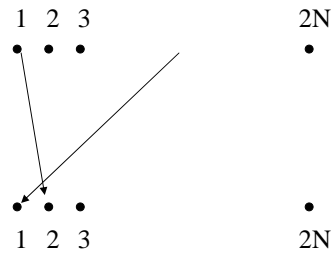


Haploid Reproduction Model (i.e. no recombination)

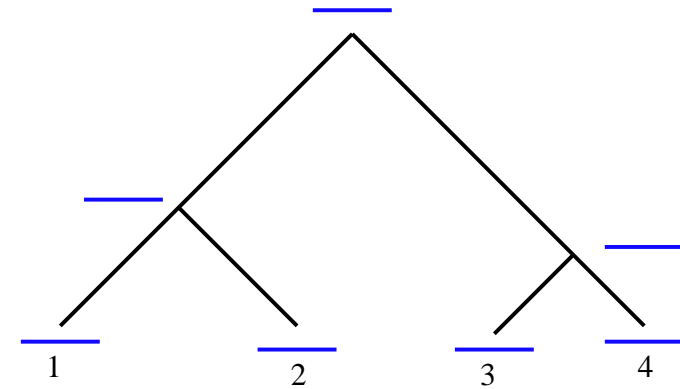


Individuals are made by sampling with replacement in the previous generation.

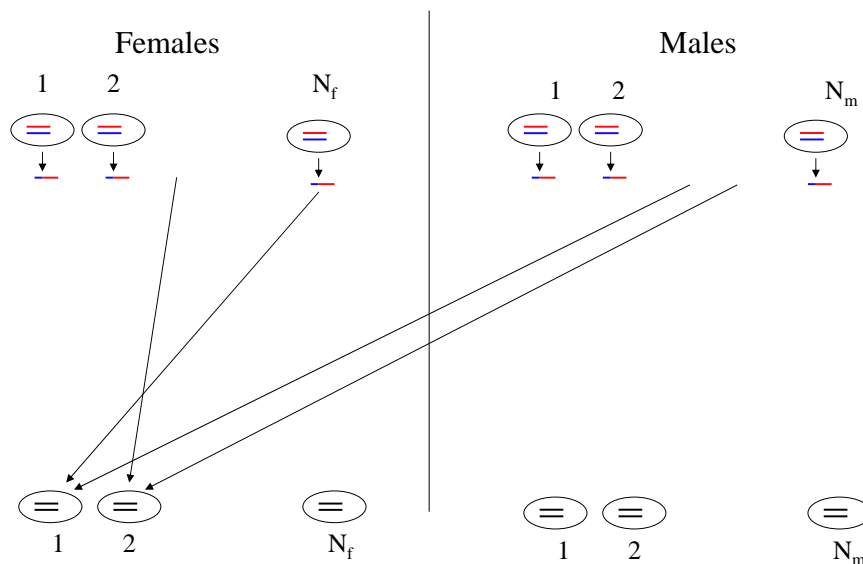
The probability that 2 alleles have same ancestor in previous generation is $1/2N$.

The probability that k alleles have less than $k-1$ ancestors in previous generation is vanishing.

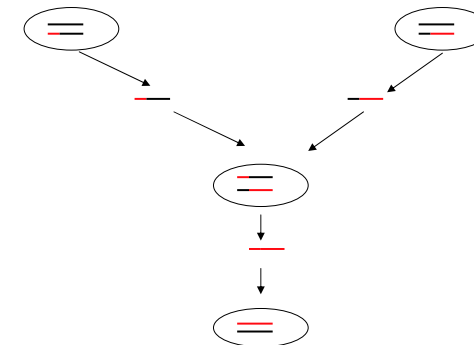
0 recombinations implies traditional phylogeny



Diploid Model with Recombination



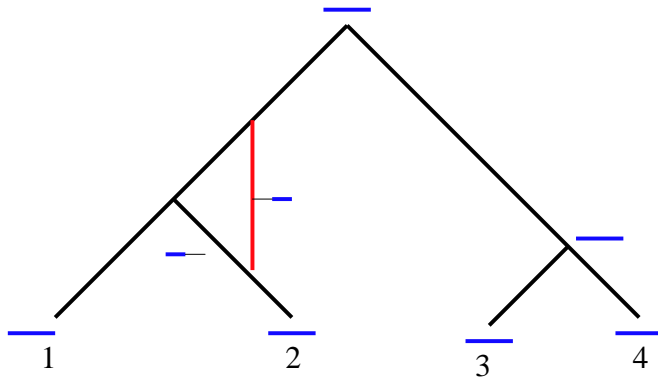
The Diploid Model Back in Time.



