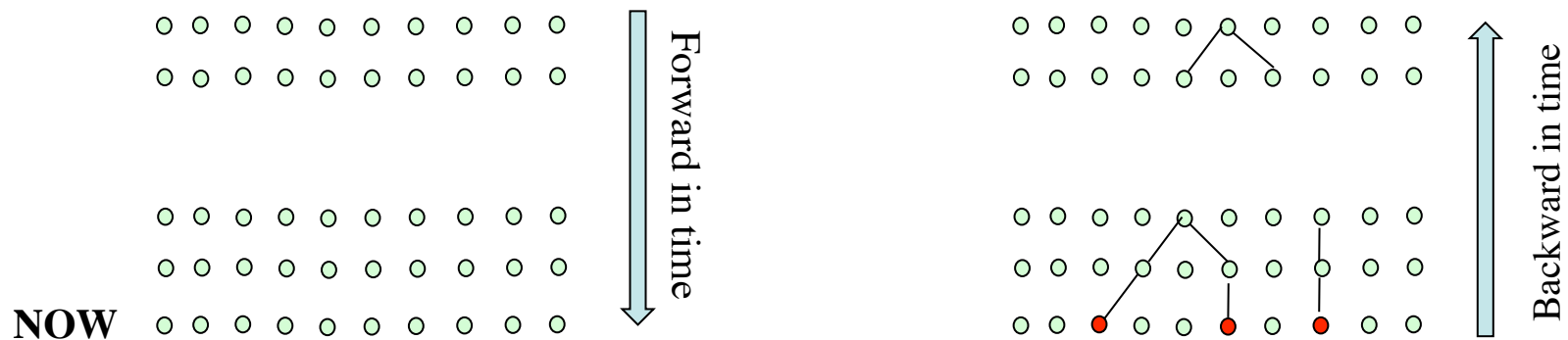
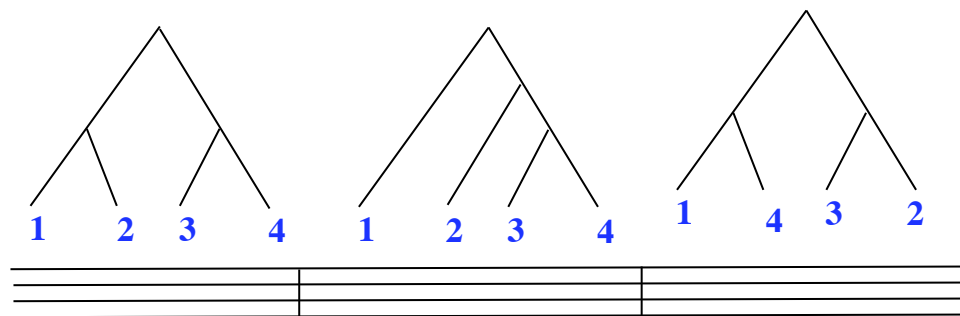


Population Genetics, Recombination Histories & Global Pedigrees

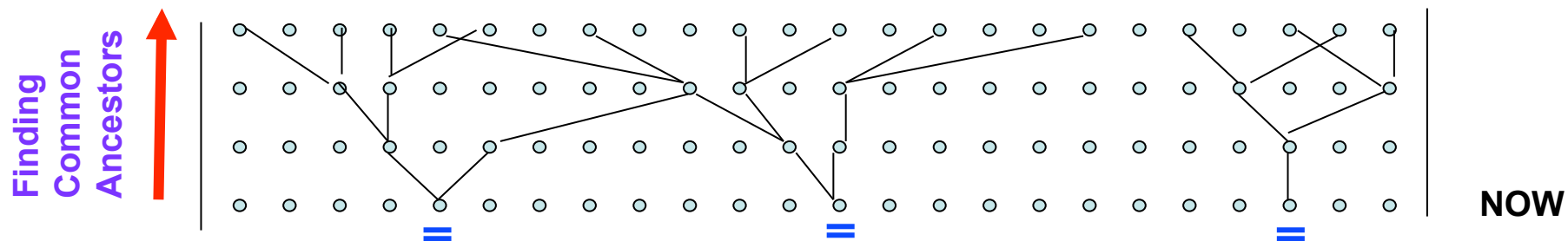
Population Genetics and Genealogies



Finding Minimal Recombination Histories



Global Pedigrees

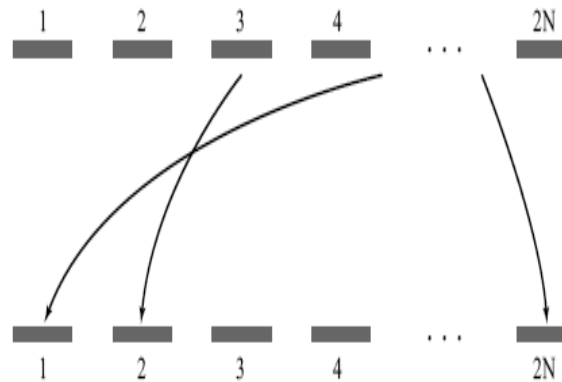


Wright-Fisher Model of Population Reproduction

Haploid Model

i. Individuals are made by sampling with replacement in the previous generation.

ii. The probability that 2 alleles have same ancestor in previous generation is $1/2N$

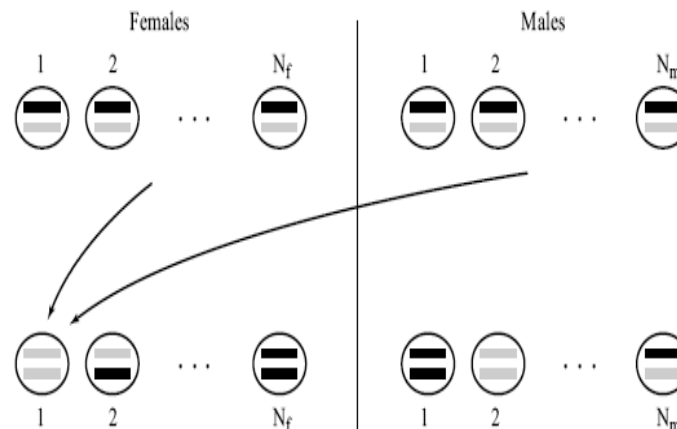


Assumptions

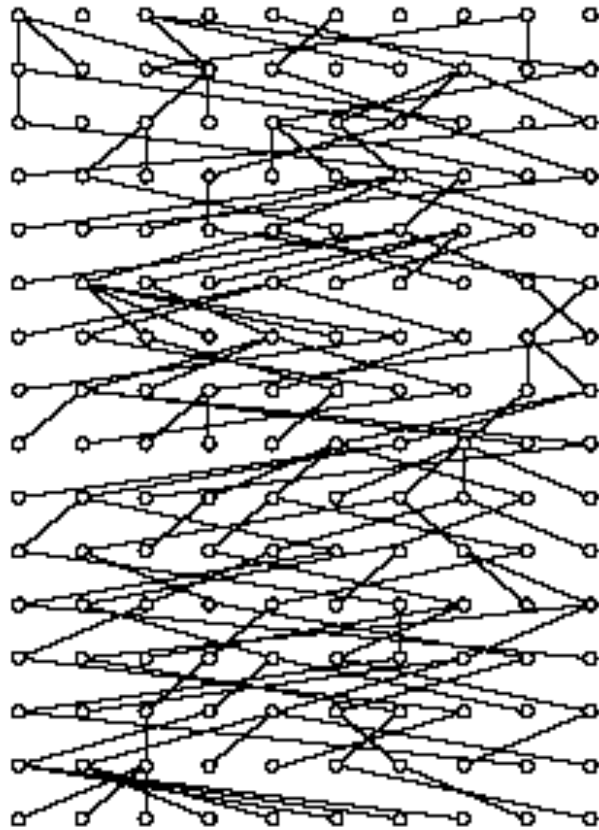
1. Constant population size
2. No geography
3. No Selection
4. No recombination

Diploid Model

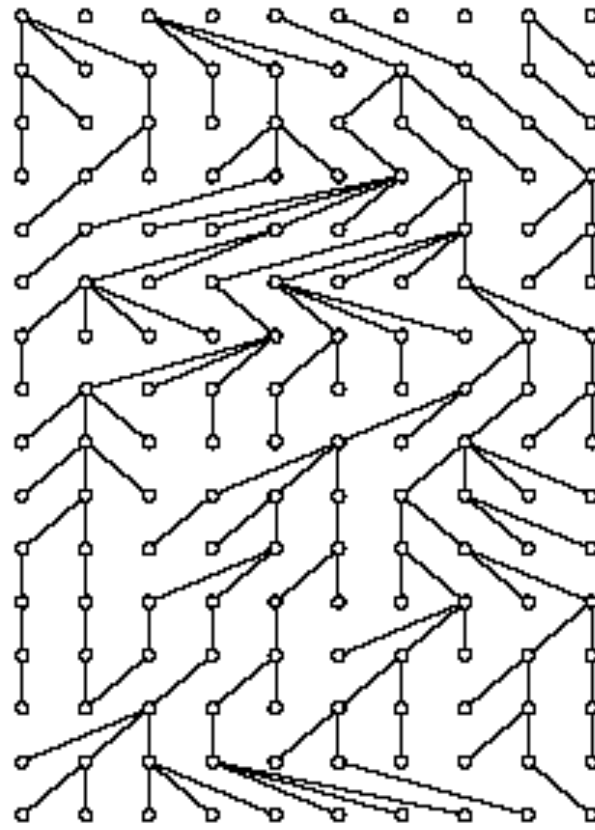
Individuals are made by sampling a chromosome from the female and one from the male previous generation with replacement



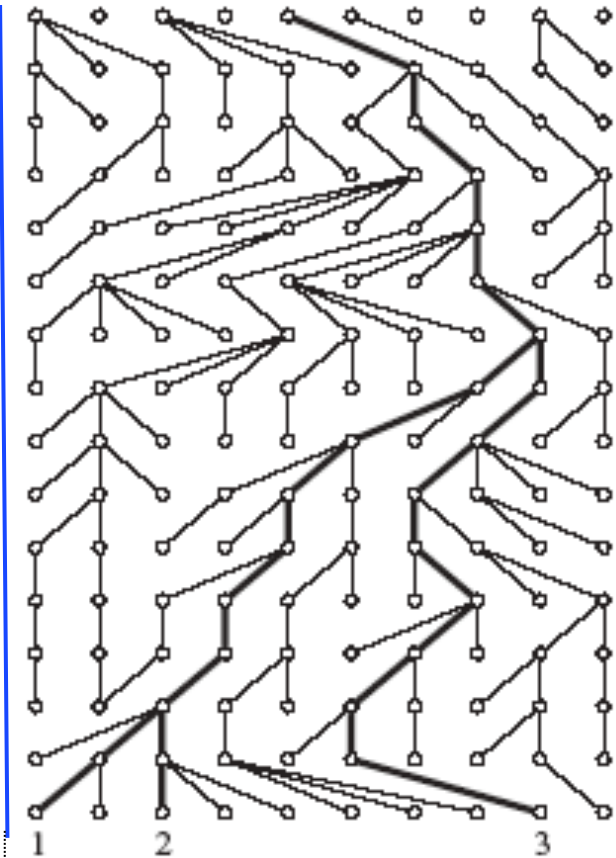
10 Alleles' Ancestry for 15 generations



Generate new generations according to the haploid model



Sort generations forward in time, such that individual 1's children in the next generation comes first, then the children of 2..

