} 4 \sum_{j=2}^{n} \frac{1}{j^2(j-1)^2} = \frac{4}{3}(\pi^2 - 9) \approx 1.159$$

Proof of this limit

This proof contains two parts:

- $$\lim_{n \to +\infty} \sum_{j=1}^{n} \frac{1}{j^2} = \frac{\pi^2}{6}$$

Proof of this limit

This proof contains two parts:

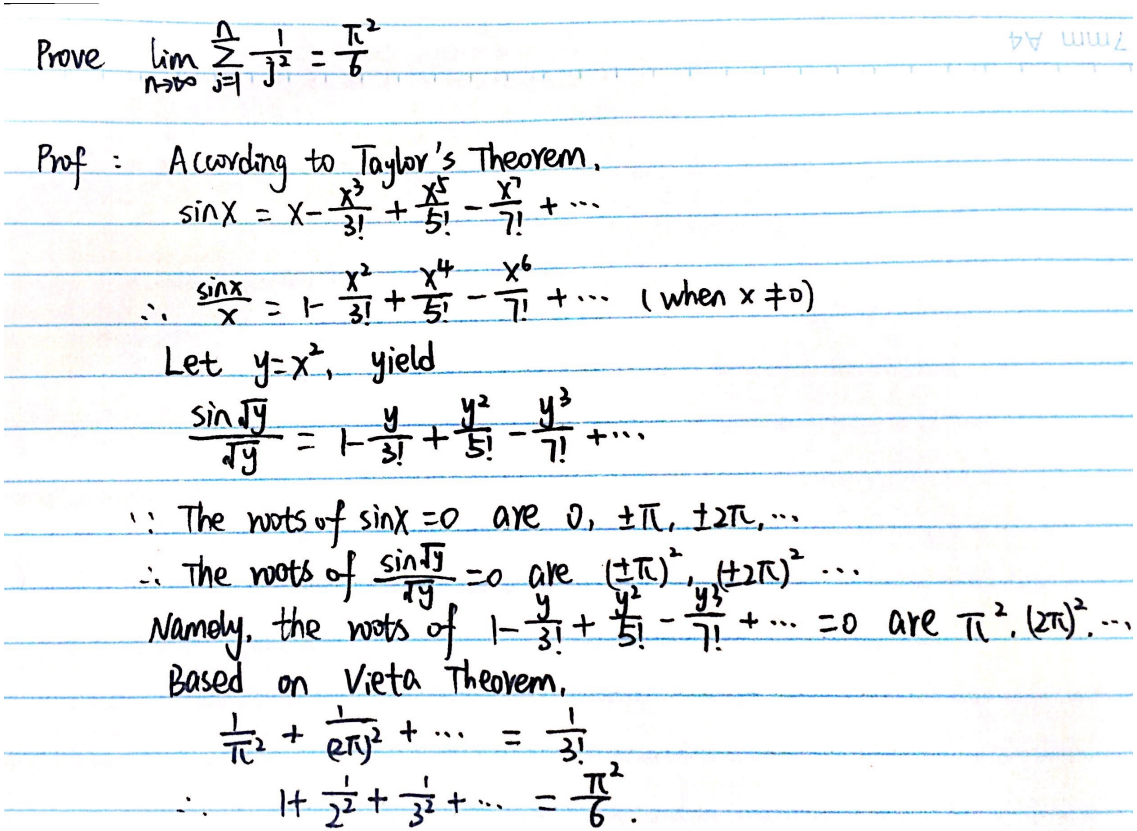
- $$\lim_{n\to+\infty}\sum_{j=1}^{n}\frac{1}{j^2}=\frac{\pi^2}{6}$$

  and

$$\lim_{n\to+\infty}\sum_{j=2}^{n}\frac{1}{j^2(j-1)^2}=\frac{1}{3}(\pi^2-9)$$

Proof of this limit

This proof of first part:



Prove $\lim_{n\to\infty} \sum_{j=1}^{n} \frac{1}{j^2} = \frac{\pi^2}{6}$

Prof : According to Taylor's Theorem,

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$\therefore \frac{\sin x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \cdots \quad (\text{when } x \neq 0)$$

Let $y = x^2$, yield

$$\frac{\sin\sqrt{y}}{\sqrt{y}} = 1 - \frac{y}{3!} + \frac{y^2}{5!} - \frac{y^3}{7!} + \cdots$$

$\because$ The roots of $\sin x = 0$ are $0, \pm\pi, \pm 2\pi, \cdots$

$\therefore$ The roots of $\frac{\sin\sqrt{y}}{\sqrt{y}} = 0$ are $(\pm\pi)^2, (\pm 2\pi)^2 \cdots$

Namely, the roots of $1 - \frac{y}{3!} + \frac{y^2}{5!} - \frac{y^3}{7!} + \cdots = 0$ are $\pi^2, (2\pi)^2, \cdots$

