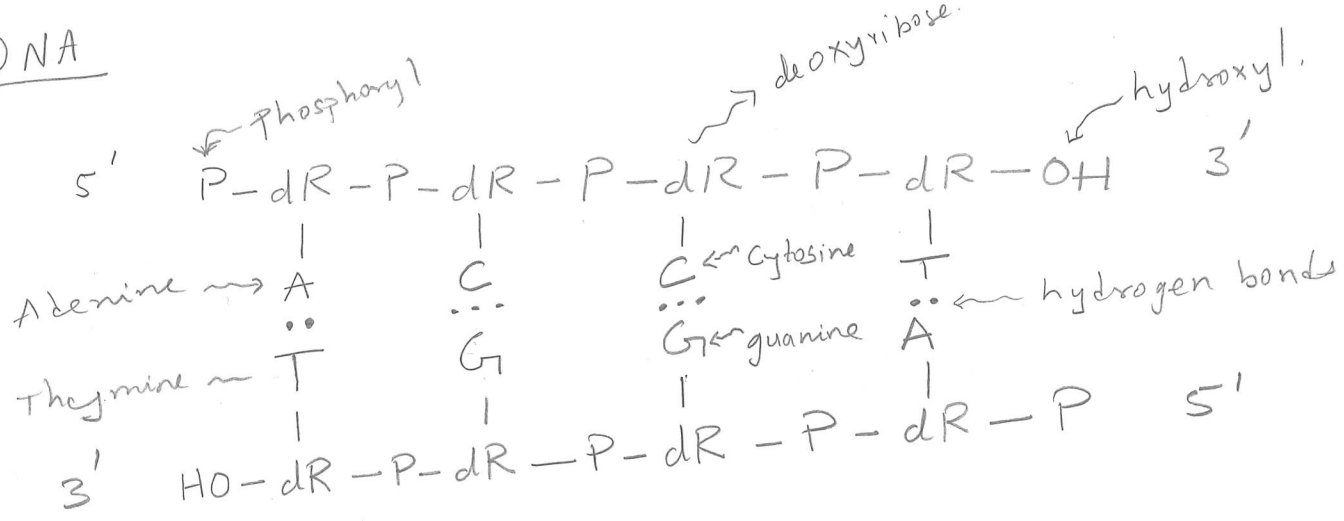
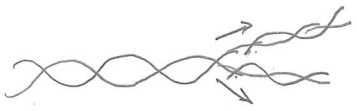
DNA



- Hereditary information of living things is carried by DNA sequences $\in \{A,C,G,T\}^N$ (its genome)
  - N: set of nucleotides
  - For yeast $N \approx 12M$
- Double-stranded DNA can replicate! by copying from each strand