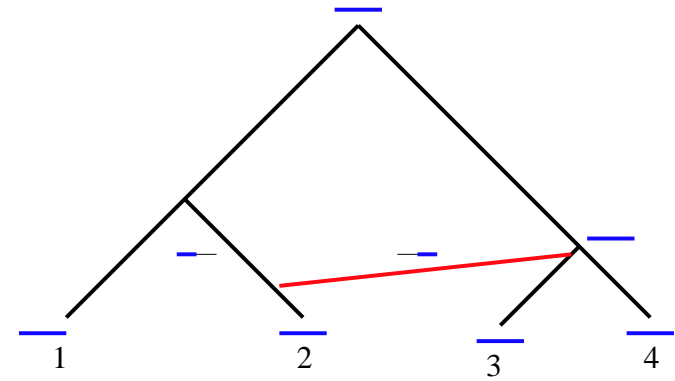
A recombinant sequence will have two different ancestor sequences in the grandparent.

1- recombination histories I:

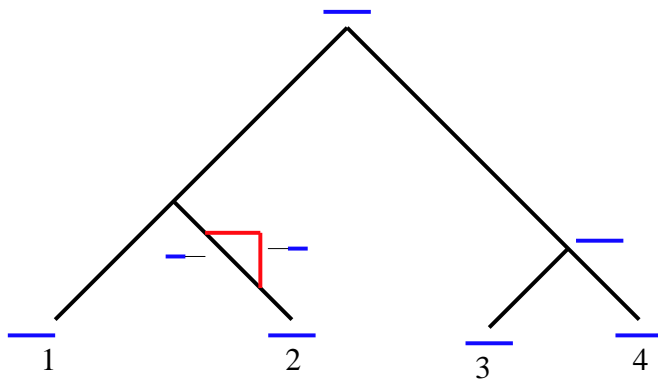
Branch length change



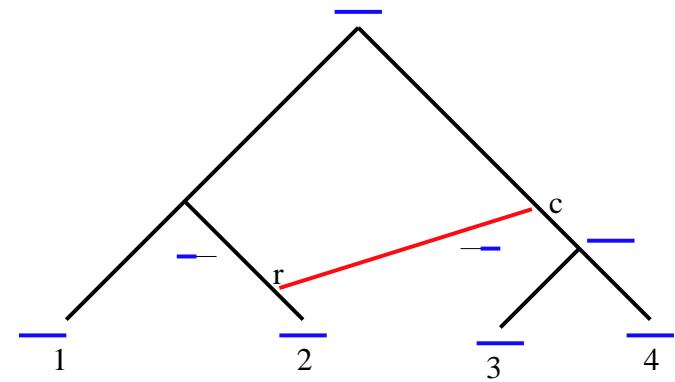
1- recombination histories II: Topology change



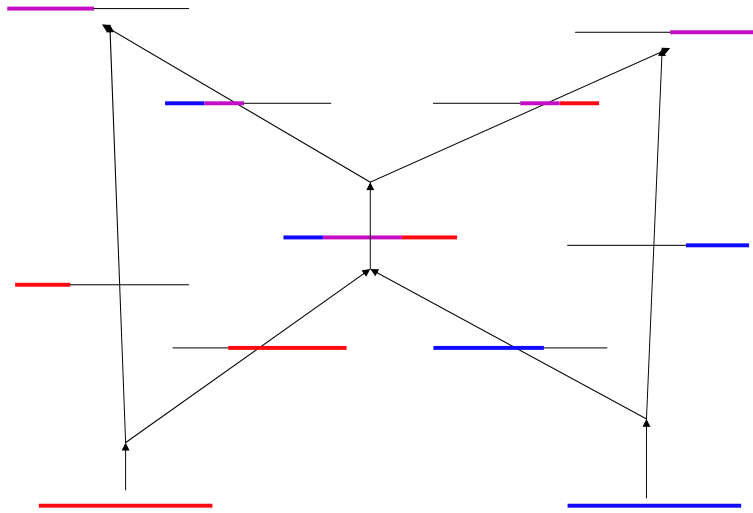
1- recombination histories III: Same tree