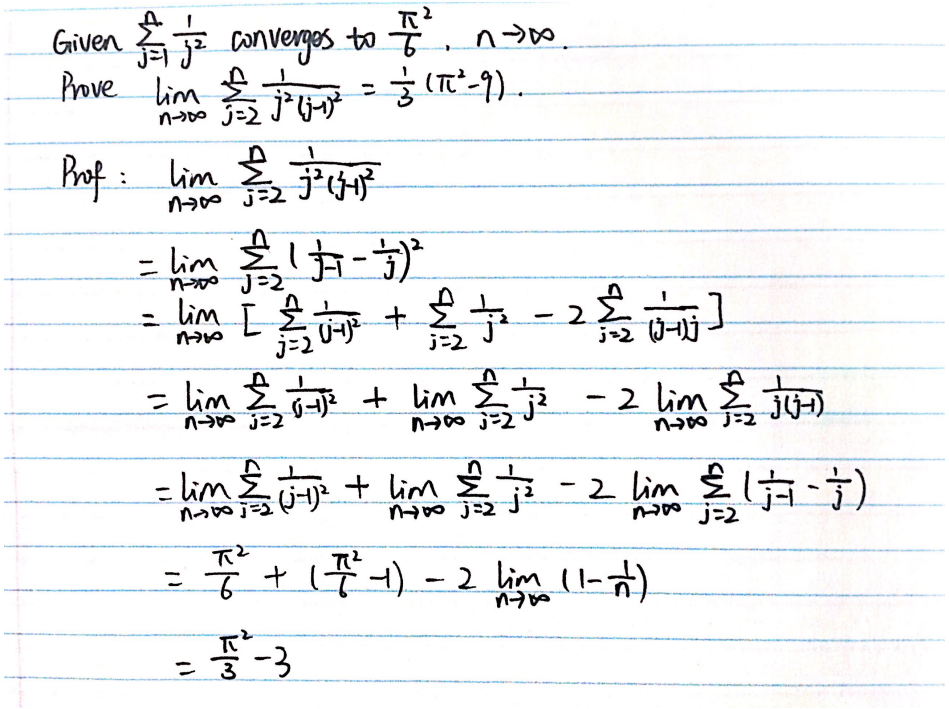
Based on Vieta Theorem,

$$\frac{1}{\pi^2} + \frac{1}{(2\pi)^2} + \cdots = \frac{1}{3!}$$

$$\therefore \quad 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots = \frac{\pi^2}{6}.$$

Proof of this limit

This proof of second part:



Total branch length of the tree L_n

The mean L_n is obtained by weighting the the coalescent times by the number of lineages.

- The mean L_n is

$$E(L_n) = \sum_{j=2}^{n} jE(T_j) = 2\sum_{j=2}^{n-1} \frac{1}{j} \approx log(n)$$

Total branch length of the tree L_n

The mean L_n is obtained by weighting the the coalescent times by the number of lineages.

- The mean L_n is

$$E(L_n) = \sum_{j=2}^{n} jE(T_j) = 2\sum_{j=2}^{n-1} \frac{1}{j} \approx log(n)$$

and

$$Var(L_n) = \sum_{j=2}^{n} j^2 Var(T_j) = 4\sum_{j=1}^{n} \frac{1}{j^2}$$

which converges to $2\pi^2/3 \approx 6.579$

Table

n	2	3	4	5	6	10	15	20
$\mathrm{Var}(H_n)$	1.000	1.111	1.139	1.149	1.153	1.158	1.159	1.159
T_2 contribution	1.000	0.900	0.877	0.870	0.867	0.864	0.863	0.863
$\mathrm{Var}(L_n)$	4.000	5.000	5.444	5.694	5.854	6.159	6.304	6.375
T_2 contribution	1.000	0.800	0.734	0.702	0.683	0.649	0.635	0.627

- For example:

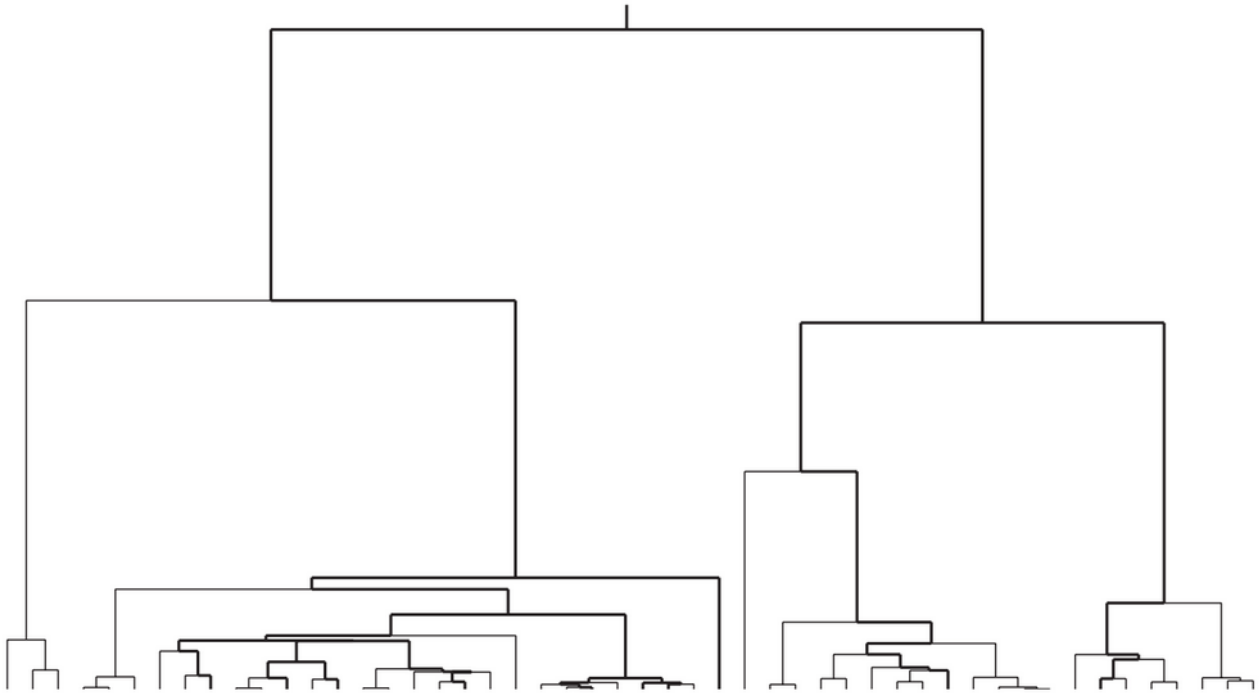
$$0.877 = \frac{1}{1.139}$$

$$0.8 = \frac{2^2 Var(T_2)}{2^2 Var(T_2) + 3^2 var(T_3)} = \frac{4}{4+1} = 0.8$$

- $Var(L_n) \approx 6.579$, T_2 contribution converges to 0.627, as n increase.

The effect of sampling more sequences

10 (**in bold**) versus 40 sequences



Mathematical models of alleles

A configuration of n genes can be also simulated in a two-step procedure: simulate the genealogy of n genes, then add mutations to this genealogy according to chosen mutation model. Three types of mutation models:

1. The infinite alleles model
2. The infinite sites model
3. Finite sites model

Infinite alleles model

Infinite alleles model was proposed by Kimura and Crow (1964)

Theorem (definition)

Each mutation creates a new allele.