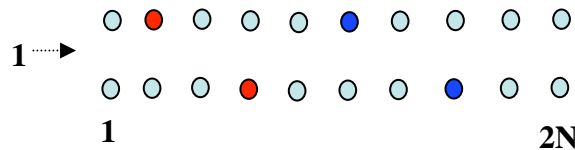


Pick individuals in the present and label edges to their ancestors back in time

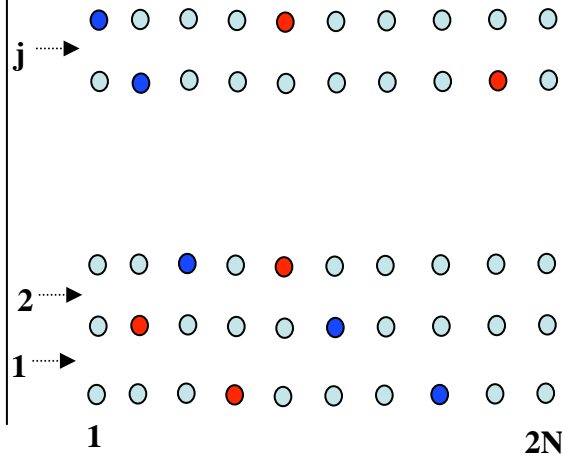
Waiting for most recent common ancestor - MRCA

Distribution of time until 2 alleles had a common ancestor, X_2 ?:

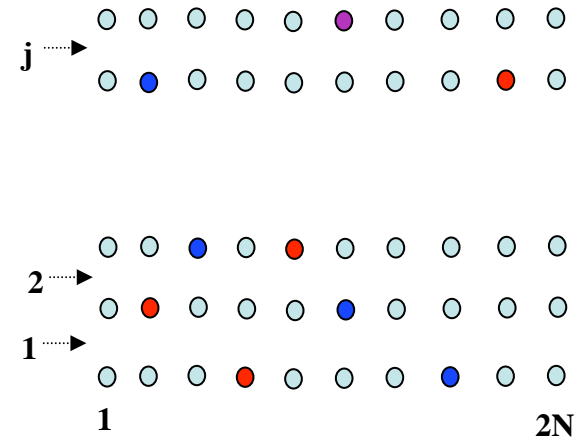
$$P(X_2 > 1) = (2N-1)/2N = 1-(1/2N)$$



$$P(X_2 > j) = (1-(1/2N))^j$$



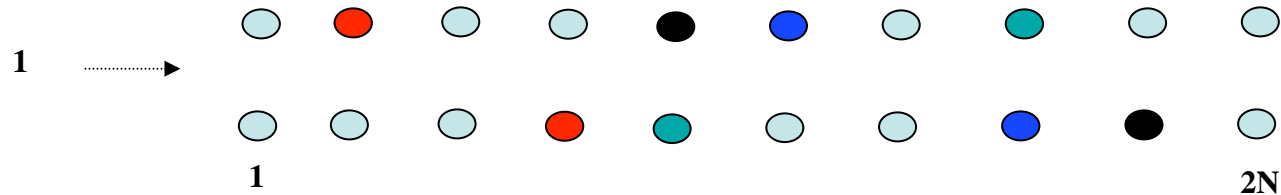
$$P(X_2 = j) = (1-(1/2N))^{j-1} (1/2N)$$



Mean, $E(X_2) = 2N$.

Ex.: $2N = 20.000$, Generation time 30 years, $E(X_2) = 600000$ years.

$P(k) := P\{k \text{ alleles had } k \text{ distinct parents}\}$



Ancestor choices:

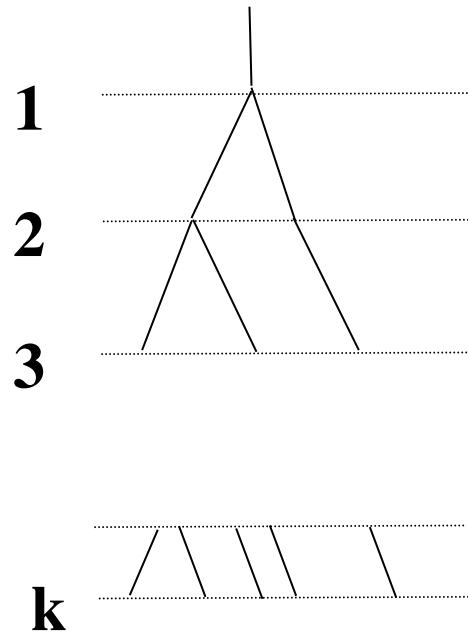
$k \rightarrow \text{any}$	$k \rightarrow k$	$k \rightarrow k-1$	$k \rightarrow j$
$(2N)^k$	$2N * (2N-1) * \dots * (2N-(k-1))$ $=: (2N)_{[k]}$	$\binom{k}{2} (2N)_{[k-1]}$	$S_{k,j} (2N)_{[j]}$

$S_{k,j}$ - the number of ways to group k labelled objects into j groups. (Stirling Numbers of second kind).

For $k \ll 2N$:

$$P(k) = \frac{2N_{[k]}}{(2N)^k} \approx (k^2 < 2N) \left(1 - \binom{k}{2} / 2N\right) \approx e^{-\binom{k}{2} / 2N}$$

Expected Height and Total Branch Length



Time Epoch

1

1/3

$$\frac{1}{\binom{k}{2}} = \frac{2}{k(k-1)}$$

Branch Lengths

2

1

2/(k-1)

Expected Total height of tree: $H_k = 2(1 - 1/k) + \dots + 1/(k-1)$ ca= $2 \ln(k-1)$

i. Infinitely many alleles finds 1 allele in finite time.

ii. It takes less than twice as long for k alleles to find 1 ancestor as it does for 2 alleles.

Expected Total branch length in tree, $L_k: 2(1 + 1/2 + 1/3 + \dots + 1/(k-1))$ ca= $2 \ln(k-1)$

6 Realisations with 25 leaves



Observations:

Variation great close to root.

Trees are unbalanced.

Sampling more sequences



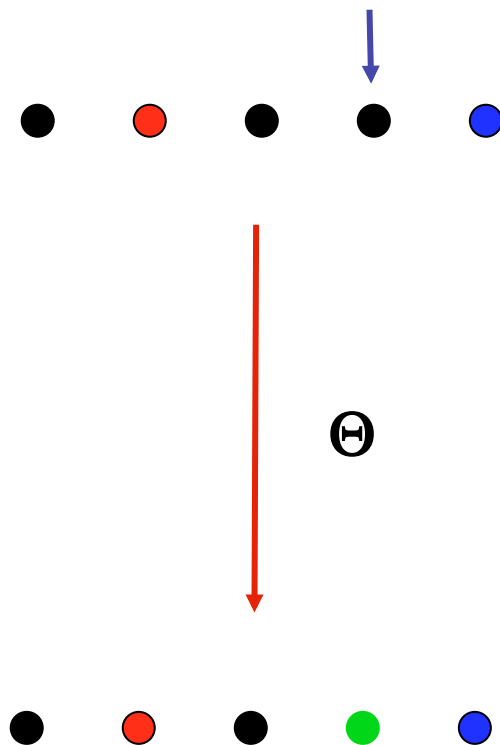
The probability that the ancestor of the sample of size n is in a sub-sample of size k is

$$\frac{(n+1)(k-1)}{(n-1)(k+1)}$$

Letting n go to infinity gives $(k-1)/(k+1)$, i.e. even for quite small samples it is quite large.

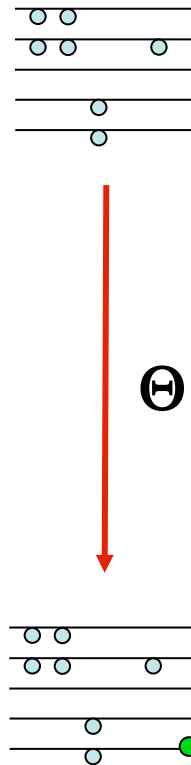
Three Models of Alleles and Mutations.

Infinite Allele



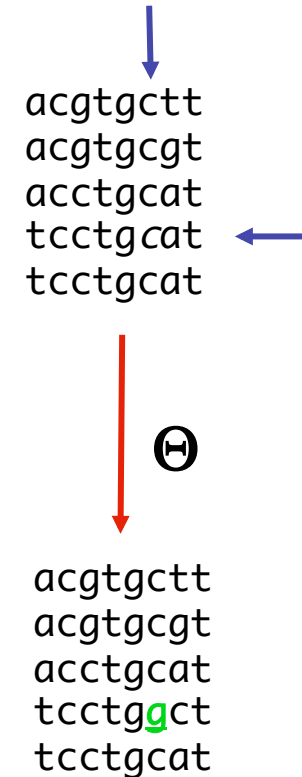
- i. Only identity, non-identity is determinable
- ii. A mutation creates a new type.

Infinite Site



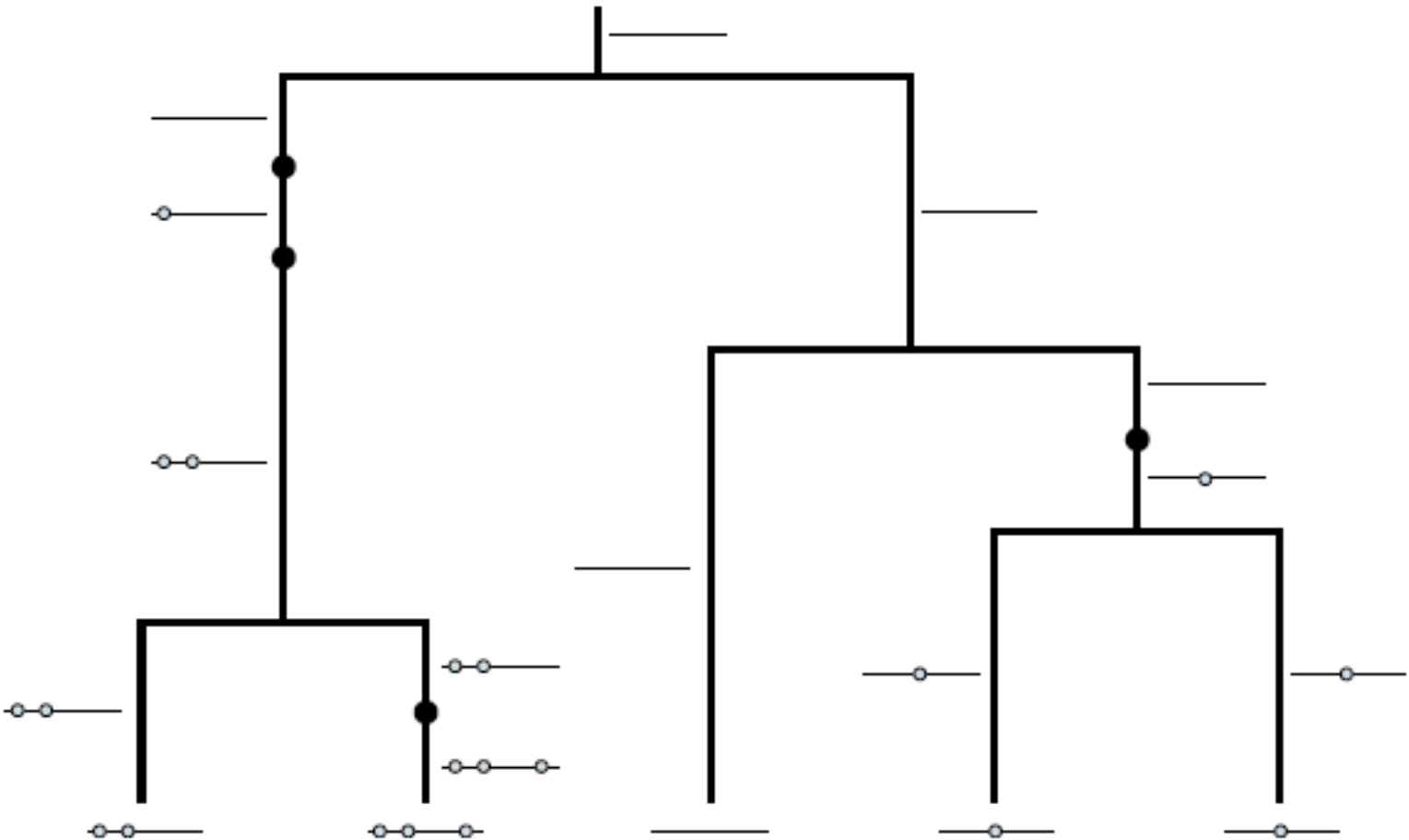
- i. Allele is represented by a line.
- ii. A mutation always hits a new position.