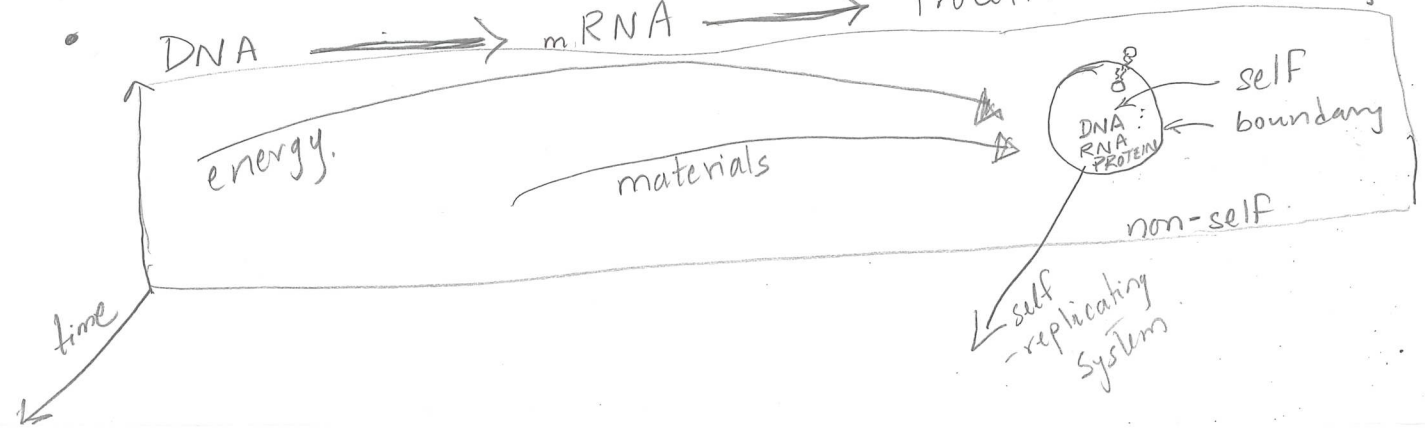


- Nucleotide frequencies is non-uniform for Yeast

$$\begin{cases} A = 0.3090 \\ T = 0.3078 \\ C = 0.1917 \\ G = 0.1913 \end{cases}$$

- organisms have different number of copies of their genome (organized by partitions called chromosomes)
  - 1 (haploid) eg. bacteria
  - 2 (diploid) eg. humans, (most animals)
  - 4 (tetraploid) eg. plants.
  - 6 (hexaploid) eg. wheat
  - >6 (polyploid) eg. sorghum has 100 chromosomes of 8 types?

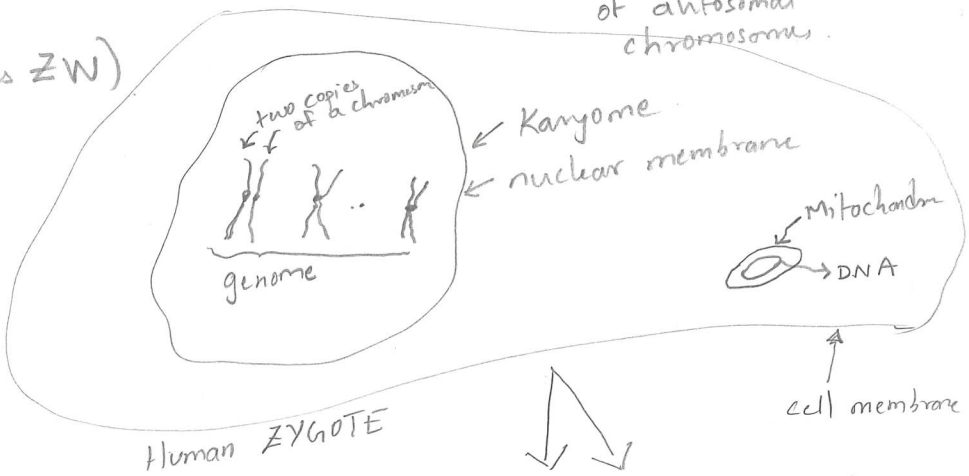
- DNA → mRNA → Protein $:= \{20 \text{ Amino Acids}\}^m$

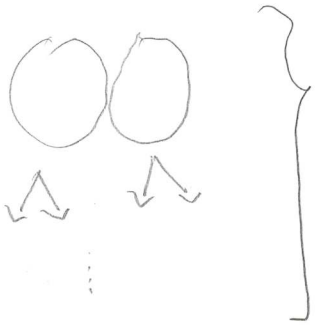


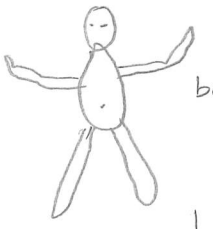
Sexual Reproduction in diploids & recombination

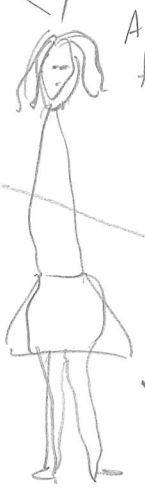
humans → 1 sex chromosome, females XX ; males XY
→ 22 autosomes, females & males have two copies AA.
of autosomal chromosomes.