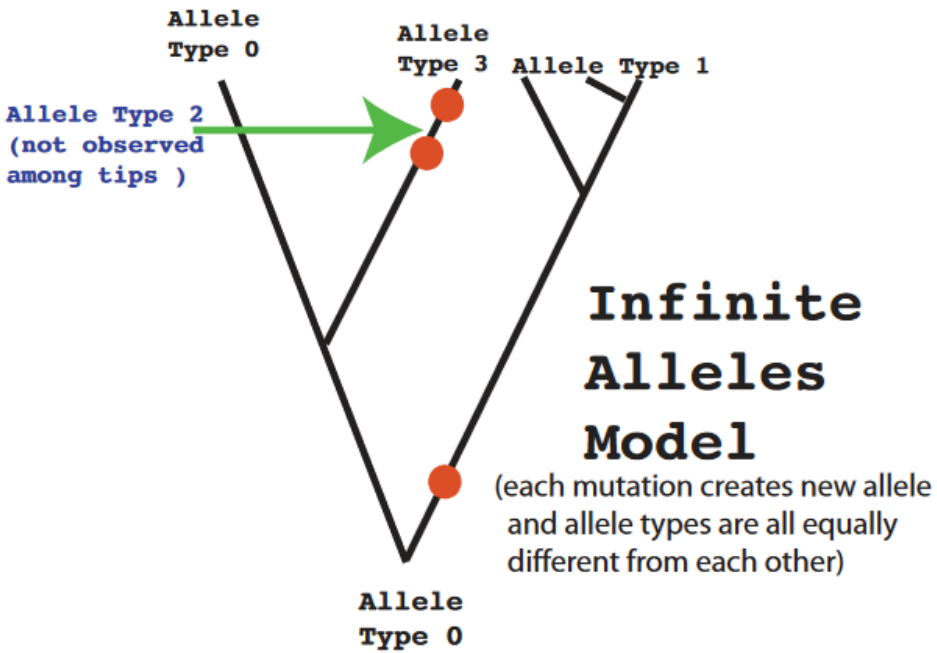
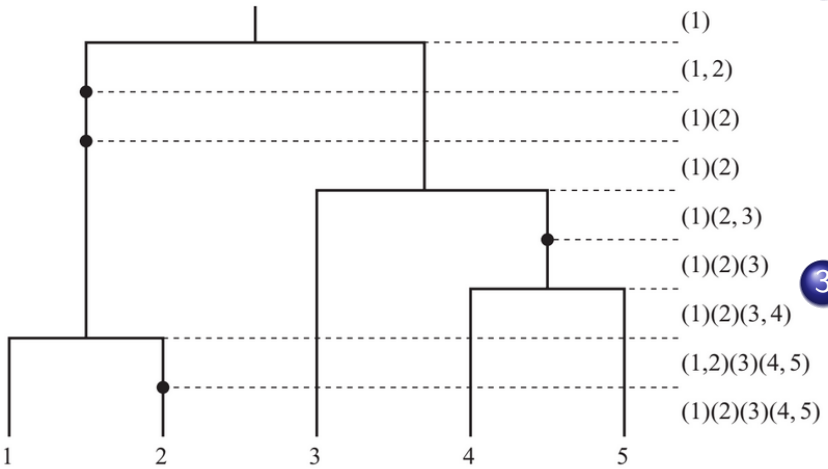


Figure: Adapted from this tutorial

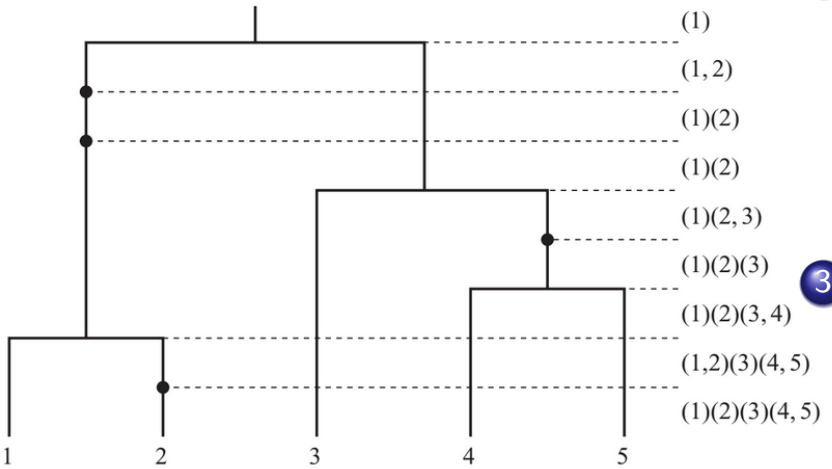
Infinite alleles model

Infinite alleles model



1. Data from the infinite alleles model can be represented as a vector $\vec{a} = (a_1, a_2, ..)$
2. $\vec{a}$ is called the **allelic partition** of the data(Xavier Didelot,Statistical population genetics tutorial,Univ of Oxford)
3. $n = \sum_1^n i a_i$;$K = \sum_i^n a_i$ is the number of allele types

Infinite alleles model



1. Data from the infinite alleles model can be represented as a vector $\vec{a} = (a_1, a_2, ..)$
2. $\vec{a}$ is called the **allelic partition** of the data(Xavier Didelot,Statistical population genetics tutorial,Univ of Oxford)
3. $n = \sum_1^n i a_i$;$K = \sum_i^n a_i$ is the number of allele types
4. For example: In the left example, $\vec{a} = (3, 1, 0, 0, 0); n = 5, K_5 = 4$

In addition, $P(K_n = k)$ is easily obtained; Ewens' sampling formula also provides the $P_n(\vec{a})$.