Finite Site

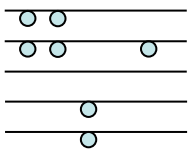


- i. Allele is represented by a sequence.
- ii. A mutation changes nucleotide at chosen position.

Infinite Site Model



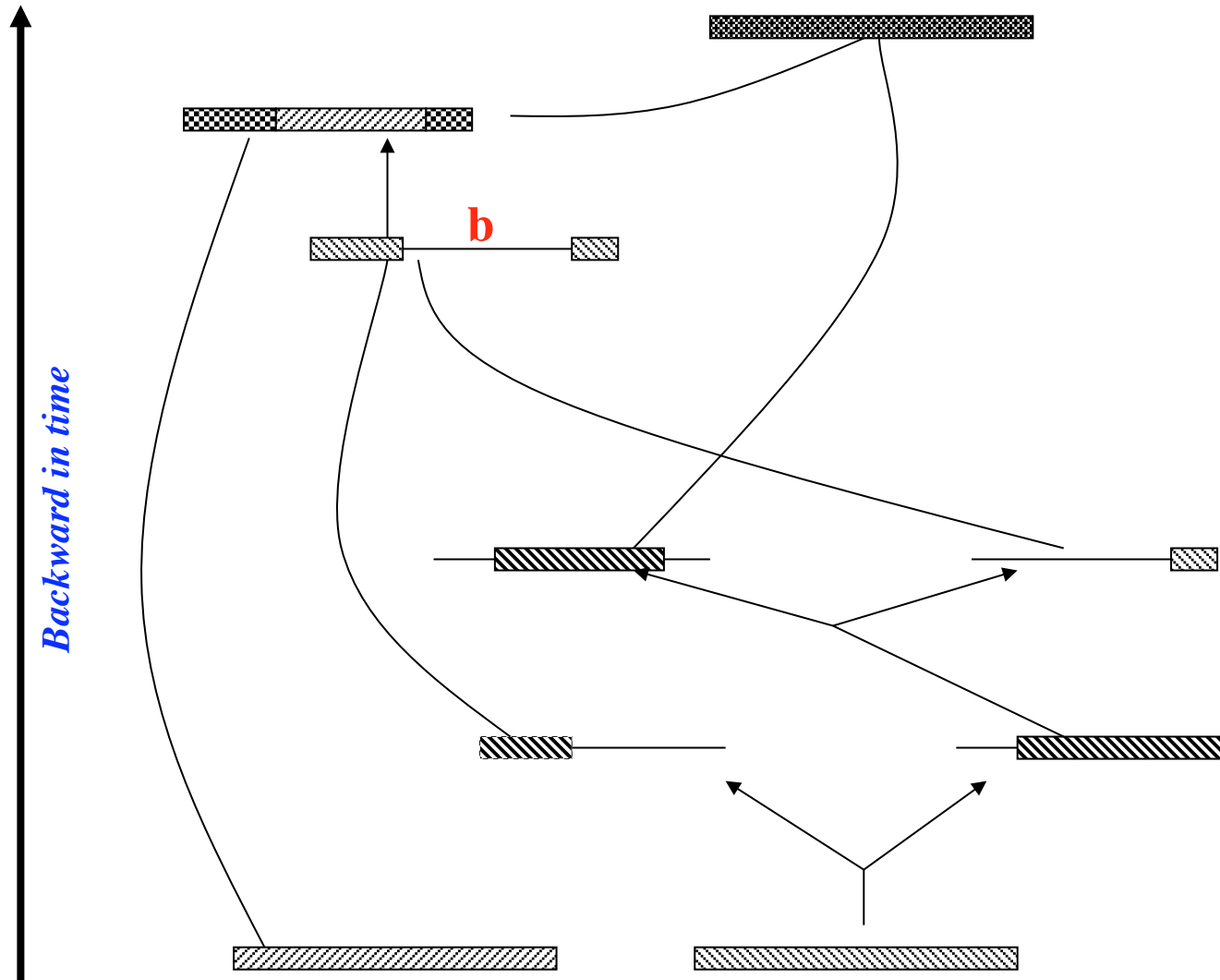
Final Aligned Data Set:



Recombination-Coalescence Illustration

Copied from Hudson 1991

Intensities



Coales.

Recomb.

0

ρ

1

$(1+b)\rho$

3

$(2+b)\rho$

6

2ρ

3

2ρ

1

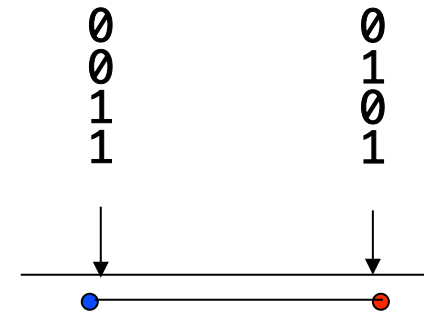
2ρ

Local Inference of Recombinations

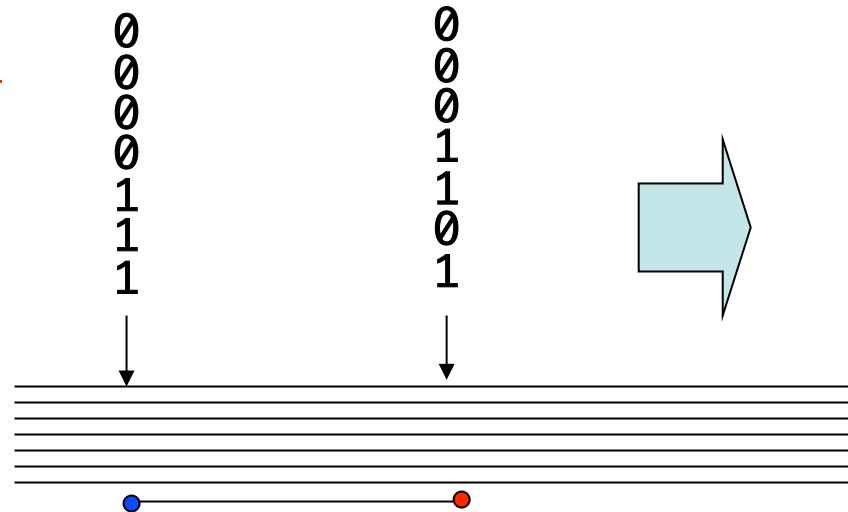
T . . . G
 T . . . C
 A . . . G
 A . . . C

Recoding

- At most 1 mutation per column
- 0 ancestral state, 1 derived state



Incompatibility:



Four combinations

00
 10
 01
 11

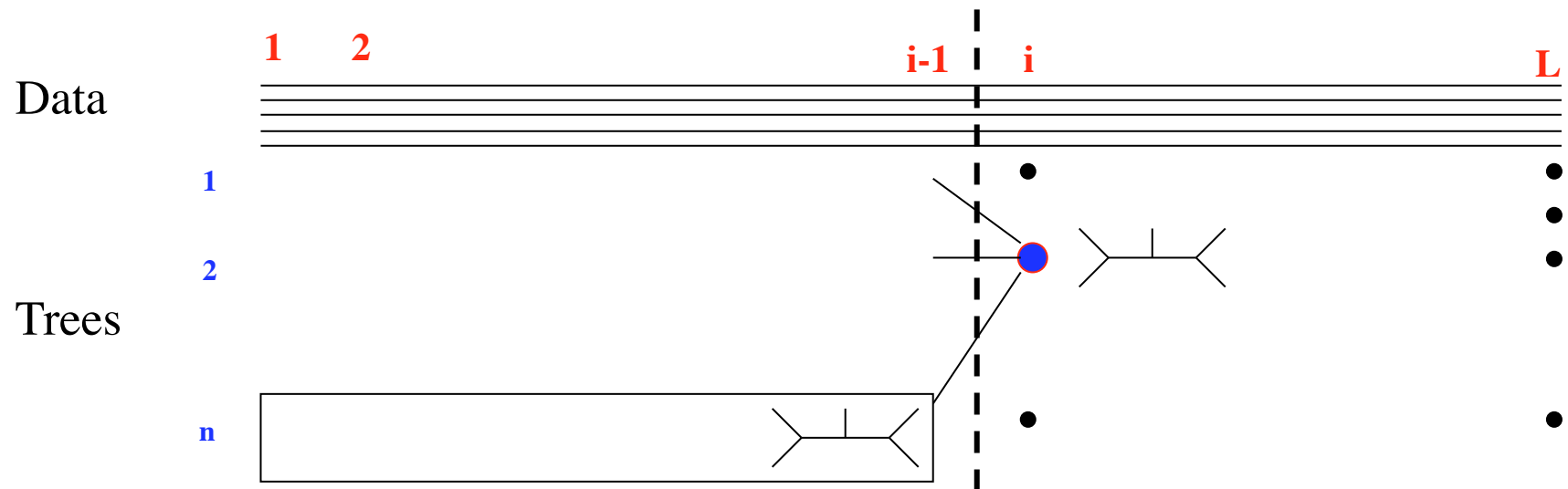
Myers-Griffiths (2002): Number of Recombinations in a sample, N_R , number of types, N_T , number of mutations, N_M obeys:

$$N_R \geq N_T - N_M - 1$$

Minimal Number of Recombinations

The Kreitman data (1983): 11 sequences, 3200bp, 43(28) recoded, 9 different

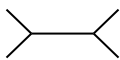
Last Local Tree Algorithm:




How many local trees?

How many neighbors?