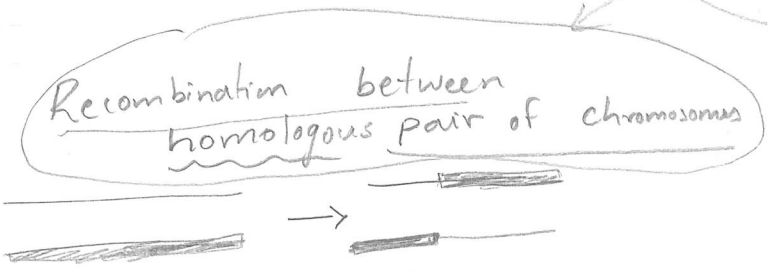
(birds: males ZZ ; females ZW)

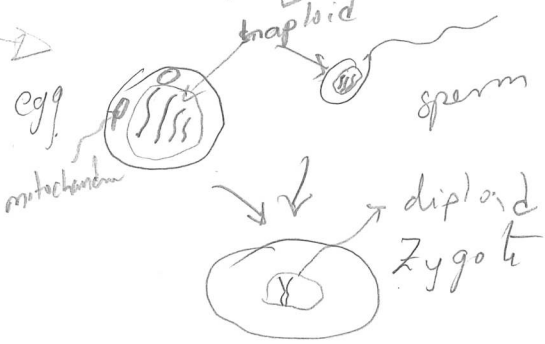












No recombination
- Y chromosome
& Mitochondrial DNA

Wright-Fisher Model (1931)

genetic locus := a location in the genome of an organism

Alleles := different versions of the genetic information encoded at a locus.

eg Alleles A and a could represent distinct DNA sequences:

posn 10193 on chrm 13

A = CTGAAATCGTAA
a = CTGATATCGTAG
X X

or just at a single position on a chromosome = **site**

site at posn. 15326 on chrm 14

A = G
a = T

Diploid organisms have two copies so they will be

A A or a a
A a a

(AA) (Aa) (aa)

Three genotypes at a biallelic locus

fitness of an individual is a measure of its ability to survive and to produce offspring.

Mutation when DNA replicates mistakes can be made with small probability. eg:- or mutations

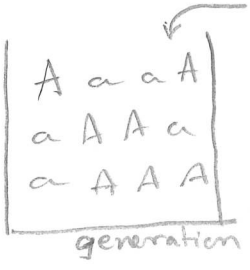
ACGTTACGGCCT
TGCAATGCTGGA
TGC

(mutations change the DNA sequence of the original template DNA)

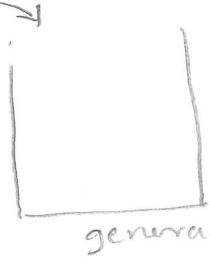
neutral evolution when mutation does not affect fitness.

W-F Model

as Sampling with replacement from an Urn with 2N balls.



urn with 2N balls



- nonoverlapping generation (deaths & rebirths every generation)
- random mating.

At gen. n i balls (individuals) have allele A and $2N-i$ balls have allele a

$2N$ = number of haploid alleles in a diploid population of size N.

To build up the $(n+1)$-th gen. choose from the urn at gen n, $2N$ times with replacement

X_n = Number of A's in gen. n

X_n is a Markov chain on $\{0, 1, 2, \cdots, 2N\}$

Since $\Pr\left(X_{n+1} = j \mid X_n = i, X_{n-1} = x_{n-1}, \cdots, X_0 = x_0\right)$

$= \Pr\left(X_{n+1} = j \mid X_n = i\right)$

$=: p(i,j)$

$= \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(\frac{2N-i}{2N}\right)^{2N-j}$

PMF of $\rightarrow$ Binomial$\left(2N, \frac{i}{2N}\right)$ RV

$\left(\frac{i}{2N}\right)$: Prob. of choosing an A allele

$\left(\frac{2N-i}{2N}\right)$: Prob. of choosing an a allele

$= \sum_{i=1}^{2N}$ Bernoulli$\left(\frac{i}{2N}\right)$

where, $\binom{2N}{j} := \frac{(2N)!}{j!\,(2N-j)!}$ = number of ways of choosing j items out of $2N$ items.

Binomial coefficient

where $j! = j(j-1)(j-2)\cdots 2 \cdot 1$

j factorial is number of ways of ordering j items in a row.

Long-term behaviour of $\{X_n\}_{n=0}^{\infty}$, the W-F Markov chain for the number of A alleles among $2N$ alleles.