Question

Look at page 46 in book, as shown in equation (2.13),I put a question mark in second term. The first term means previous event is a mutation event,then how to understand the recursion part when a coalescent even occurs.

$$P_n(\mathbf{a}) = \frac{\theta}{n-1+\theta} P_{n-1}(\mathbf{a}-\mathbf{e}_1) + \frac{n-1}{n-1+\theta} \sum_{a_{j+1}>0} \frac{j(a_j+1)}{n-1} P_{n-1}(\mathbf{a}-\mathbf{e}_{j+1}+\mathbf{e}_j), \qquad (2.13)$$

Infinite sites model

Theorem (definition)

Each gene is made of a sequence of sites, and each time a mutation occurs, it affects a site that was previously unaffected,originally proposed by Kimura (1969, 1971) and Watterson (1975)

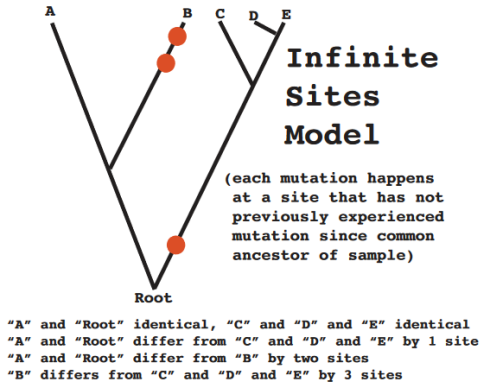


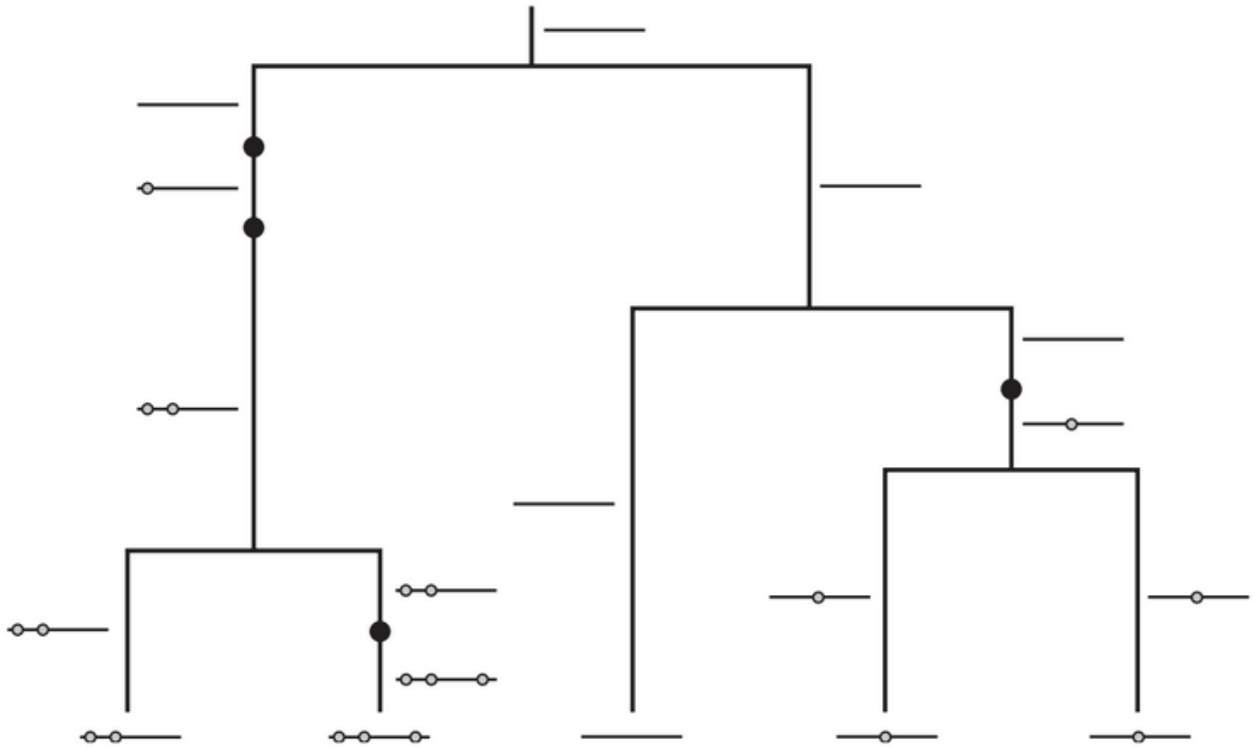
Figure: Adapted from this tutorial

Each site occurs mutation at most once, even though each site could mutate more than one times in reality.

Infinite sites model example

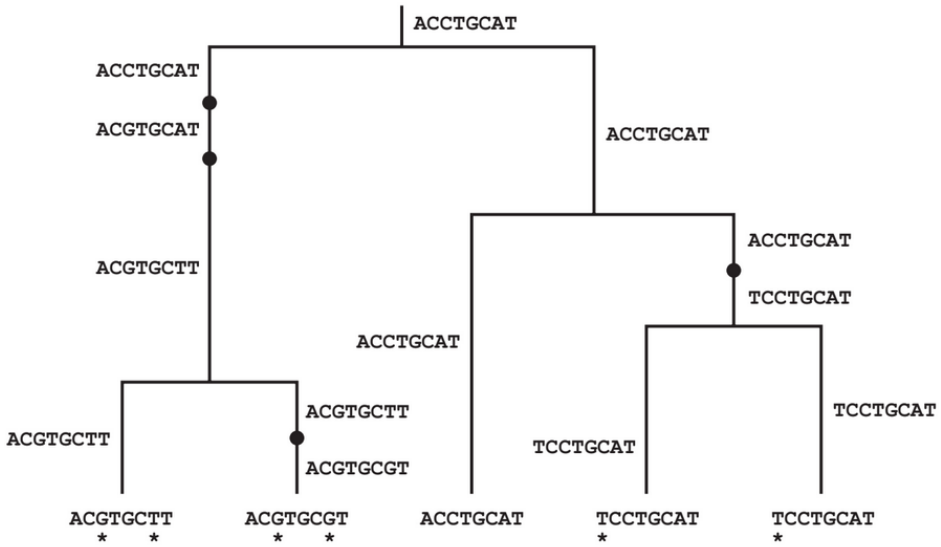
1	1	0	0
1	1	0	1
0	0	0	0
0	0	1	0
0	0	1	0

Figure: incidence matrix



Finite sites model

1. Recurrent mutations are allowed, namely, same position could happen over two mutations.
2. **Simplest** finite sites model is introduced by Jukes and Cantor in 1969.
3. The difference between finite sites model and previous two models is former only consider the substitutions, while infinite ones also take indels rates into account.