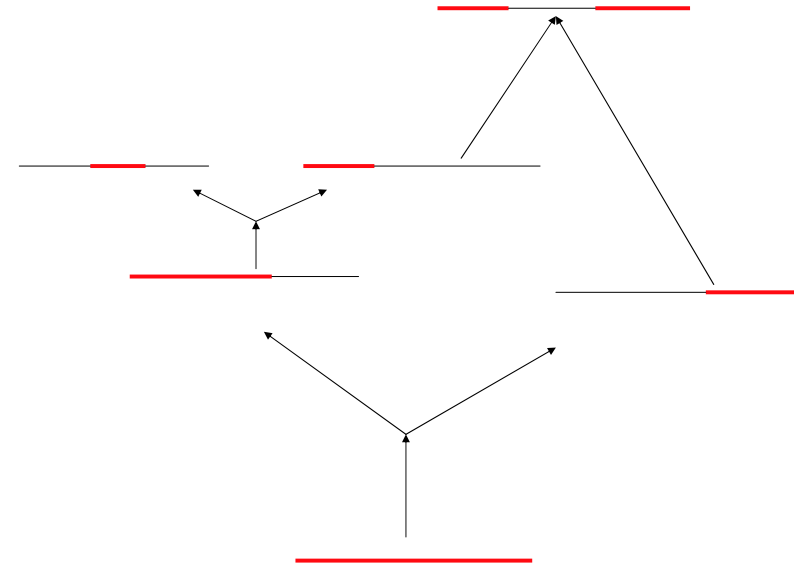
1- recombination histories IV: Coalescent time must be further back in time than recombination time.



Recombination Histories V: Multiple Ancestries.



Recombination Histories VI: Non-ancestral bridges



Summarising new phenomena in recombination-phylogenies

Consequence of 1 recombination

Branch length change

Topology change

No change

Time ranking of internal nodes

Multiple Ancestries

Non-ancestral bridges

What is the probability of different histories?

Coalescence +Recombination (Hudson(1983))

r = probability for a recombination within a dinucleotide pr. generation.

$\rho = r * (L-1) * 2N =$ Expected number of recombinations/(gene*2N generations).

1. Waiting time backward until first recombination is $\text{expo}(\rho)$ distributed.

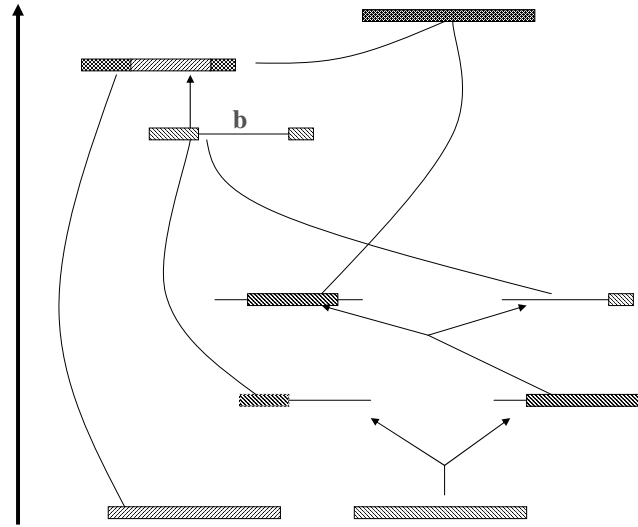
ex. gene 1000 bp $r = 10^{-8}$, $N = 10^4$, generation span 30 years.

Waiting time for a recombination/coalescence: $10^5/2 * 10^4$ generations.

2. The position will be chosen uniformly on the gene.

Recombination-Coalescence Illustration

Copied from Hudson 1991

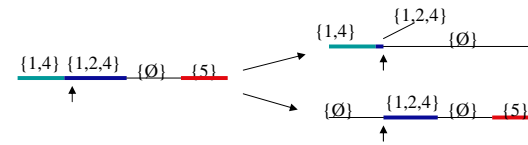


Intensities	
Coales.	Recomb.
0	ρ
1	$(1+b)\rho$
3	$(2+b)\rho$
6	2ρ
3	2ρ
1	2ρ

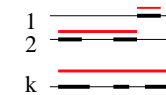
Back-in-Time Process

Two kinds of operations on sequence sets going backward in time. Each sequence consists of **intervals** and each interval is **labelled with subsets of $\{1, \dots, k\}$** - possibly the empty set.