Either $X_n \xrightarrow{n\to\infty} \begin{cases} 0 & \text{with no } A \text{ alleles} \\ 2N & \text{with no } a \text{ alleles} \end{cases}$

Because 0 & $2N$ are absorbing states.
(the chain can never leave once it enters either 0 or $2N$)

We say fixation has occurred when X_n enters an absorbing state
(whole population is fixated on one allele).

fixation Time τ, i.e first time the popn is all a's or all A's.

$$\tau := \min\{n : X_n = 0 \text{ or } X_n = 2N\}$$

Thm 1 In the W-F Model, the prob of fixation in the all A's state, given you start with i A's is:

$$\Pr(X_\tau = 2N \mid X_0 = i) = \frac{i}{2N}$$

Proof: Since $2N < \infty$, fixation with all A's or a's will eventually occur (i.e. 0 & $2N$ are absorbing states).

Since the expectation of the Binomial(n,p) RV is np

$$E(X_{n+1} \mid X_n = i) = 2N\left(\frac{i}{2N}\right) = i = X_n$$

Taking expected value on both sides, we get

$$E(X_{n+1}) = E(X_n) \rightsquigarrow$$ so average value of the number of A's remains the same throug time.

Now, since $X_n = X_\tau$ when $n > \tau$

$$(\star) \quad i = E(X_n \mid X_0 = i) = E(X_\tau;\ \tau \le n \mid X_0 = i) + E(X_\tau;\ \tau > n \mid X_0 = i)$$

where, $E(X;A)$ = Expected value over the set A.

Now let $n \to \infty$ with the fact that $|X_n| \leq 2N$, we get that:

$$E(X_\tau ; \tau \leq n \mid X_0 = i) \to E(X_\tau \mid X_0 = i)$$

and $$E(X_\tau ; \tau > n \mid X_0 = i) \to 0.$$

So, from (✡):

$$i = E(X_\tau \mid X_0 = i) = 2N \cdot \Pr(X_\tau = 2N \mid X_0 = i)$$

and $\therefore$ $$\Pr(X_\tau = 2N \mid X_0 = i) = \frac{i}{2N}$$ ▨

[Thus, the prob. of being fixed in all A's state having started with i A alleles is simply $\frac{i}{2N}$, The proportion of A alleles at the start.]

Mutation Now suppose that mutations occur in the population, whereby $A \to \alpha$ (mutates to) or $a \to \alpha$, at rate μ. Then from Thm 1 we get Kimura's result:

Thm 2 Under W-F model, the rate of fixation of neutral mutations in a haploid pop^n of size $2N$ is the mutation rate μ.

Proof. Note that mutations occur to some individual in the population at rate $2N\mu$ and since this is the only individual with this mutation, it goes to fixation (by Thm 1) with prob. $\frac{1}{2N}$.

Heterozygosity is the prob. that two copies of the locus chosen (without replacement) at time n are different:

$$H_n^0 := \frac{X_n(2N-X_n)}{\frac{2N(2N-1)}{2}}$$

← number of choosing an A and an a allele

← $\binom{2N}{2} = \frac{2N!}{(2N-2)!\,2!} = \frac{2N(2N-1)}{2}$

The number of ways of choosing any two out of $2N$ items

$$= \frac{2X_n(2N-X_n)}{2N(2N-1)}$$

Thm 3 Let $h(n) := E(H_n^0)$, the expected heterozygosity at time n in the W-F. model. Then,

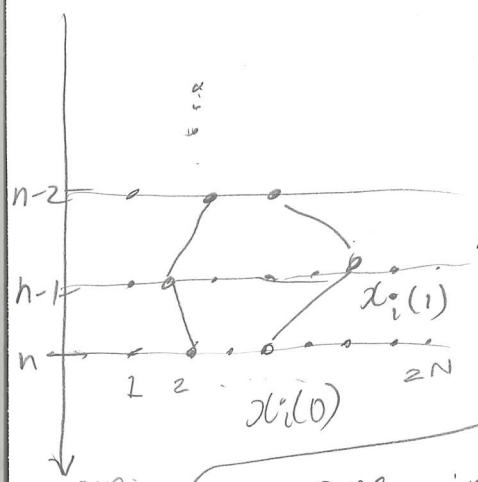
$$h(n) = \left(1-\frac{1}{2N}\right)^n h(0)$$

Proof Let us label the alleles (2N copies of the locus.) or individuals by $1, 2, \dots, 2N$

Suppose we pick two individuals at time n labelled by $x_1(0)$ and $x_2(0)$.