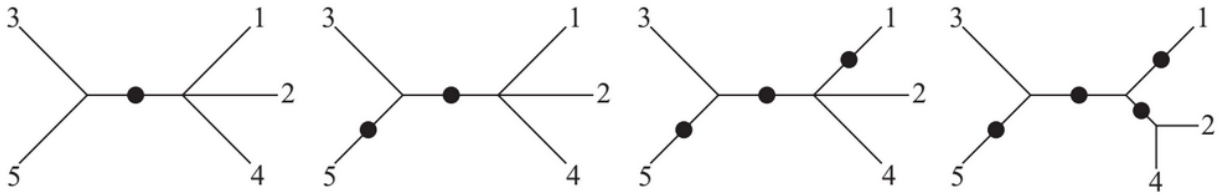
gene tree

1. gene tree is a visualization of the data.
2. gene tree is built under the infinite sites model.
3. gene tree **is not a coalescent tree**, since it doesn't provide information about the relative time of two internal nodes, however, all coalescent trees consistent with the sample must agree with clusters in the gene tree.

Build gene tree approach 1

index	a	b	c	d
1	0	0	0	0
2	0	0	1	1
3	1	0	1	0
4	0	0	1	1
5	1	1	1	0

Table: 1



Build gene tree approach 2

Gusfield(1991) provided an elegant algorithm to construct a gene tree.Initially, organsize the incidence matrix by changing table 1 to table 2, removing identical columns is also necessary. The first step is to check whether the data is compatible with a gene tree.In short,as long as all possible pairs 00,01,10,11 are present in any two columns,gene trees cannot be built, otherwise proceed.

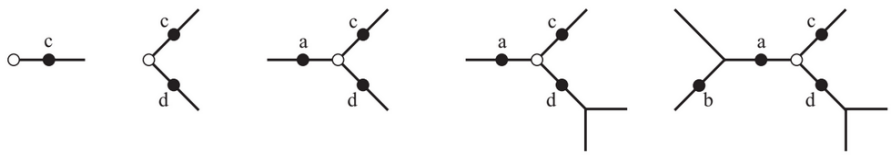
index	a	b	c	d
1	0	0	1	0
2	0	0	0	1
3	1	0	0	0
4	0	0	0	1
5	1	1	0	0

Build gene tree approach 2

Then consider each column as a binary number, sort the columns into decreasing order.Here how to consider each column as a binary number? Why column c is the largest binary number?(book page 75)

	c	d	a	b
1.	1	0	0	0
2.	0	1	0	0
3.	0	0	1	0
4.	0	1	0	0
5.	0	0	1	1.

The last step is to add sequence one at a time.



Wright-Fisher models

haploid reproduction model

The genes making up the present generation are drawn randomly with replacement from the parental generation.

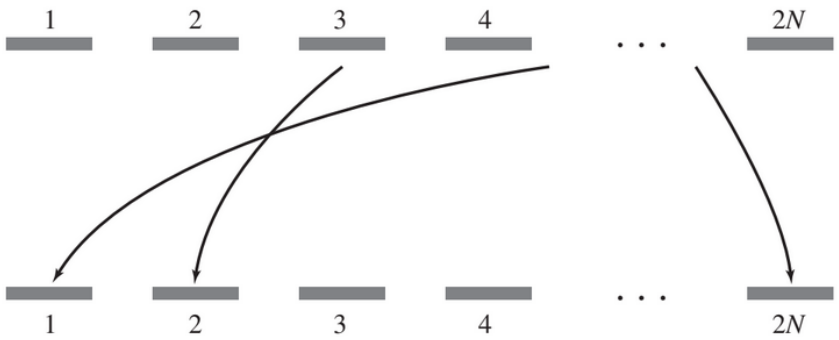
diploid reproduction model

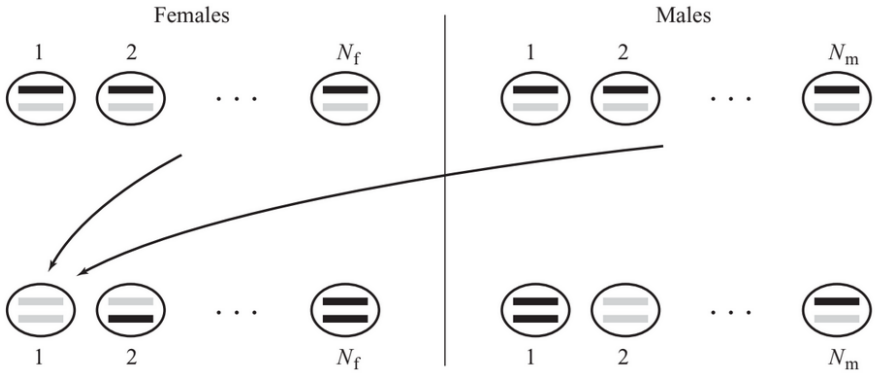
An individual in the present generation draws randomly with replacement one of its genes from the female population and the other gene from the male population.

Moran model

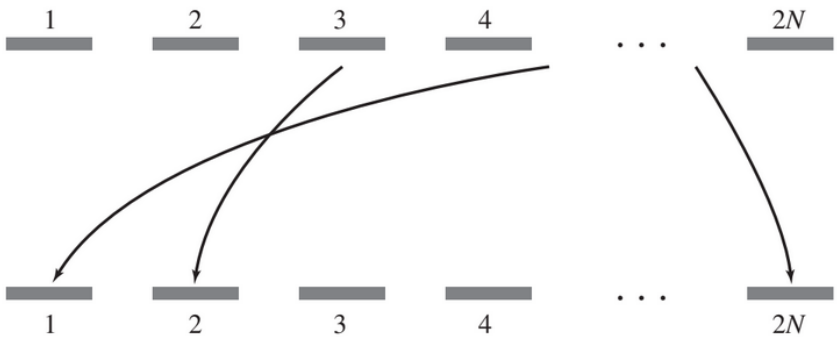
In each generation,one gene is chosen to given birth to one gene and one gene is chosen to die. All other gene survive to the next generation.