A **recombination** takes one sequence and a position and generates two sequences:



Example:

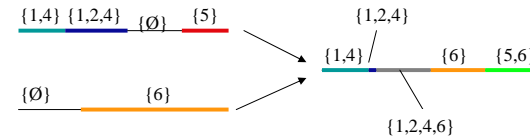


Rates:

Coalescent: $\binom{k}{2}$

Recombination: $\rho * \text{length of red}$

Coalescent:



Grand Most Recent Common Ancestor: GMRCA

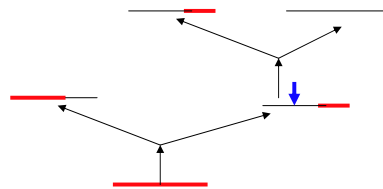
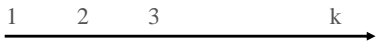
(griffiths & marjoram, 96)

- Track all sequences including those that have lost all ancestral material.
- The G-ARG contains the ARG. The graph is too large, but the process is simpler.

Sequence number - k .

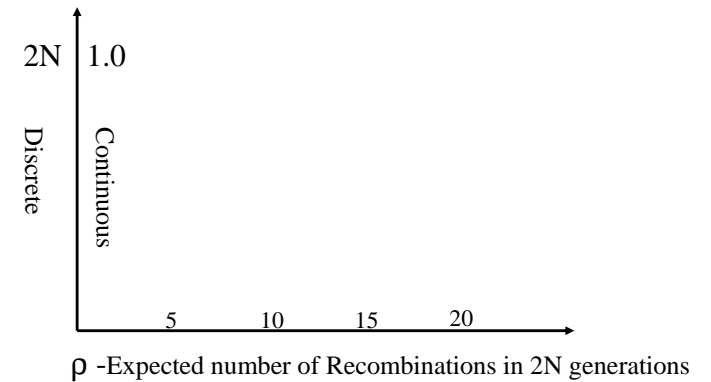
Birth rate: $\rho * k/2$

Death rate: $\binom{k}{2}$



$$E(\text{events until } \{1\}) = (\text{asymp.}) \exp(\rho) + \rho \log(n)$$

Parametrization



H.sapiens - Kbases

20

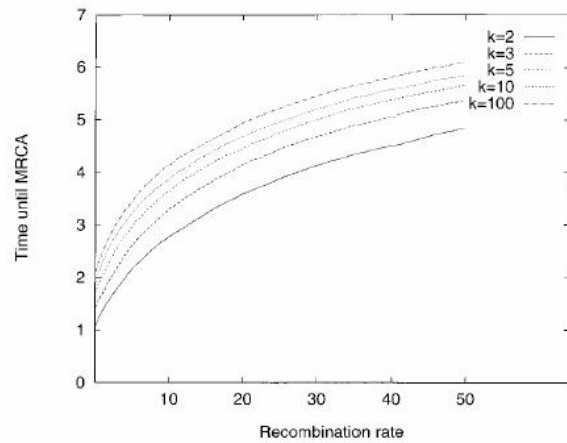
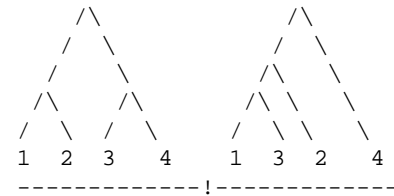


FIGURE 10.—The expected time until all positions have found a MRCA. This expected time becomes quickly independent of sample size: The difference between sample size 25 and sample size 100 is <2%. For $\rho = 0$, the time until a MRCA is distributed according to the coalescent process and the expectation is $2(1 - 1/k)$, k denoting sample size.

Properties of Neighboring Trees.

(partially from Hudson & Kaplan 1985)