Each indiv. $x_i(0)$, $i = 1, \dots, 2N$ is a descendant of some individual $x_i(1)$ at time $n-1$, who is a descendant of $x_i(2)$ at time $n-2$, etc.

So, $(x_i(m) : 0 \le m \le n)$ gives the genealogy or ancestral lineage of $x_i(0)$ (ancestors back in time)



Let $x_1(0)$ and $x_2(0)$ be two randomly chosen indivs.

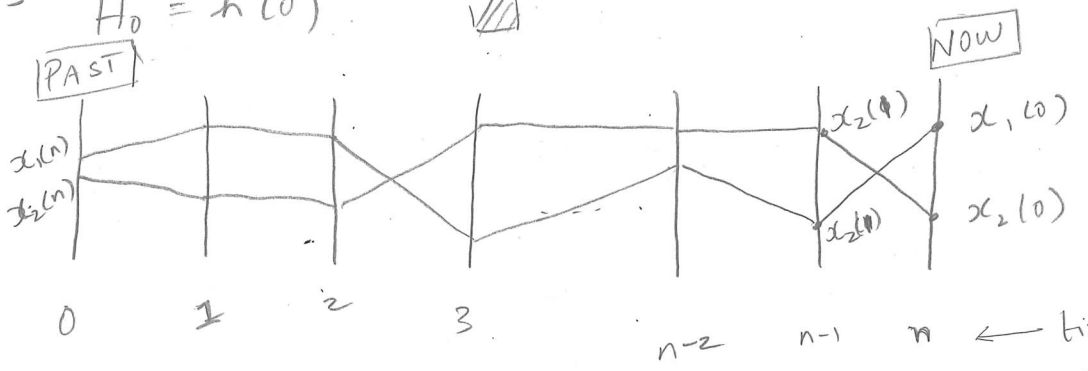
Note: If $x_1(m) = x_2(m)$ then $x_1(l) = x_2(l)$ for $m \le l \le n$.

• If $x_1(m) \neq x_2(m)$ the parental choices are indep. with $\Pr\{x_1(m+1) \neq x_2(m+1)\} = 1 - \frac{1}{2N}$

For $x_1(n) \neq x_2(n)$, different parents should be chosen for all times $1 \leq m \leq n$,

$$\Pr\left\{ x_1(m) \neq x_2(m) : 1 \leq m \leq n \right\} = \left(1 - \frac{1}{2N}\right)^n$$

Finally $x_1(n)$ and $x_2(n)$ are two random indivs from time 0, so the prob they are different is $H_0^0 = h(0)$ ▨



A pair of genealogies remaining distinct for n gens.

—x—

The Coalescent (Kingman 1982)

When x is small $\boxed{1 - x \approx e^{-x}}$ if $2N$ is large

Thm 3 (with $\frac{1}{2N}$ small)

$$h(n) = \left(1 - \frac{1}{2N}\right)^n h(0) \approx e^{-n/2N} h(0)$$

Suppose, instead of 2, we sample K random individuals, then the prob. 2 of the k will pick the same parent in prev. gen. is.

$$\approx \underbrace{\frac{k(k-1)}{2}}_{\text{number of ways of picking 2 from K indivs}} \cdot \underbrace{\frac{1}{2N}}_{\text{The prob. that These two indivs will choose thy same parent}}$$

it's only approximate because we ignore events with prob. of $O\left(\frac{1}{N^2}\right)$ eg: three indivs. choose same parent.

Thm 4 When measured in units of $2N$ generations, the amount of time during which there are k lineages, t_k, has approximately exponential distribution with rate $k(k-1)/2$.