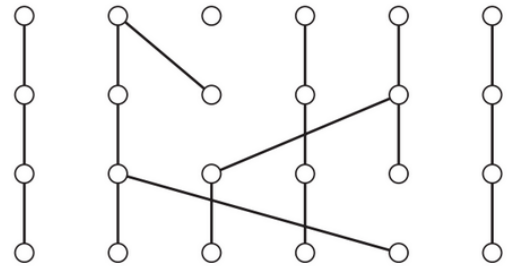
Visualize models

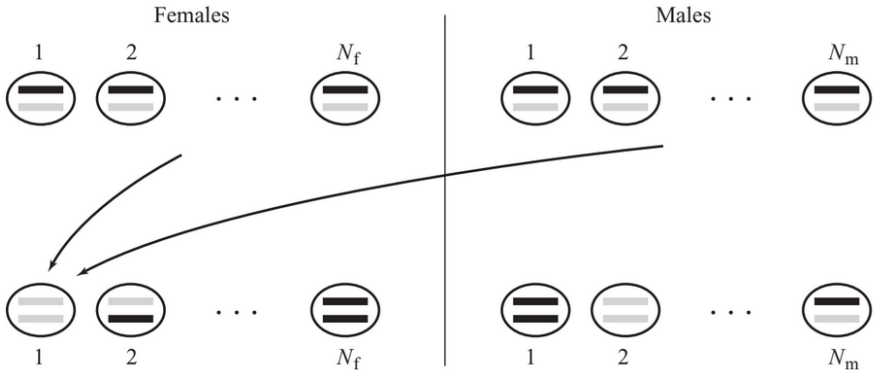




Visualize models



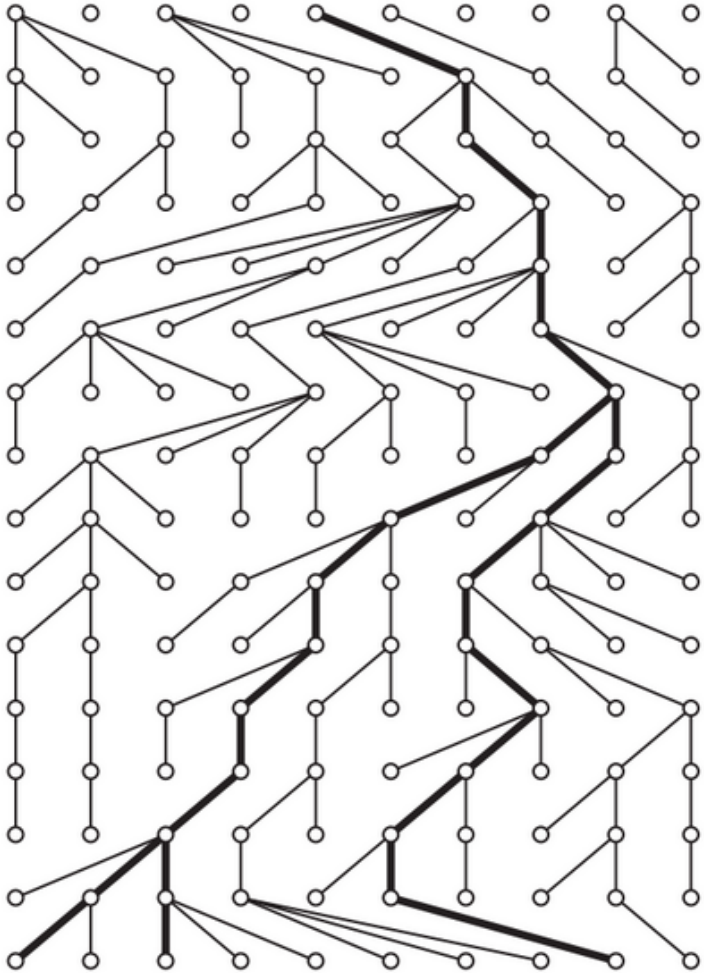




An example

MRCA;TMRCA

Wright-Fisher animator:http://www.coalescent.dk/



The number of descendants of a gene in one generation

Let v_i be the number of descendants of a gene i in generation t,$i = 1, 2, 3...2N$.

- $$p(v_i = k) = \binom{2N}{k}(\frac{1}{2N})^k(1 - \frac{1}{2N})^{2N-k}$$

The number of descendants of a gene in one generation

Let v_i be the number of descendants of a gene i in generation t,$i = 1, 2, 3...2N$.

- $$p(v_i = k) = \binom{2N}{k}(\frac{1}{2N})^k(1 - \frac{1}{2N})^{2N-k}$$

  and

  $$E(v_i) = 2N\frac{1}{2N} = 1$$

  which is the consequence of the population size being constant.