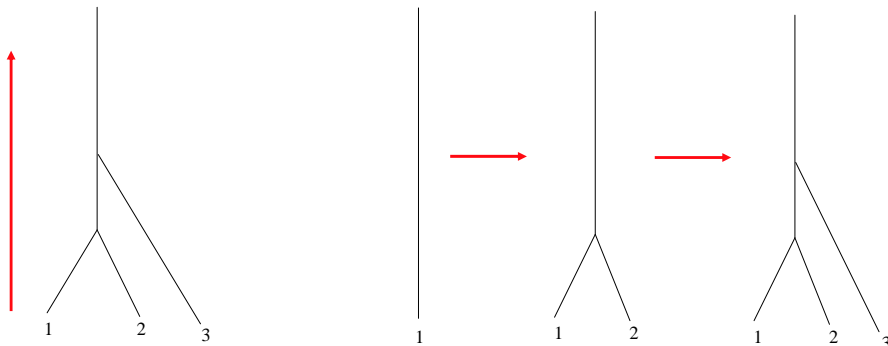


Leaves	Root	Edge-Length	Topo-Diff	Tree-Diff
2	1.0	2.0	0.0	.666
3	1.33	3.0	0.0	.694
4	1.50	3.66	0.073	.714
5	1.60	4.16	0.134	.728
6	1.66	4.57	0.183	.740
10	1.80	5.66	0.300	.769
15	1.87	6.50	0.374	.790
500	1.99		0.670	

Old + Alternative Coalescent Algorithm

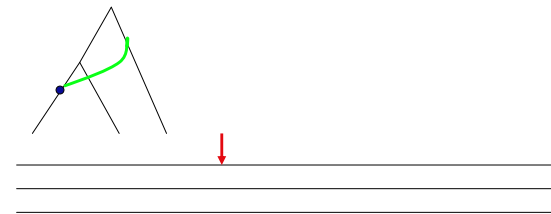
Old

Adding alleles one-by-one to a growing genealogy



Spatial Coalescent-Recombination Algorithm

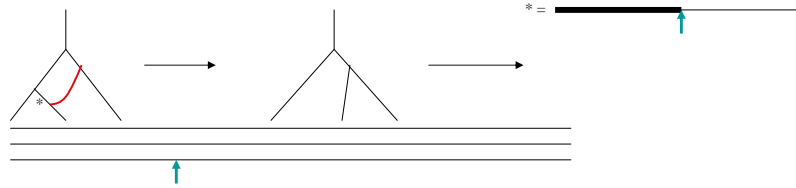
(Wiuf & Hein 1999 TPB)



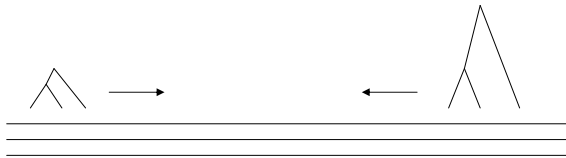
1. Make coalescent for position 0.0.
2. Wait $\text{Expo}(\text{Total Branch length})$ until recombination point, p .
3. Pick recombination point (*) uniformly on tree branches.
4. Let new sequence coalesce into genealogical structure. Continue 1-4 until $p > L$.

Properties of the spatial process

i. The process is non-Markovian



ii. The trees cannot be reduced to Topologies



How many Genetic Ancestors does a population have?

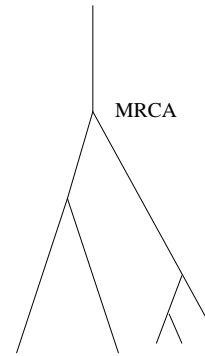
No recombination

Mitochondria

Y-chromosome

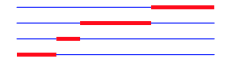
Recombination

X + autosomal chromosomes



Recombination-Coalescence Equilibrium:

(Sample independent)



MRCA at each position:

(Sample dependent)

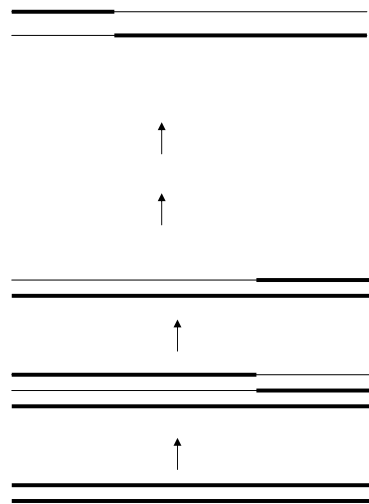


Present sample:

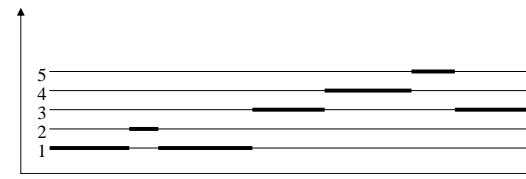


Tracing one sequence back in time.