Proof $\Pr\{k \text{ lineages remain distinct for } n \text{ generations}\}$

$$\approx \left(1 - \frac{k(k-1)}{2} \cdot \frac{1}{2N}\right)^n \approx \exp\left(-\frac{k(k-1)}{2} \cdot \frac{n}{2N}\right)$$

Recall $T \sim \text{Exponential}(\lambda)$ ← rate parameter

$P(T > t) = e^{-\lambda t}$ $\quad E(T) = \frac{1}{\lambda}$

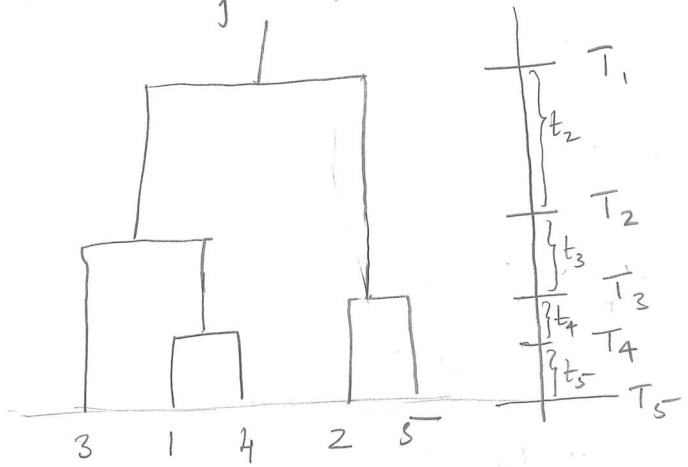
By letting $N \to \infty$ and expressing time in terms of $2N$ generations, i.e letting $t = n/2N$, we get

The time to the first coalescence (choosing same ancestor) event is Exponential $\left(\frac{k(k-1)}{2}\right)$ RV.

Thus k lineages coalesce to $k-1$ lineages at rate $k(k-1)/2$ using continuous time Markov chain (CTMC) terminology. ▨

The limit of genealogies in Thm 4 is called **the coalescent**

Let T_j = first time with j lineages, then we have



A realisation of the coalescent for a sample of size 5 drawn randomly from a large pop^n of size $2N$.

Expected coalescent times (we use lower case for RVs here for exposition)

$$E(t_2) = 1 \quad , \quad E(t_3) = \frac{1}{3} \quad , \quad E(t_4) = \frac{1}{6} \quad , \quad E(t_5) = \frac{1}{10}$$

$\because E(t_k) = \frac{2}{k(k-1)}$

T_1 = time of the appearance of the Most Recent Common Ancestor (MRCA) of the sample

$$= t_n + t_{n-1} + \cdots + t_3 + t_2$$

$$E(T_1) = E\left(\sum_{k=2}^{n} t_k\right) = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\left(\frac{1}{k-1} - \frac{1}{k}\right)$$

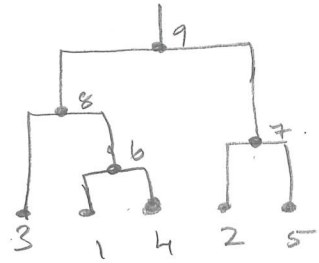
$$= 2\left(\sum_{k=1}^{n-1} \frac{1}{k} - \sum_{k=2}^{n} \frac{1}{k}\right) = 2\left(1 - \frac{1}{n}\right) \xrightarrow[n\to\infty]{} 2$$

for large samples

But $E(t_2) = 1 \approx \frac{E(T_1)}{2}$

Simulating The Coalescent

Let's label internal nodes.



$V_3 = \{7, 8\}$
$V_2 = \{3, 6, 7\}$
$V_1 = \{2, 3, 5, 6\}$
$V_0 = \{1, 2, 3, 4, 5\}$

input sample size n

initialize $V_0 = \{1, 2, 3, \cdots, n\}$, $T_n = 0$

for $k = 0, 1, \cdots, n-2$ do

- pick i_k and j_k from V_k
- $V_{k+1} \leftarrow V_k \setminus \{i_k, j_k\} \cup \{n+k+1\}$
- In tree connect $i_k \to n+k+1$ and $j_k \to n+k+1$
- Let $t_{n-k} \sim$ Exponential $\left(\binom{n-k}{2}\right)$ $\rightsquigarrow \binom{n-k}{2}^{-1} \log\left(\frac{1}{u}\right)$
- Let $T_{n-k-1} \leftarrow T_{n-k} + t_{n-k}$

Tree(2): label branch by smaller number on lower end $1 \le i \le 2n-2$

Tree(1): anc$[i_k] = n+k+1$, anc$[j_k] = n+k+1$

After generating genealogy

(+) Mutations

μ = mutation rate per gen.