Important assumptions about Wright-Fisher model

- The population size is constant.
  1) exponential growth, $\beta = 2Nb$ is the scaled growth rate.

$$N(t) = N \exp^{-\beta t}$$

Important assumptions about Wright-Fisher model

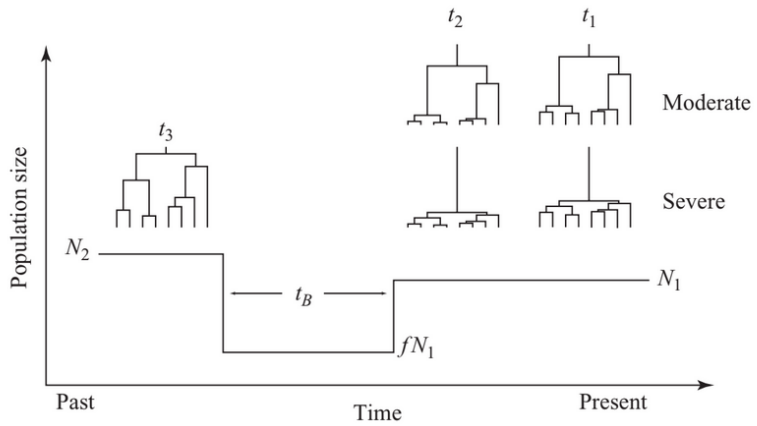
- The population size is constant.
  1) exponential growth, $\beta = 2Nb$ is the scaled growth rate.

$$N(t) = N \exp^{-\beta t}$$

and

$$t_k = \frac{1}{\beta} log(1 + \beta t_k^* \exp^{-\beta v_{k+1}})$$

  2) population bottleneck



Important assumptions about Wright-Fisher model

- The population has no geographical or social structure.
  1) Finite island model: assume a population is divided into d islands(demes),genes can only coalesce if they are in the same deme.
  2) Migration event occurs at rate:

$$I_{migr} = \frac{Mdk}{2}$$

  k is the total number of ancestors in all demes.$M = 4Nm$ is the scaled migration rate.Therefore,the probs of migration and coalescent are obtained.
  3) An algorithm for simulation is then provided. If a migration event occurs,a gene chosen randomly is migrated to a deme randomly.If a coalescent event occurs,a deme is chosen weighted, then two genes in that deme coalesce randomly.
- The genes are not recombing.

Effective population size N_e

Here we consider N_e's definitions and one generalization from Ewens(2004) since it is also suitable for variable population size(at book page 30).

- $$N_e = \frac{1}{2P(T_2 = 1)}$$

Effective population size N_e

Here we consider N_e's definitions and one generalization from Ewens(2004) since it is also suitable for variable population size(at book page 30).

- $$N_e = \frac{1}{2P(T_2 = 1)}$$

  and

  $$N_e^{(t)} = \frac{E(T_2)}{2}$$

  note:$P(T_i = j)$ means i genes find their MRCA at generation j.

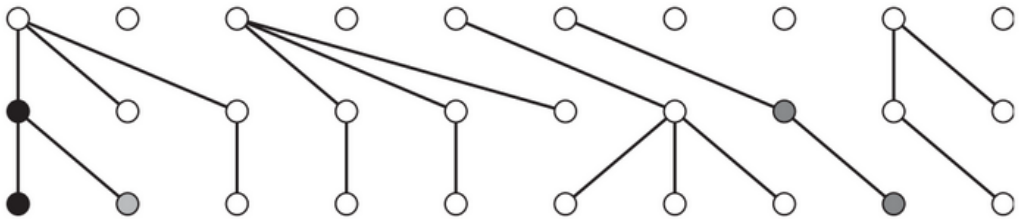
- for haploid model, $N_e = N_e^{(t)} = N$, then

  how to understand/compute the diploid situation?

Wright-Fisher model with mutation

Two independent points:

1. Whether there is a mutation in one gene in one generation is independent of whether there is another gene in the same generation;
2. Coalescent event and mutation event is independent of each other.



Wright-Fisher model with mutation

Both coalescent event and mutation event are exponentially distributed, parameters are $\binom{n}{2}$ and $\frac{n\theta}{2}$, here θ is population mutation rate$4Nu$.

- The first event is coalescent event with probability

$$\frac{\binom{n}{2}}{\binom{n}{2}+\frac{n\theta}{2}} = \frac{n-1}{n-1+\theta}$$

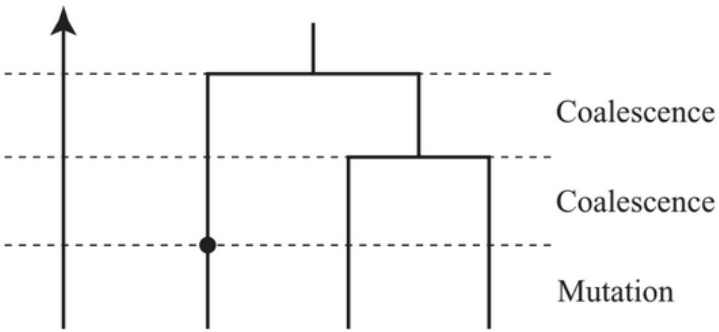
- or mutation event with probability

$$\frac{\frac{n\theta}{2}}{\binom{n}{2}+\frac{n\theta}{2}} = \frac{\theta}{n-1+\theta}$$

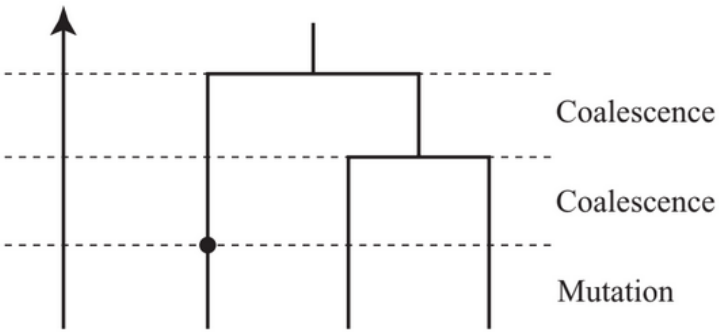
In addition,we have $\binom{n}{2}+\frac{n\theta}{2} = \frac{(n-1+\theta)n}{2}$

Wright-Fisher model with mutation

Wright-Fisher model with mutation



Wright-Fisher model with mutation



Algorithm 2

1. Start with $k = n$ genes,with previous probabilities of two different events
2. If coalescent event occurs,merge a pair randomly into one gene,$k \to k - 1$
3. If mutation event occurs,choose one lineage to mutate,leave k unchanged.
4. repeat until $k = 1$

Wright-Fisher model with recombination

Both coalescent event and recombination event are exponentially distributed, parameters are $\binom{n}{2}$ and $\frac{n\rho}{2}$, here ρ is population recombination event rate $4Nr$.