Bi-partitions

• **Unrooted**  $\frac{(2n-2)!}{2^{n-1}(n-1)!}$

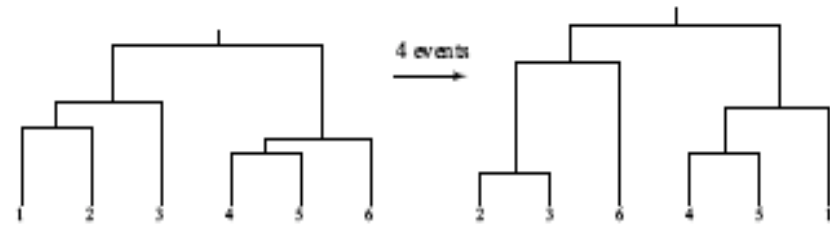
$$3n^2 - 13n + 14$$

• **Coalescent**  $\frac{n!(n-1)!}{2^{n-1}}$

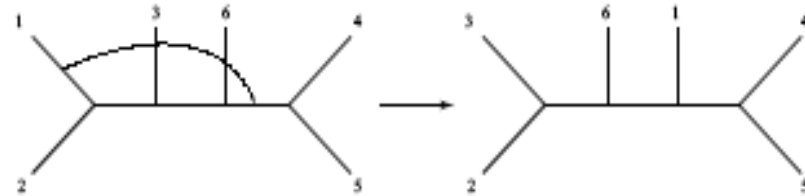
$$\sim n^3$$

Metrics on Trees based on subtree transfers.

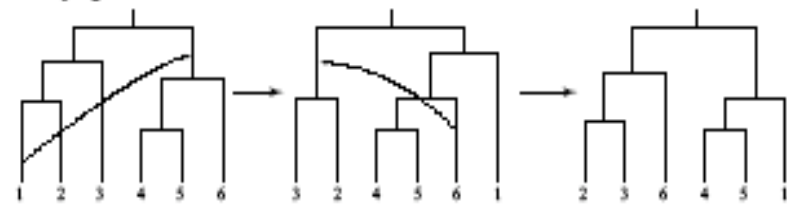
Trees including branch lengths



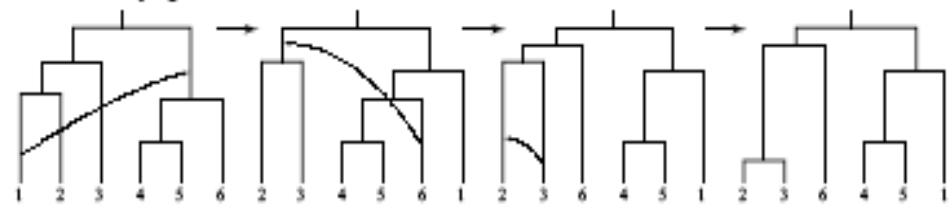
Unrooted tree topologies



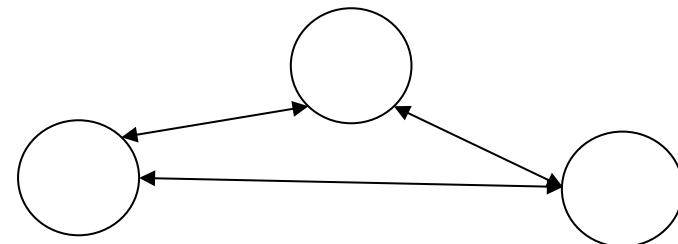
Rooted tree topologies



Tree topologies with age ordered internal nodes



Pretending the **easy** problem (unrooted) is the **real** problem (age ordered), causes violation of the triangle inequality:



Tree Combinatorics and Neighborhoods

Observe that the size of the unit-neighbourhood of a tree does not grow nearly as fast as the number of trees

$\delta(T)$:= number of trees one SPR operation away from a given tree T .

	Unrooted		Rooted			Dendrograms		
n	# of trees	δ	# of trees	δ_{\max}	δ_{\min}	# of trees	δ_{\max}	δ_{\min}
4	3	2	15	12	10	18	12	13
5	15	12	105	28	24	180	33	37
6	105	30	945	52	44	2,700	71	79
7	945	56	10,395	84	70	56,700	128	143
8	10,395	90	135,135	124	102	1,587,600	210	233
9	135,135	132	2,027,025	170	140	57,153,600	?	?
10	2,027,025	182	34,459,425	224	184	2,571,912,000	?	?

Due to Yun Song

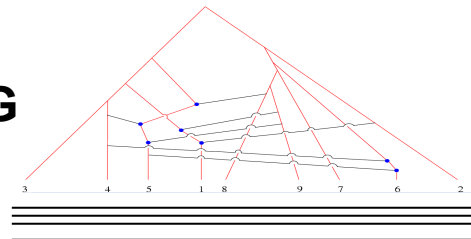
$$\begin{aligned}
 (2n-3)!! &= \frac{(2n-2)!}{2^{n-1}(n-1)!} & 3n^2 - 13n + 14 & \frac{n!(n-1)!}{2^{n-1}} & \frac{1}{3}(2n^3 - 3n^2 - 20n + 39) \\
 2(n-3)(2n-7) & 4(n-2)^2 - 2 \sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor & \frac{1}{6} \left\{ 4n^3 - 9n^2 - 13n + 42 - 3(2n+3) \left\lfloor \frac{n-1}{2} \right\rfloor + 9 \left(\left\lfloor \frac{n-1}{2} \right\rfloor \right)^2 \right\}
 \end{aligned}$$

Allen & Steel (2001)

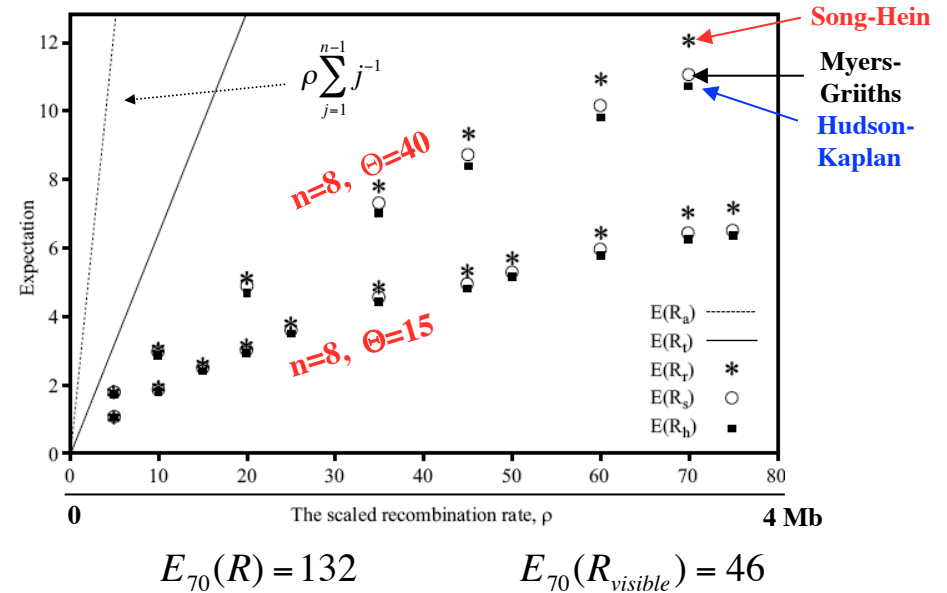
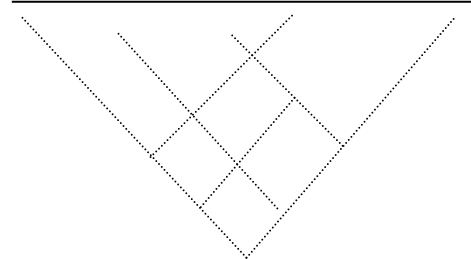
Song (2003+)

minARGs: Recombination Events & Local Trees

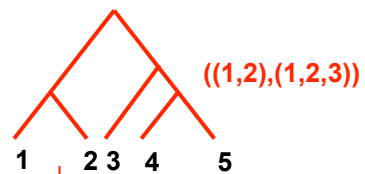
Minimal ARG



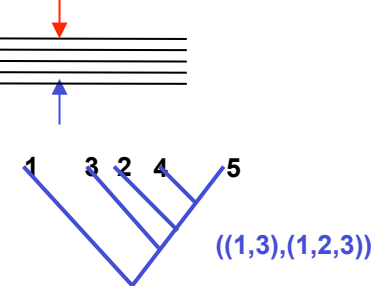
True ARG



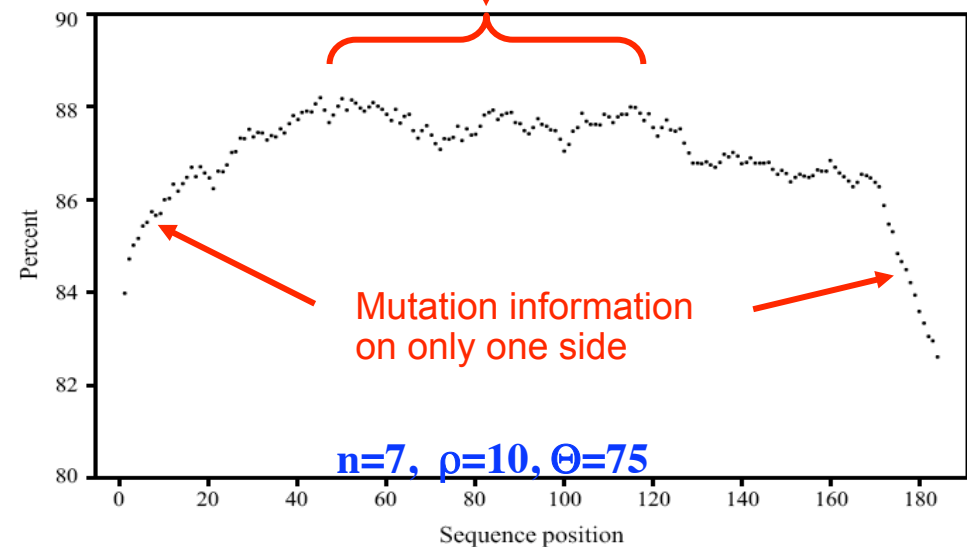
True ARG



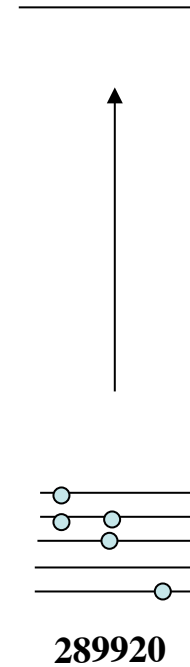
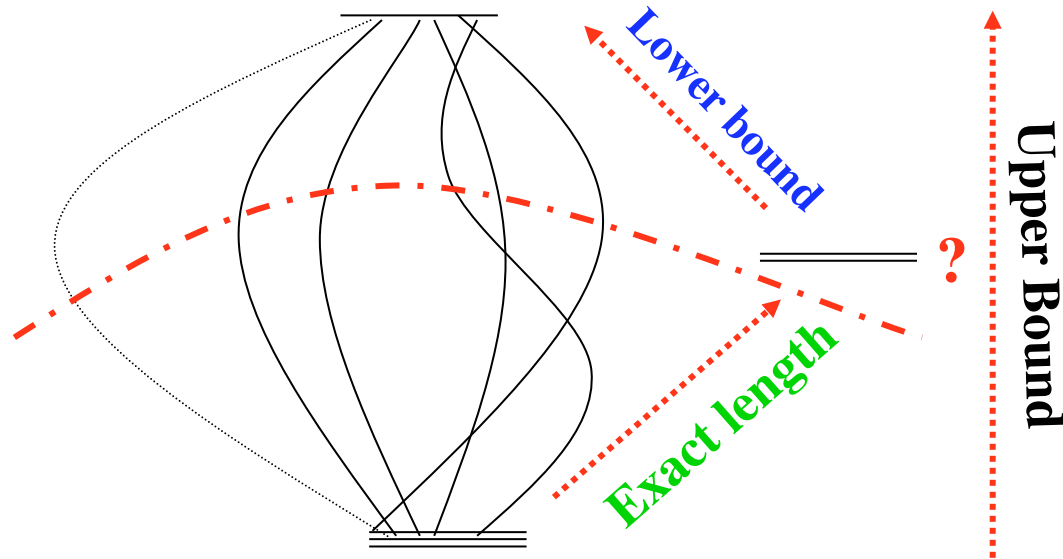
Reconstructed ARG