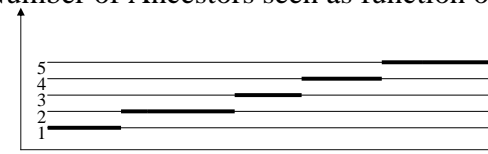
From Wiuf & Hein 1997



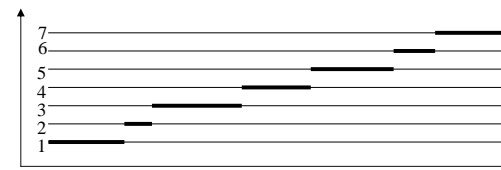
One realisation of a set of ancestors



Number of Ancestors seen as function of sequence length:



Number of Segments seen as function of sequence length:



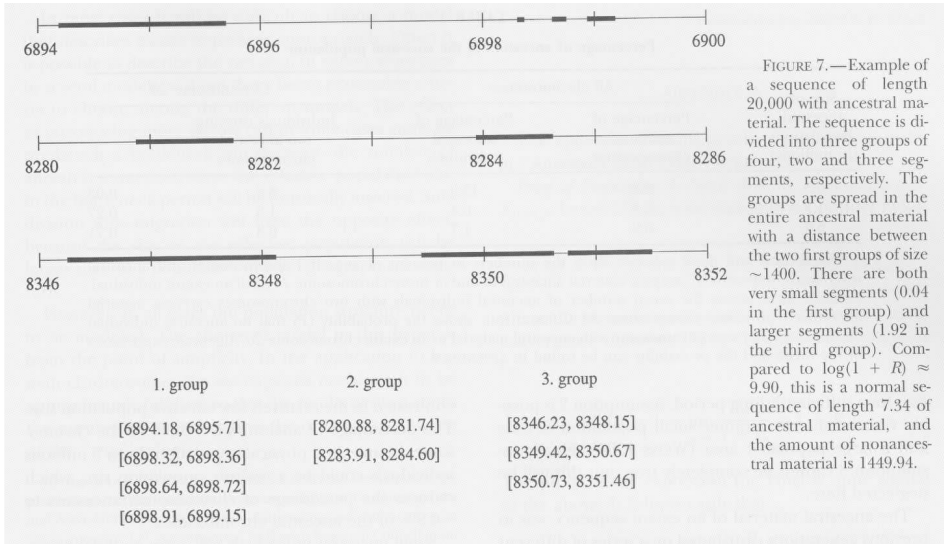


FIGURE 7.—Example of a sequence of length 20,000 with ancestral material. The sequence is divided into three groups of four, two and three segments, respectively. The groups are spread in the entire ancestral material with a distance between the two first groups of size ~ 1400 . There are both very small segments (0.04 in the first group) and larger segments (1.92 in the third group). Compared to $\log(1 + R) \approx 9.90$, this is a normal sequence of length 7.34 of ancestral material, and the amount of nonancestral material is 1449.94.

Number of ancestors as function of sequence length

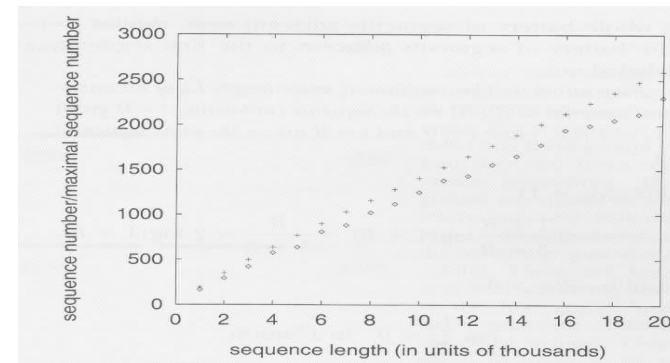


FIGURE 5.—Actual sequence number A_R vs. the total number of sequences C_R . The actual sequence number is the sequence number at the endpoint of the sequence, *i.e.*, the value of A_i in the point $t = R$. One hundred simulations were performed with sequence length varying from 1000 to 20,000. + denotes the mean number of sequences, and \diamond denotes the mean of the actual sequence number. The actual number is between 10–15% less than the total, which shows that a return to sequences with small sequence numbers rarely occurs.

Number of ancestors to the Human Genome

S_p – number of Segments

L_p – amount of ancestral material on sequence 1.

$$\rho = 4N_e * r -$$

N_e : Effective population size, r : expected number of recombinations per generation.