- The first event is coalescent event with probability

$$\frac{\binom{n}{2}}{\binom{n}{2}+\frac{n\rho}{2}} = \frac{n-1}{n-1+\rho}$$

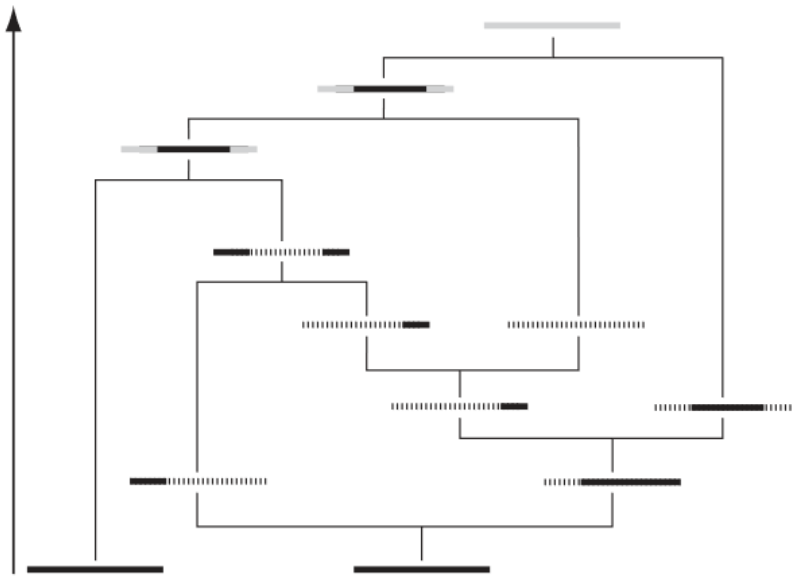
- or recombination event with probability

$$\frac{\frac{n\rho}{2}}{\binom{n}{2}+\frac{n\rho}{2}} = \frac{\rho}{n-1+\rho}$$

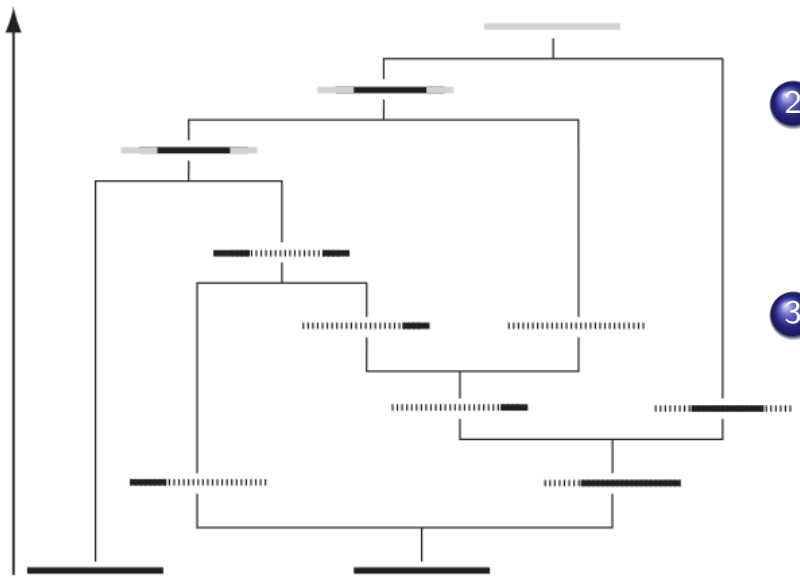
In addition,we have $\binom{n}{2}+\frac{n\rho}{2} = \frac{(n-1+\rho)n}{2}$

Wright-Fisher model with recombination(ARG)

Wright-Fisher model with recombination(ARG)



Wright-Fisher model with recombination(ARG)



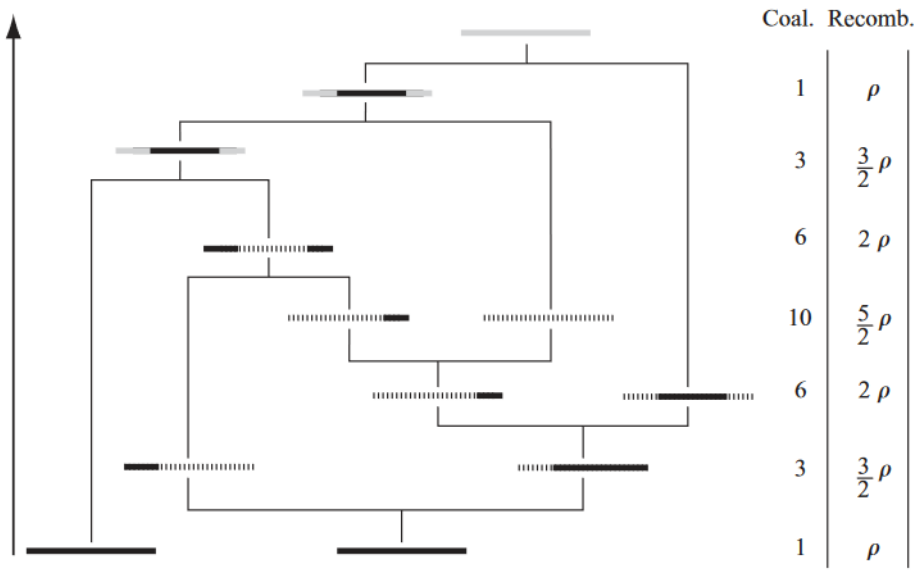
Algorithm 3

1. Start with $k = n$ genes,with previous probabilities of two different events
2. If coalescent event occurs,merge a pair randomly into one gene,$k \to k - 1$
3. If recombination event occurs,choose breakpoint uniformly in random sequence,$k \to k + 1$.
4. repeat until $k = 1$

Wright-Fisher model with recombination

Two noteworthy points:

1. It has low probability ($\frac{1}{2N} * \frac{\rho}{4N} = \frac{\rho}{8N^2}$)for two sequences experiencing coalescent and recombination events at the same time;
2. Finally GMRCA will be found,since coalescent intensity $\propto k^2$,recombination intensity $\propto k$.



The End