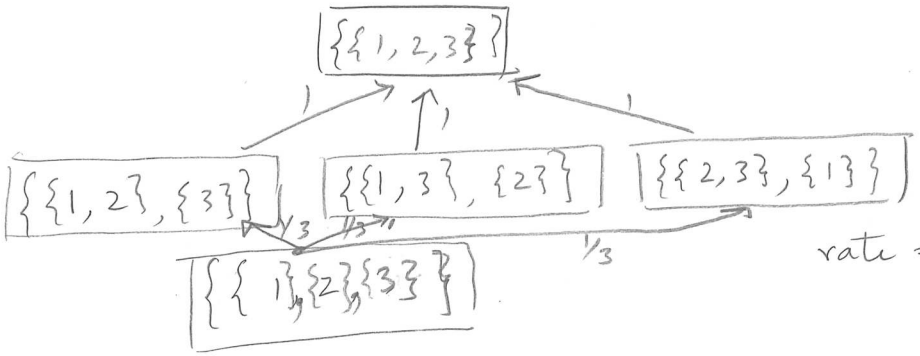
$\theta = 4N\mu$

for branch $i = 1$ to $2n-2$

$X(i) \sim \text{Poi}\left(\frac{\theta}{2}(T_{\text{anc}[i]} - T_i)\right)$ distributed uniformly across m sites

State space of the coalescent

ξ_n = set of all set partitions of $\{1, 2, \ldots, n\}$

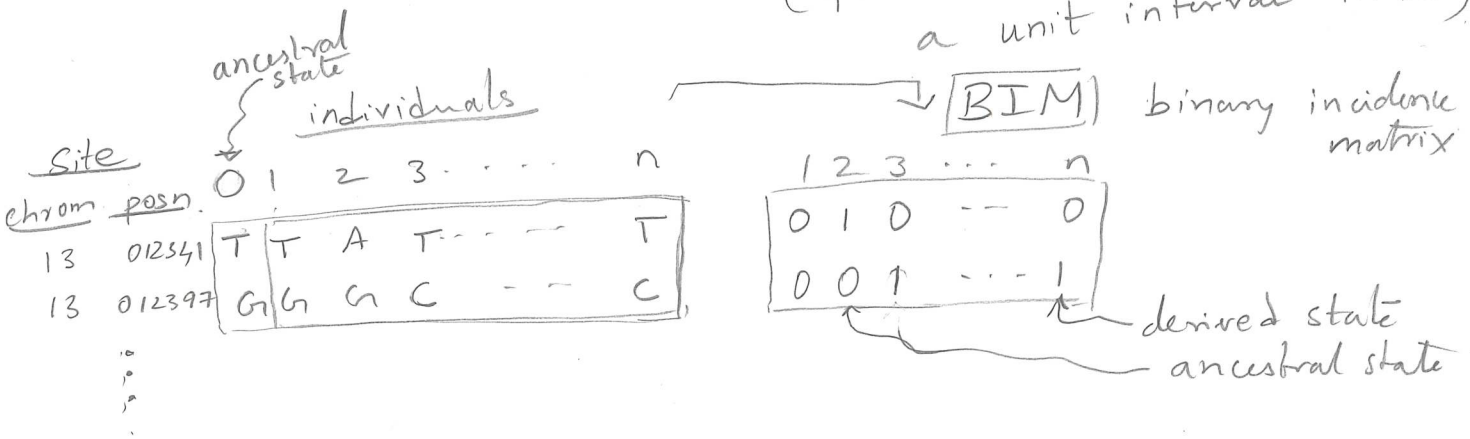


$$\text{rate} = \binom{3}{2} = \frac{3!}{1!\,2!} = 3$$

Infinite Sites Model (Kimura, 1969)

Mutations always occur at distinct sites.

(picture DNA sequence as a unit interval in $\mathbb{R}$)



Segregating Sites

S_n = number of segregating sites = number of site positions where some pair of individual DNA sequences differ.