Mutation information on both sides

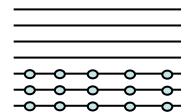


Counting + Branch and Bound Algorithm

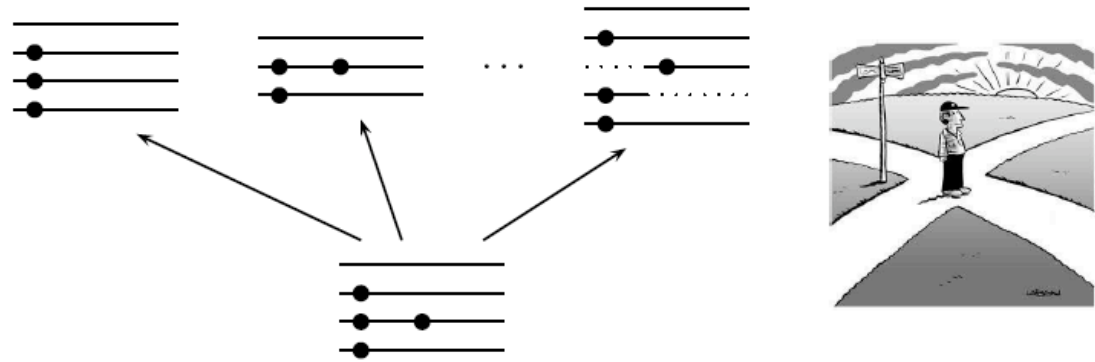


k	k-recombination neighborhood
0	3
1	91
2	1314
3	8618
4	30436
5	62794
6	78970
7	63049
8	32451
9	10467
10	1727

n	Number of segregating sites			
	2	3	4	5
2	30	573	16 875	689 175
3	108	6 286	743 387	149 861 079
4	330	62 589	32 482 009	35 523 729 489
5	866	445 137	893 479 326	4 938 627 635 669
6	2 143	3 302 506	29 521 615 942	962 962 451 049 968
7	4 611	17 409 443	568 860 072 916	91 812 561 254 804 105
8	9 728	98 432 218	13 273 296 248 617	
9	18 378	420 106 717	195 515 335 378 914	
10	34 552	1 917 604 869		
11	59 577	6 985 275 356		



BB & Heuristic minimal ancestral recombination graphs



Beagle

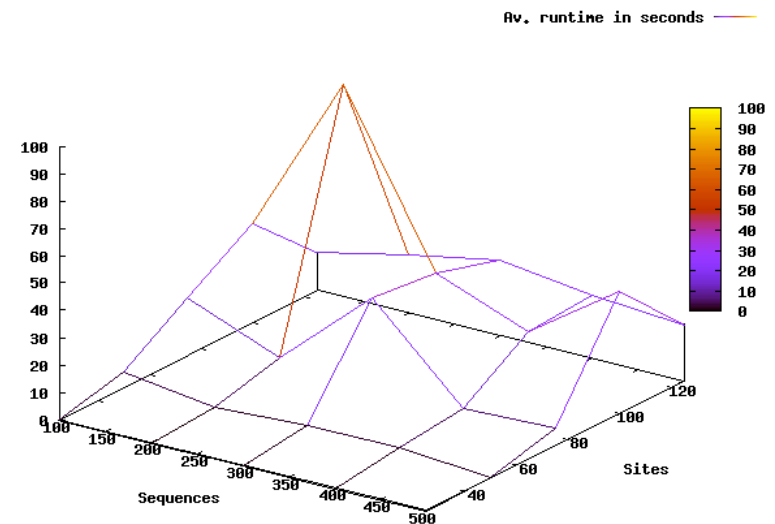
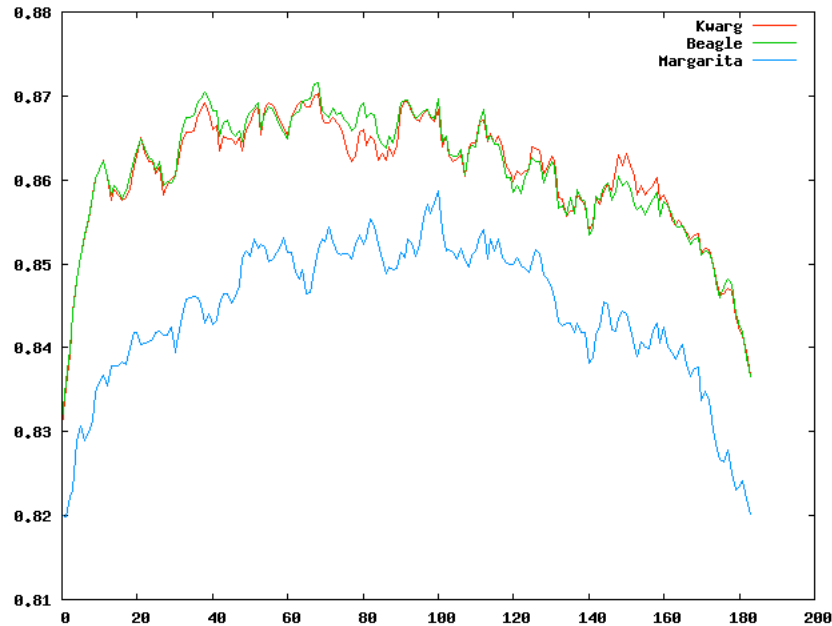
Try each in turn
until shortest route
is determined

Margarita

Just follow road
seeming to lead in
the right direction

Kwarg

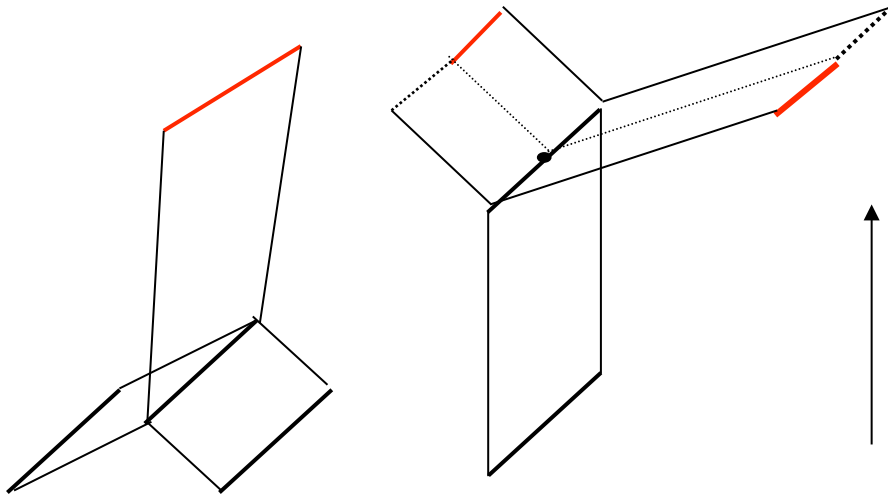
Choice based on
location of next
crossroads



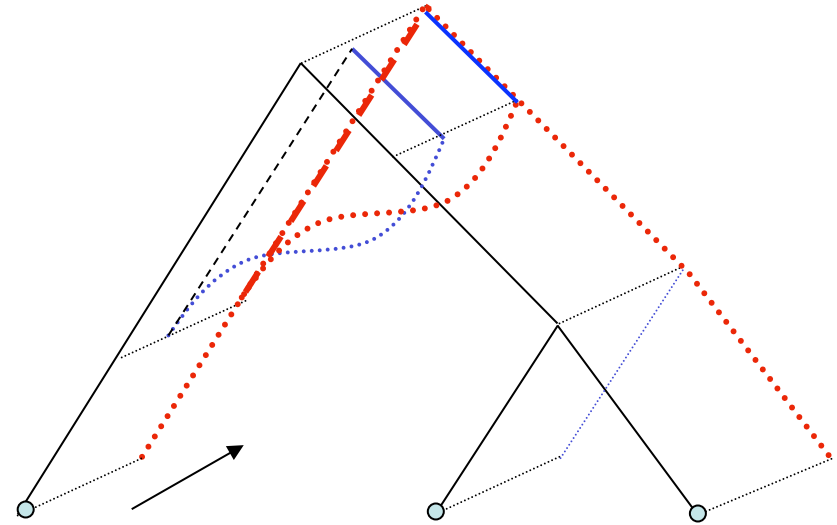
Time versus Spatial 1: Coalescent-Recombination

(Griffiths, 1981; Hudson, 1983 - Wiuf & Hein, 1999)

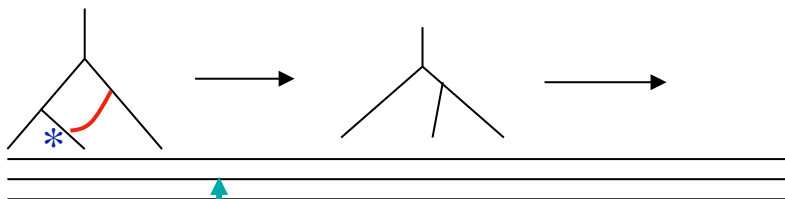
Temporal Process



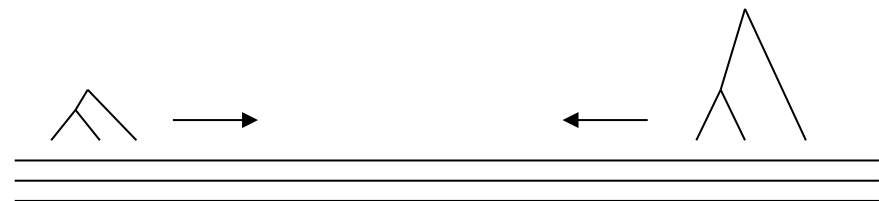
Spatial Process



i. The process is non-Markovian

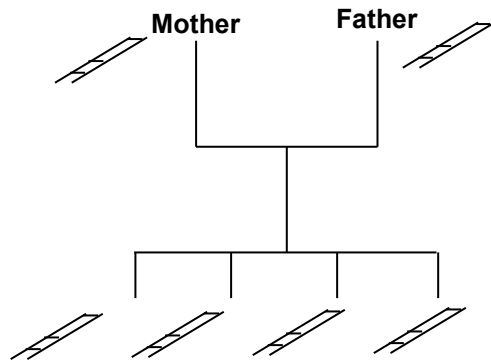


ii. The trees cannot be reduced to Topologies



Time versus Spatial 2: Pedigrees

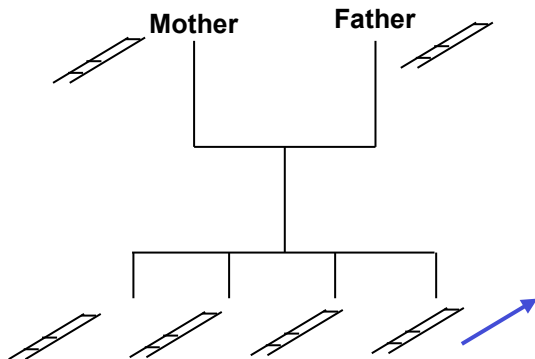
Elston-Stewart (1971) - Temporal Peeling Algorithm:



Condition on parental states

Recombination and mutation are Markovian

Lander-Green (1987) - Genotype Scanning Algorithm:



Condition on paternal/maternal inheritance

Recombination and mutation are Markovian

Time versus Spatial 3: Phylogenetic Alignment

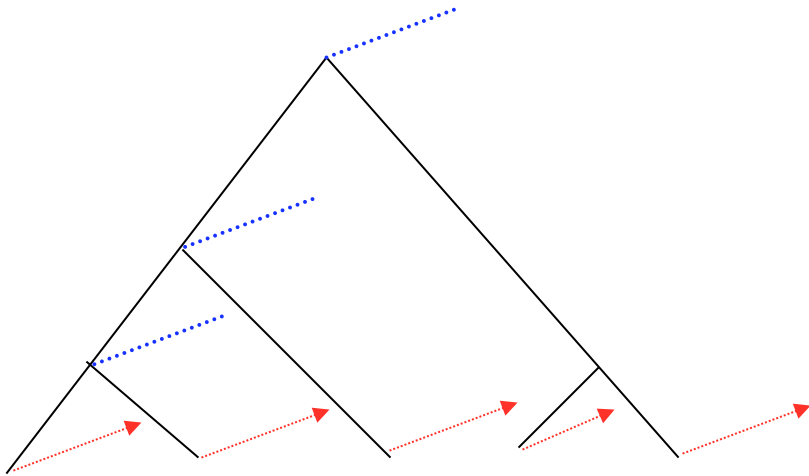
- **Optimisation Algorithms**

 - indels of length 1 (David Sankoff, 1973) **Spatial**

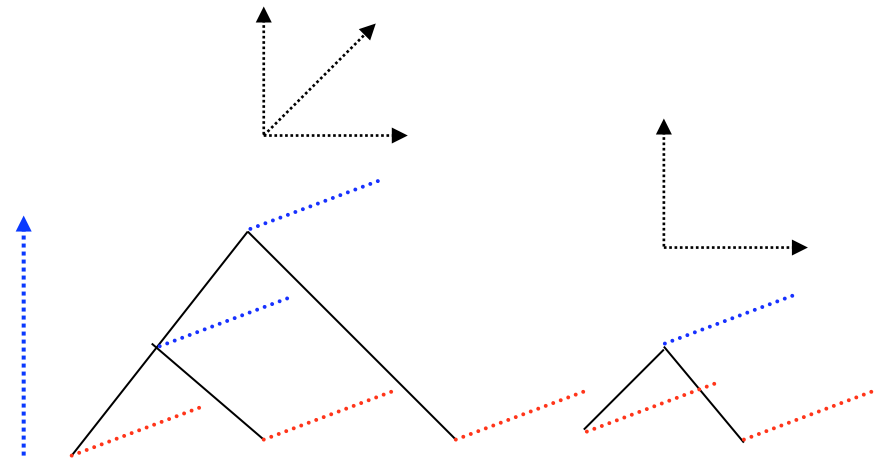
 - indels of length k (Bjarne Knudsen, 2003) **Temporal**

- **Statistical Alignment**

Spatial:



Temporal:

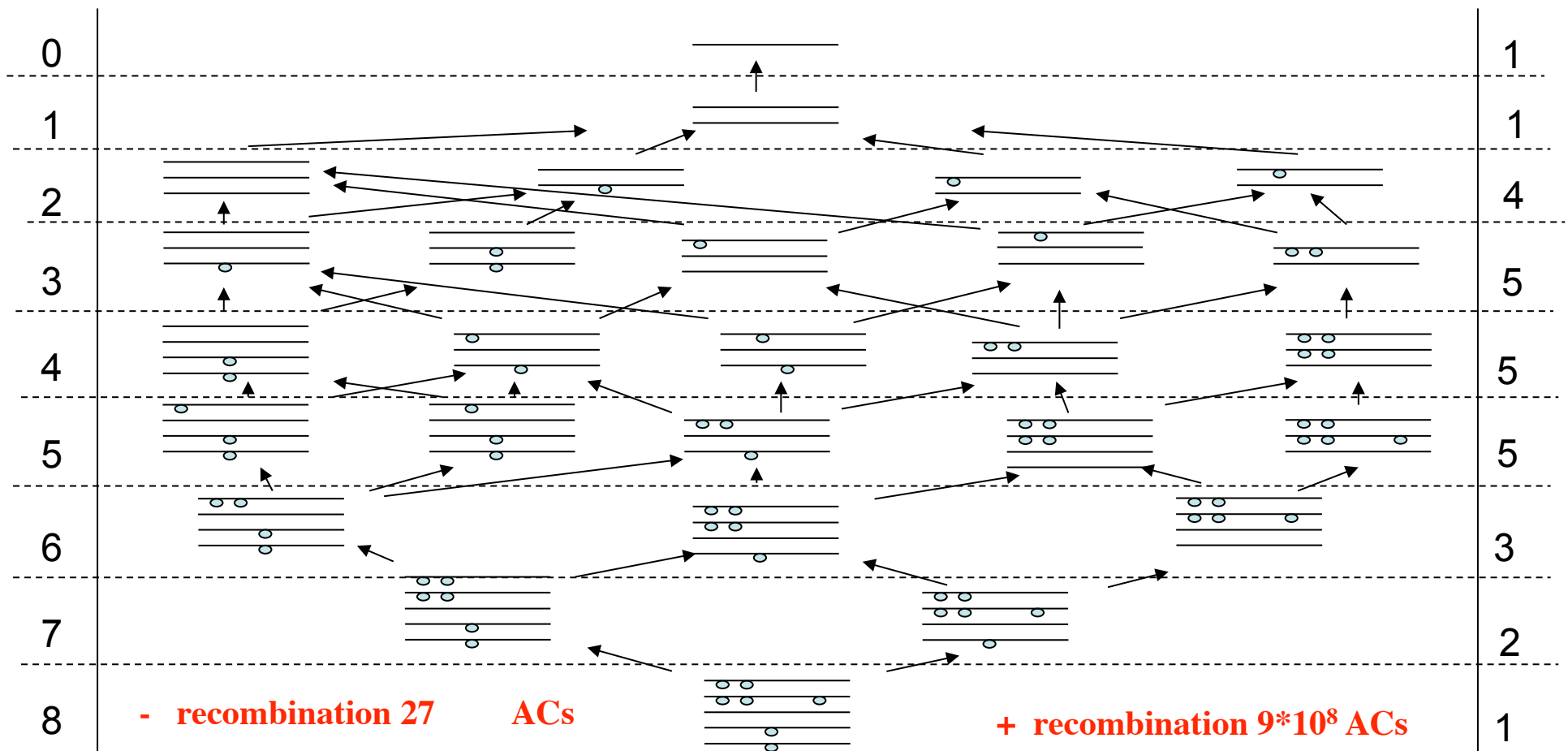


The Griffiths-Ethier-Tavare Recursions

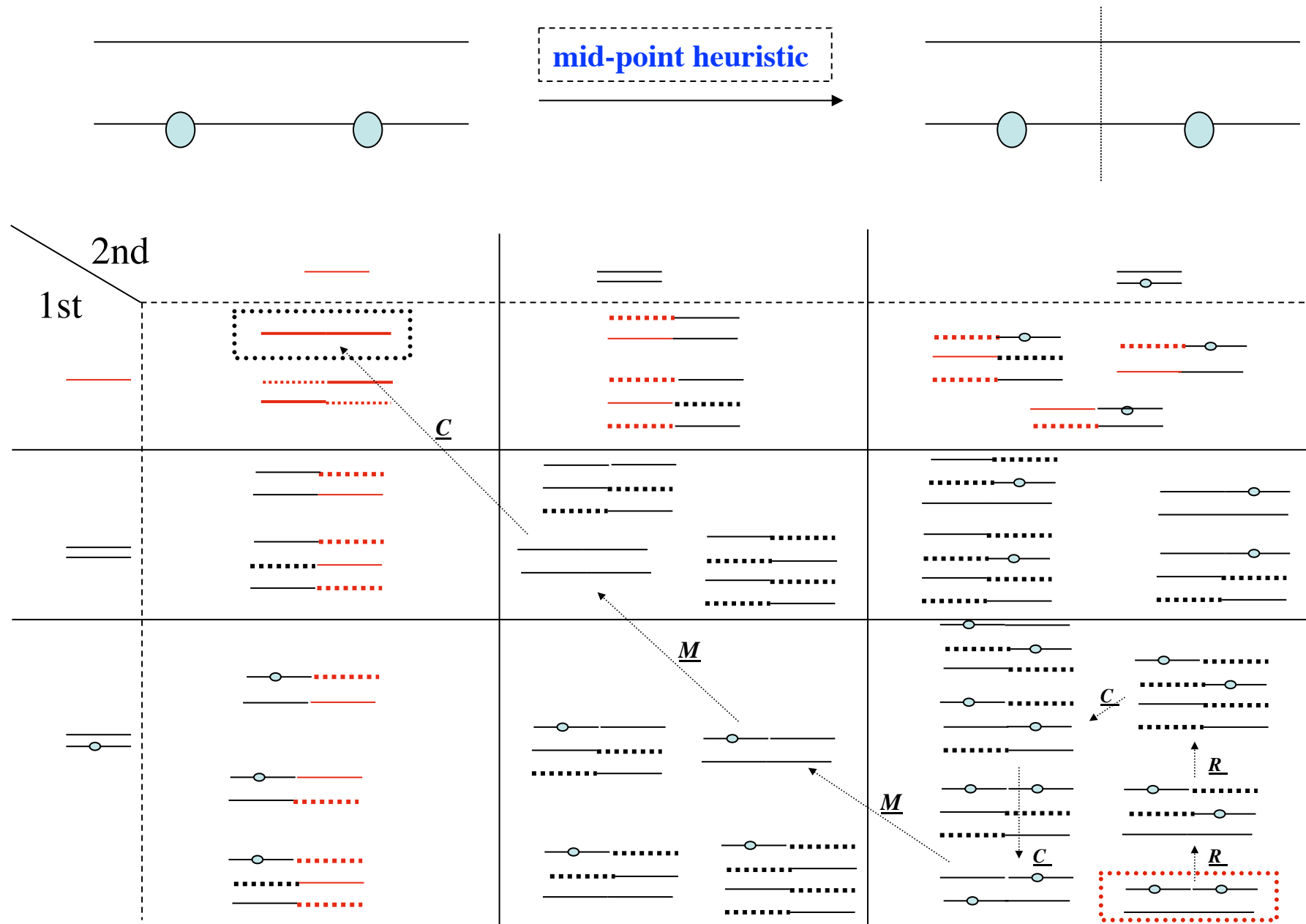
No recombination: Infinite Site Assumption
Ancestral State Known