Theoretical Results

$$E(S_p) = 1 + \rho$$

$$E(L_p) = \log(1+\rho)$$

$P(\text{number of segments in } [0, \rho])$ as $\rho \rightarrow \infty > 0$

Applications to Human Genome

Parameters used $4N_e$ 20.000 Chromos. 1: 263 Mb. 263 cM

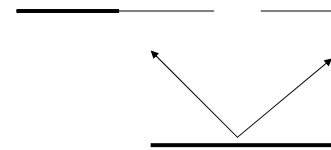
Chromosome 1: Segments 52.000 Ancestors 6.800

All chromosomes Ancestors 86.000

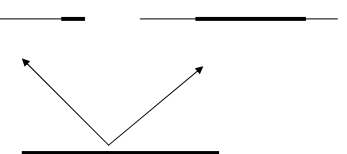
Physical Population. 1.3-5.0 Mill.

Gene Conversion

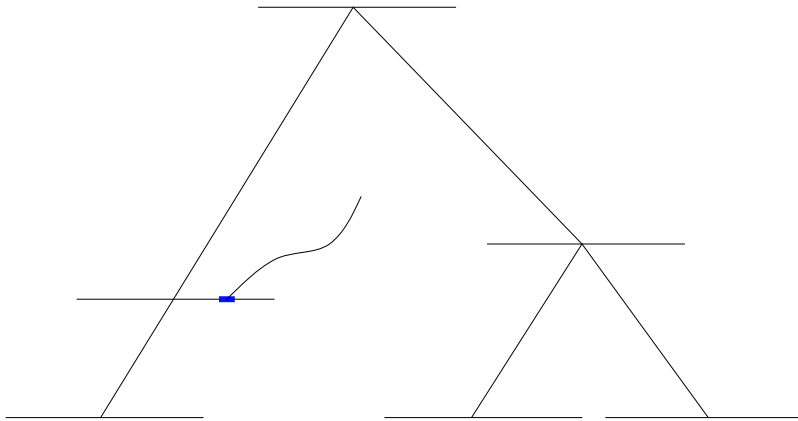
Recombination:



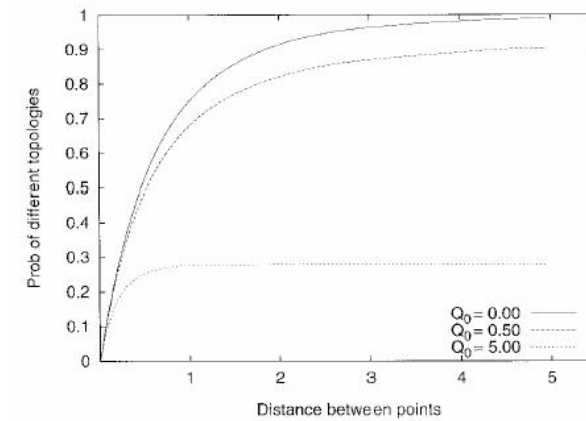
Gene Conversion:



Treeness



From Wiuf & Hein 2000



Consequences of Recombination

Incompatible Sites

00
01
11
10

Varying Divergence Along Sequence.

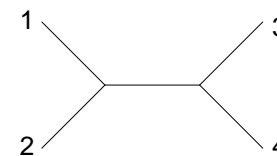
1.5 0.4

Convergence from high correlation to no correlation along sequence.

Topology Shifts along sequences.

Compatibility

	1	2	3	4	5	6	7
1	A	T	G	T	G	T	C
2	A	T	G	T	G	A	T
3	C	T	T	C	G	A	C
4	A	T	T	C	G	T	A
		i	i	i			



i. 3 & 4 can be placed on same tree without extra cost.

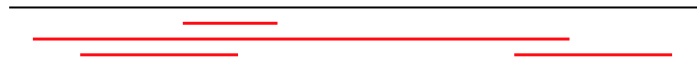
ii. 3 & 6 cannot.

Definition: Two columns are **incompatible**, if they are more expensive jointly, than separately on the cheapest tree.

Compatibility can be determined without reference to a specific tree!!

Hudson's R_M

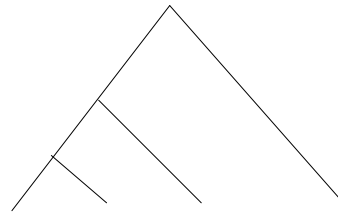
(k positions can at most have (k+1) types without recombination)
ex. Data set:



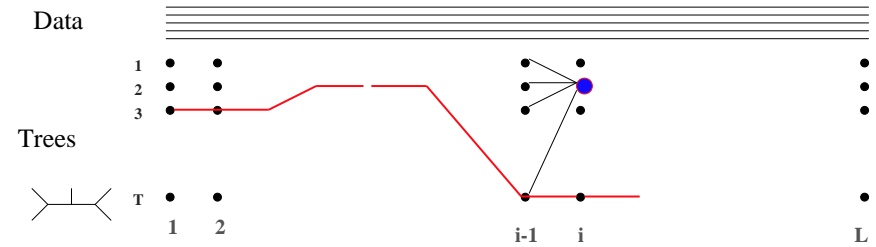
A underestimate for the number of recombination events:



If you equate R_M with expected number of recombinations, this would be an analogue to Watterson's estimators. Unfortunately, R_M is a gross underestimate of the real number of recombinations.



Recombination Parsimony



$$\text{Recursion: } W(T,i) = \min_{T'} \{ W(T',i-i) + \text{subst}(T,i) + d_{\text{rec}}(T,T') \}$$

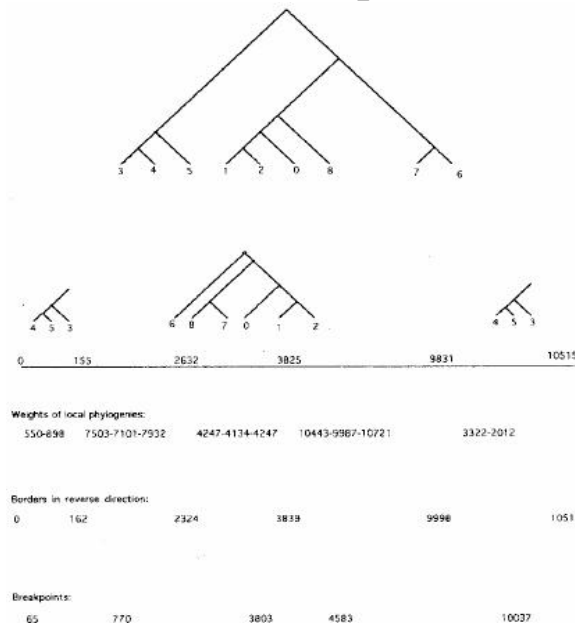
Fast heuristic version can be programmed.

Recombination Parsimony: Example - HIV

Costs:

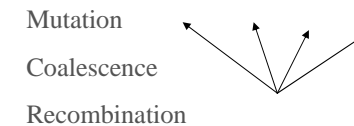
Recombination - 100

Substitutions - (2-5)



Likelihood approach to recombination

Griffiths, Tavaré (1994), Griffiths, Marjoram (1996) & Fearnhead, Donnelly (2001)



Data ATTCGTA ATGTGT ATGTGA CTTCGA
C T C

- i. Probability of Data as function of parameters (likelihood)
- ii. Statements about sequence history (ancestral analysis)
- iii. Hypothesis testing
- iv. Model Testing

Likelihood approach to recombination

(Griffiths, Marjoram (1996))

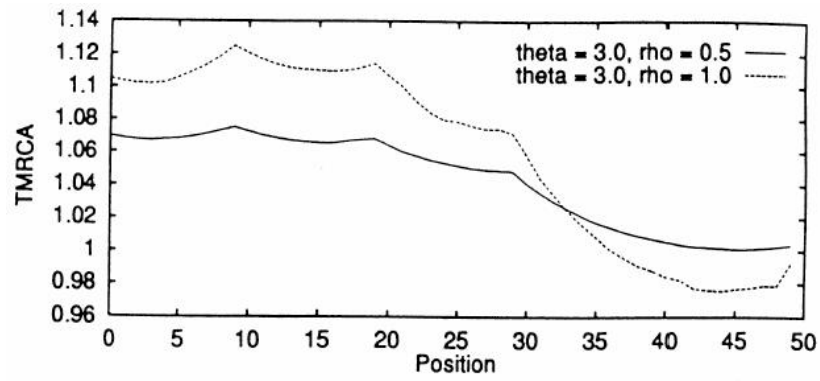


FIG. 7. TMRCA in two sequences with mutations.

Summary

What does Recombination do to Sequence Histories.

Probabilities of such histories.

Quantities of interest.

Detecting & Reconstructing Recombinations.