History Graph: Recursions Exists
No cycles

Possible Histories without Recombination for simple data example



Ancestral configurations to 2 sequences with 2 segregating sites

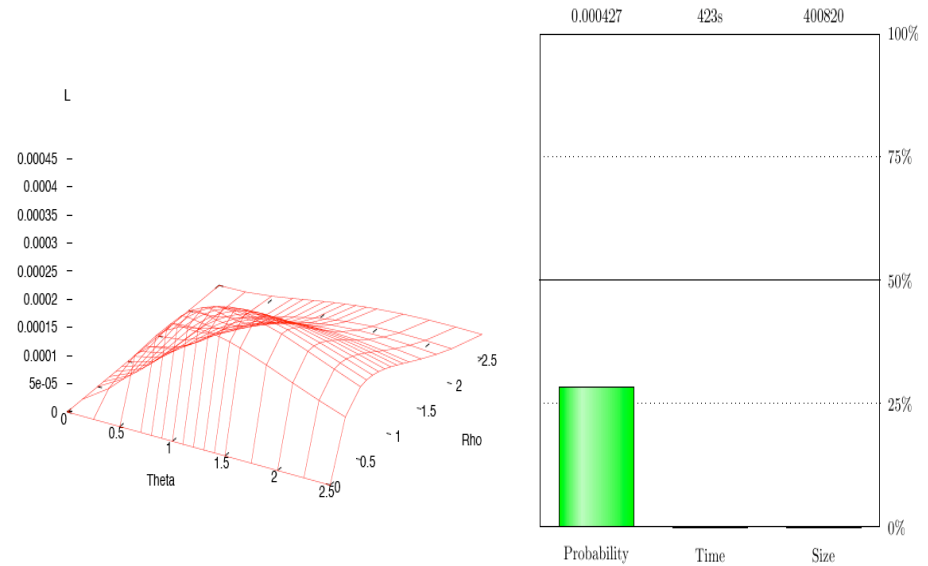


Likelihood Calculations on the ϵ -ARG

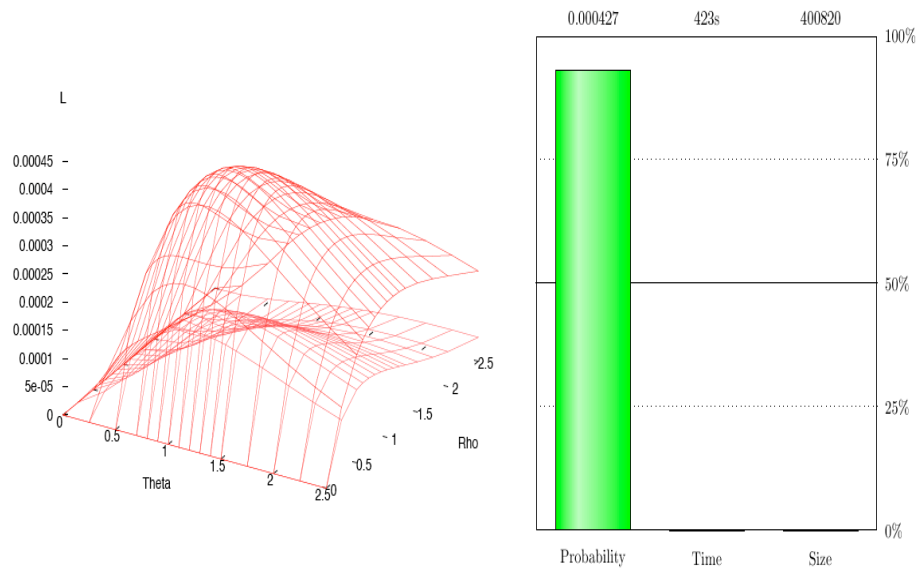
Example:

010
010
101
101
110

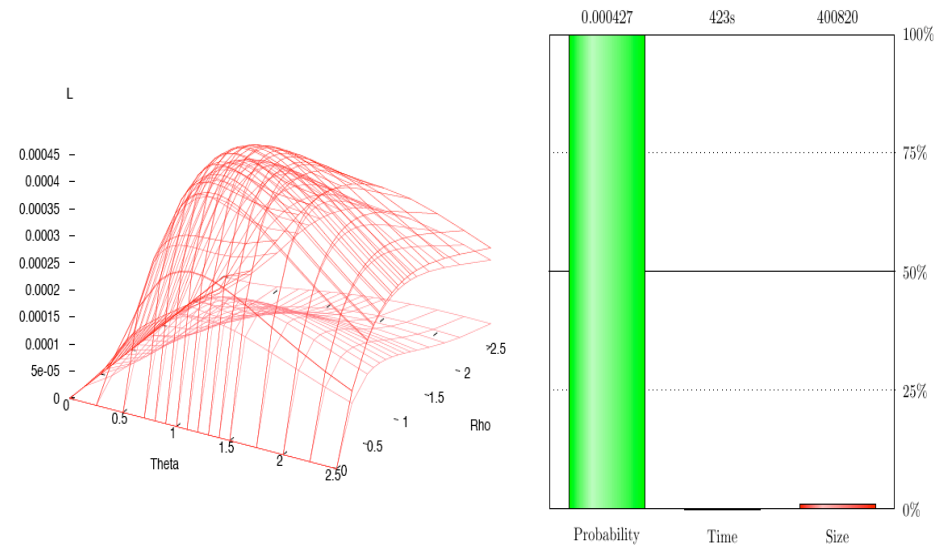
0-ARG



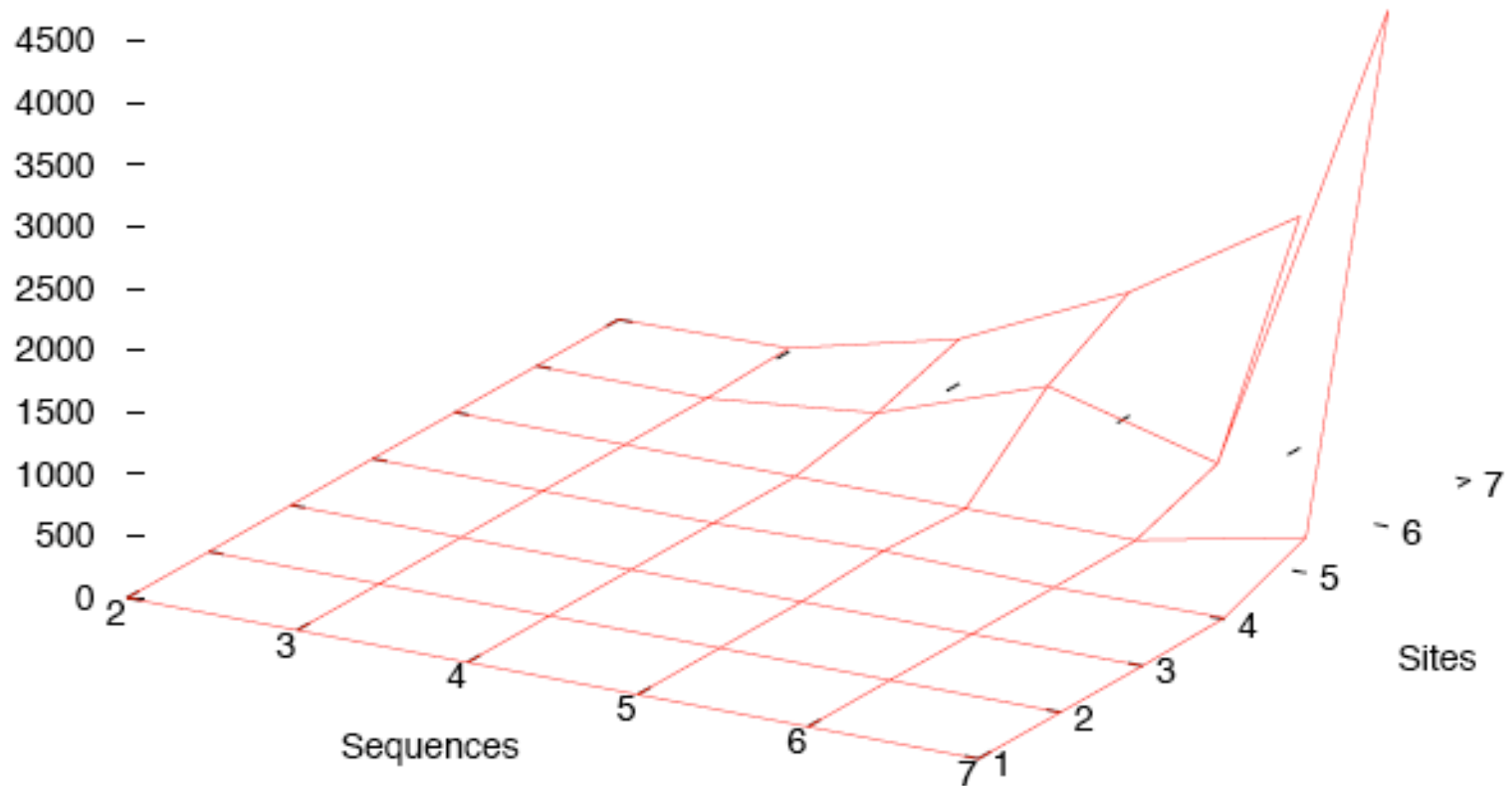
1-ARG



2-ARG



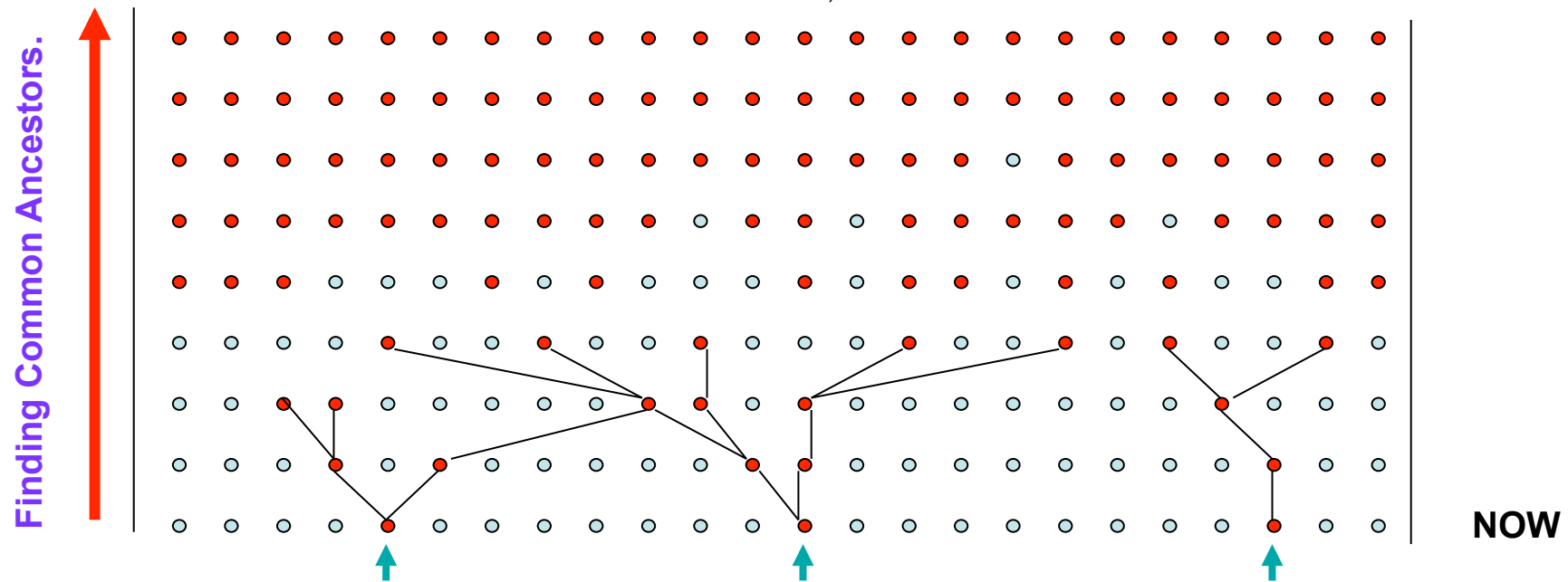
ϵ -ARG likelihood calculation



- *For sites=7, sequences=6, then 10^{-7} th of states visited*
- *Cut-off ϵ -ARG, $\epsilon=2$*

Combining Ancestral Individuals and the Coalescent

Wiuf & Hein, 2000.



Let T be the time, when somebody was everybody's ancestor.

Changs' result: $\lim T^*/\log_2(N) = 1$ prob. 1

Unify the two processes:

I. Sample more individuals II. Let each have 2 parents with probability p .

Result: A discontinuity at 1.

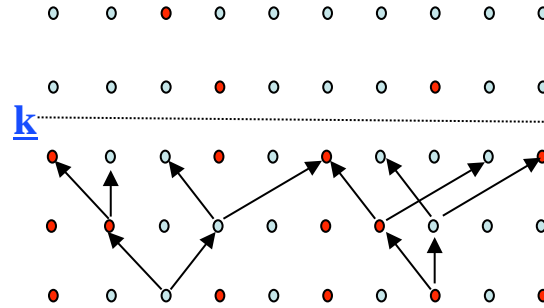
For $p < 1$ change $\log_2 \rightarrow \log_p$

Comment: Genetic Ancestors is a vanishing set within Genealogical Ancestors.

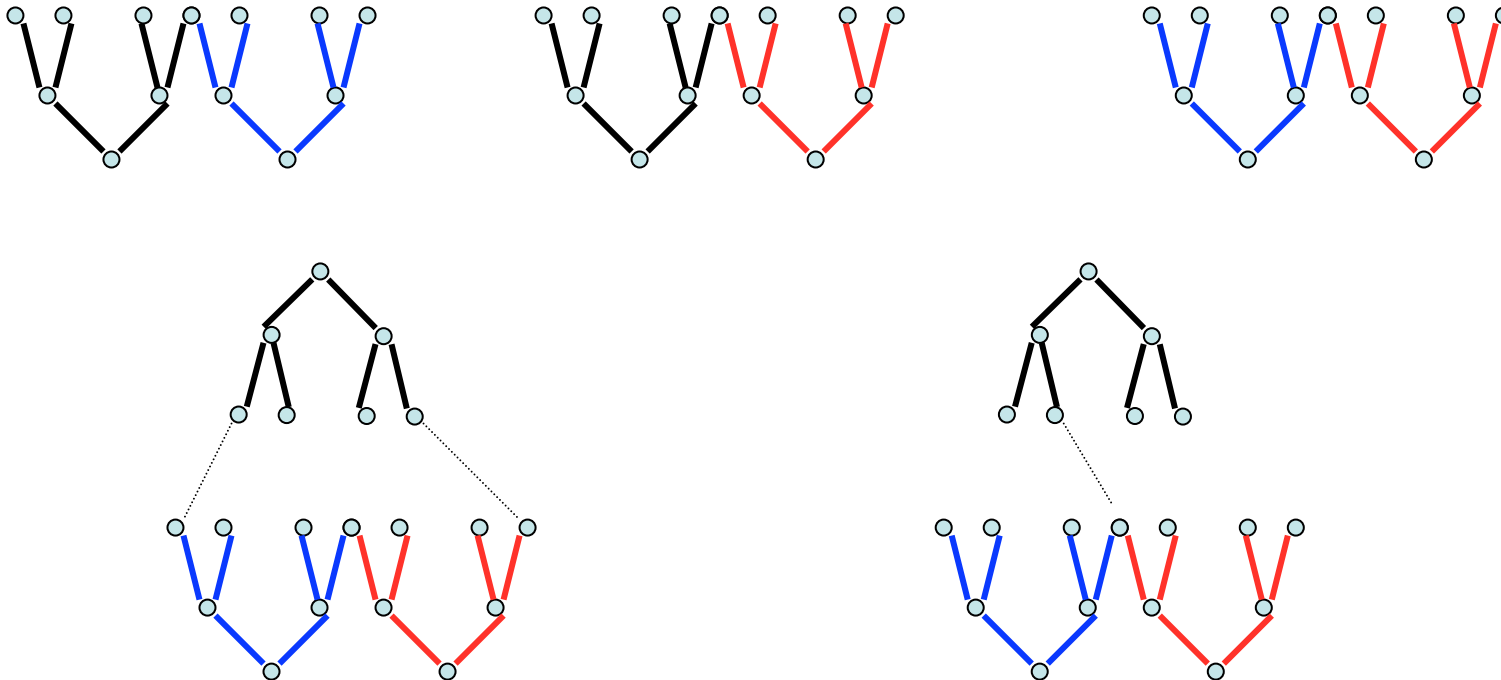
Reconstructing global pedigrees: Superpedigrees

Steel and Hein, 2006

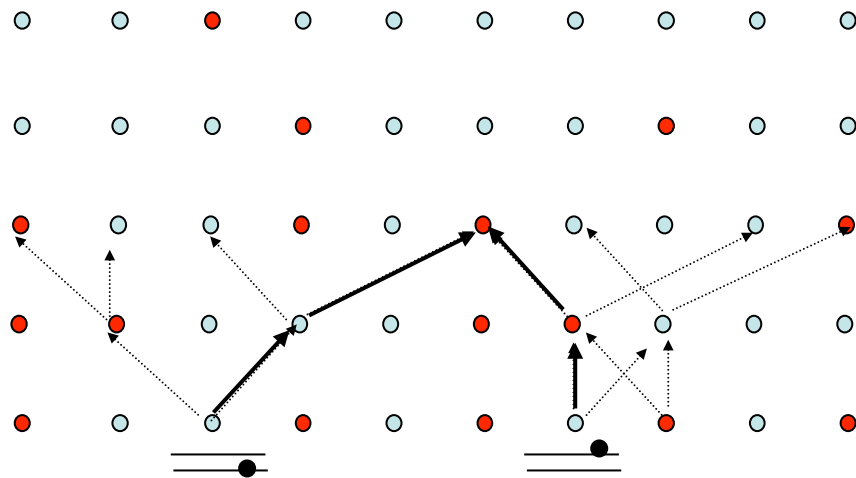
The gender-labeled pedigrees for all pairs defines global pedigree



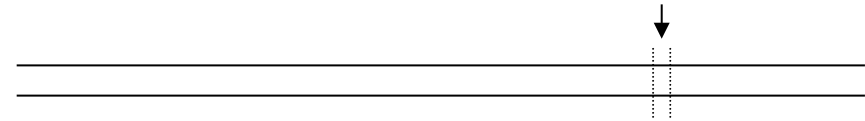
Gender-unlabeled pedigrees don't!!



Benevolent Mutation and Recombination Process

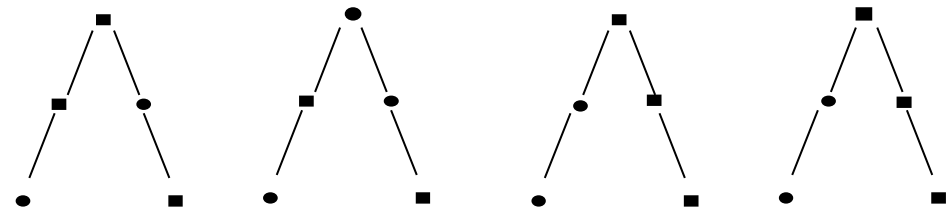
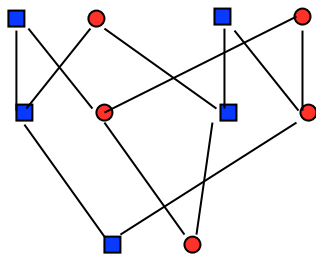
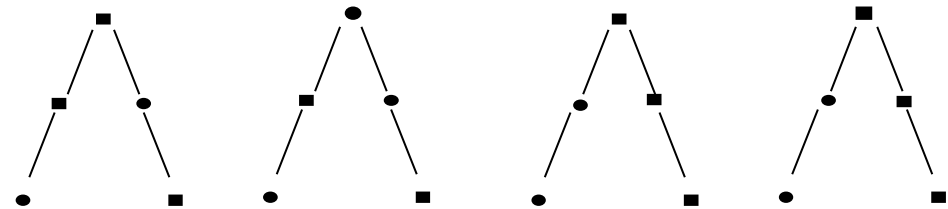
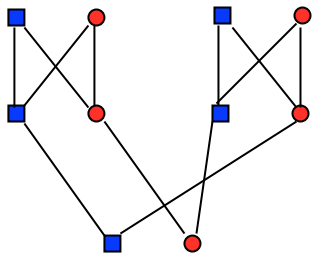


Genomes with ρ and $\mu/\rho \rightarrow \infty$
 ρ – recombination rate, μ – mutation rate



- All embedded phylogenies are observable
- Do they determine the pedigree?

Counter example:



Embedded phylogenies:

Infinite Sequences: From ARG to Pedigree

What can you observe from data (infinite sequences)?

A. The ARG?

B. Sequence of neighbor pairs of local trees with recombination points

Going to neighbor triples, quadruples,..., be more restrictive than pairs?

C. Sequence of local trees?

D. Set of local trees?

E. Set/Sequences of local unrooted tree topologies?

F. Set/Sequences of local bipartitions? (neighbor pairs...)

Given A/B/C/D/E above how much does that constrain set of pedigrees

*How **many** pedigrees are compatible with A/B/C/D/E varying over **data**?*

Infinite Sequences: From ARG to Pedigree

What can you observe from data (infinite sequences)?

A. The ARG?



B. Sequence of pairs of local trees with recombination points

Going to triples, quadruples,..., be more restrictive than pairs?

C. Sequence of local trees?

D. Set of local trees?

E. Set/Sequences of local unrooted tree topologies?

F. Set/Sequences of bipartitions

