Recombination, Phylogenies and Parsimony

Overview:

The History of a set of Sequences

The Ancestral Recombination Graph (ARG) & the minimal ARG

Dynamical programming algorithm for finding the minimal ARG

Branch and Bound algorithm for minimal ARGs

Domain of Application:

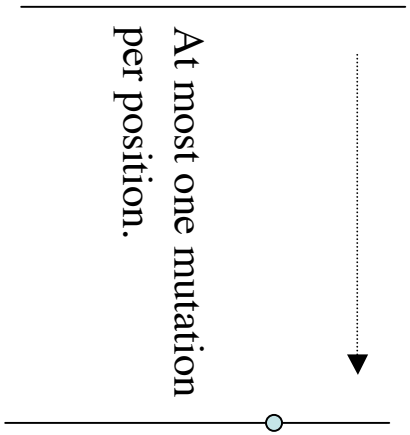
Sequence Variation

Fine scale mapping of disease genes

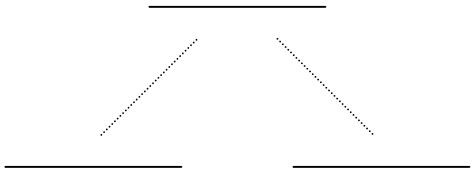
Pathogen Evolution

Mutations, Duplications/Coalescents & Recombinations

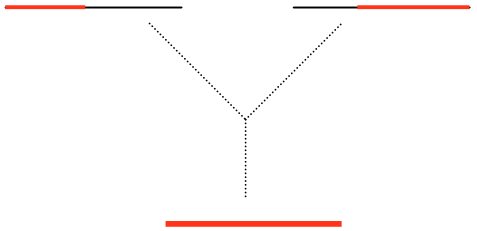
Mutation



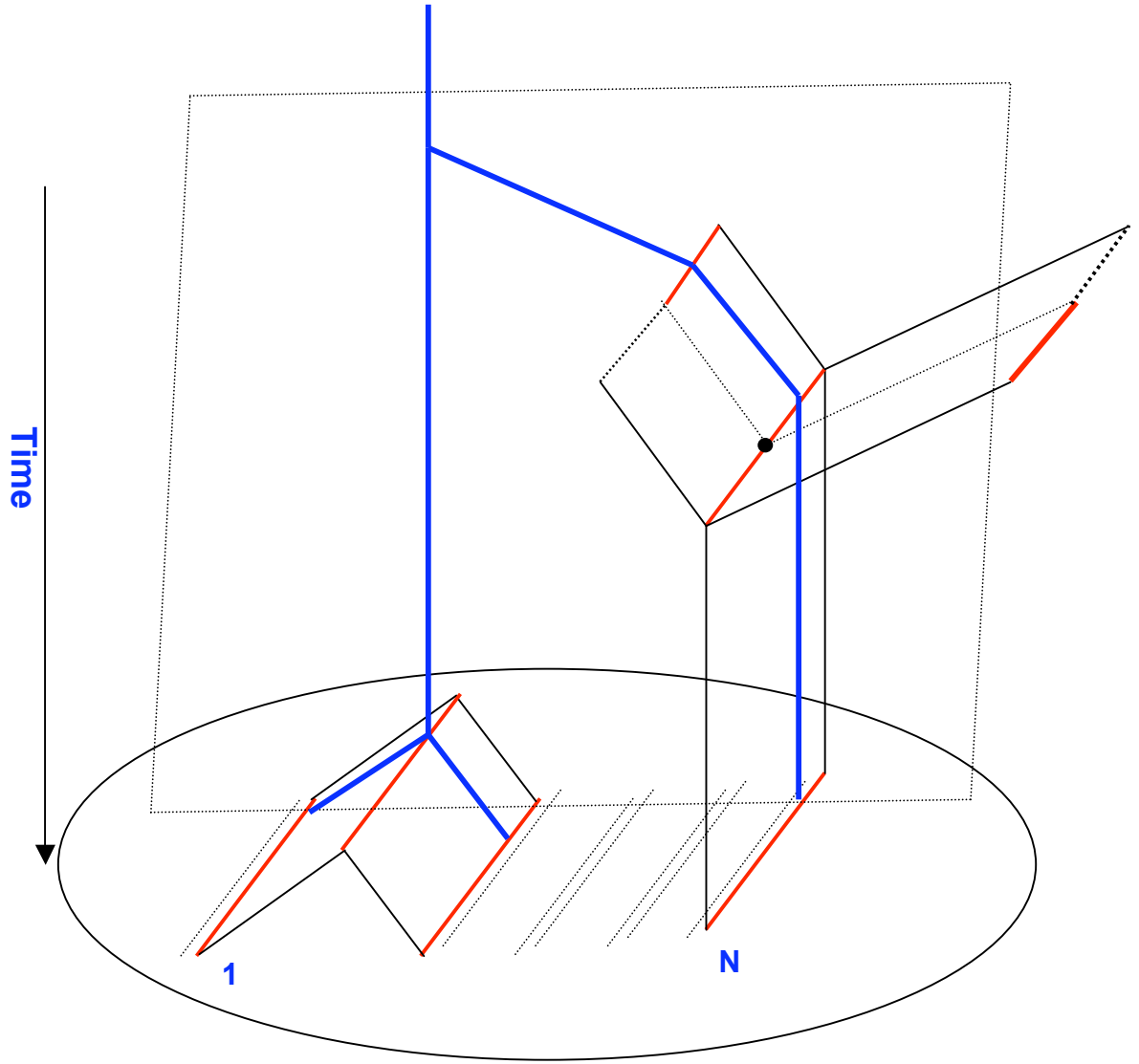
Duplication/ Coalescent



Recombination



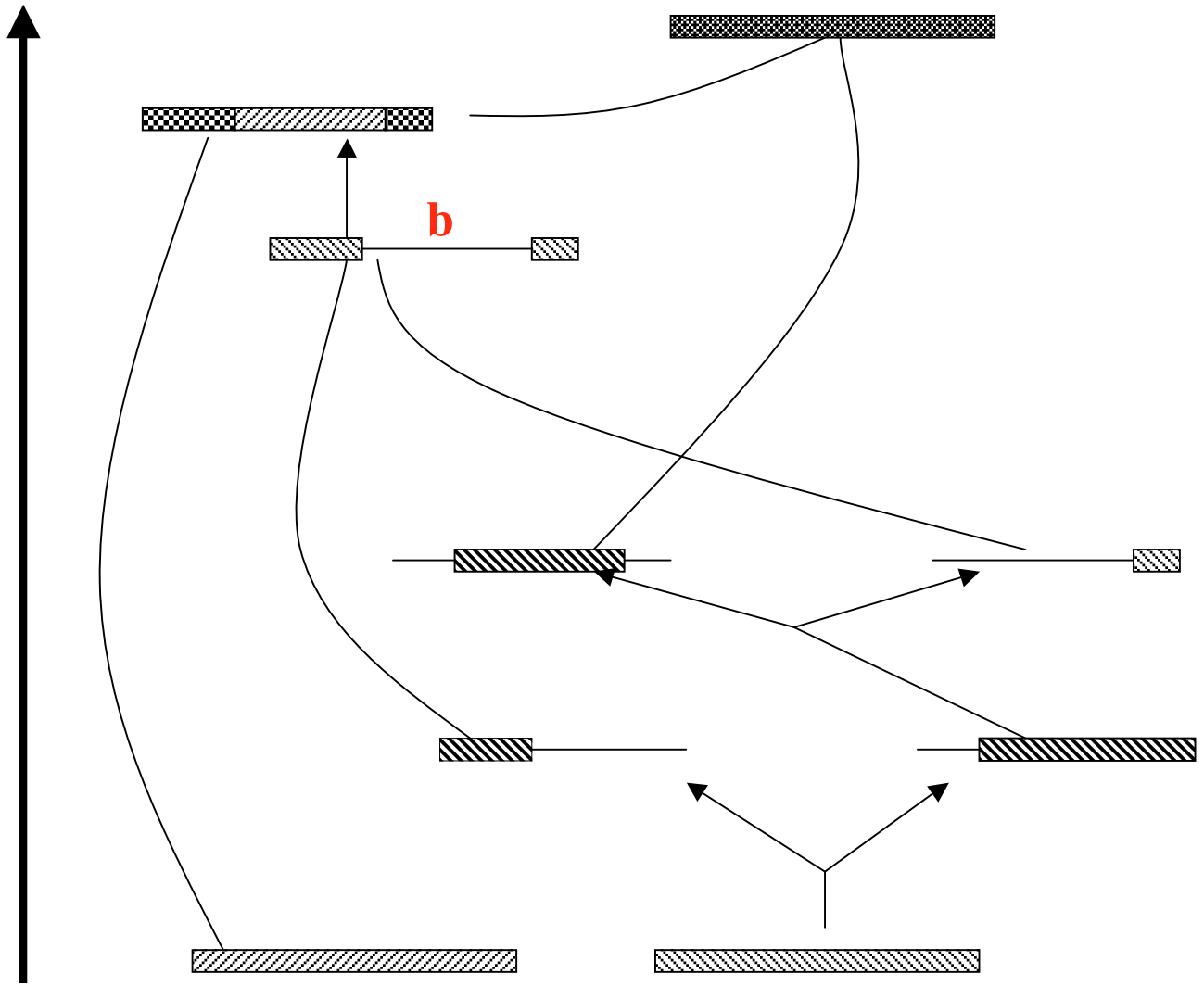
“The minimal number of recombinations for a set of sequences”



accgttgataggaaat……gta
accgttgataggaaat……gta
accgttgataggaaat……gta

Recombination-Coalescence Illustration

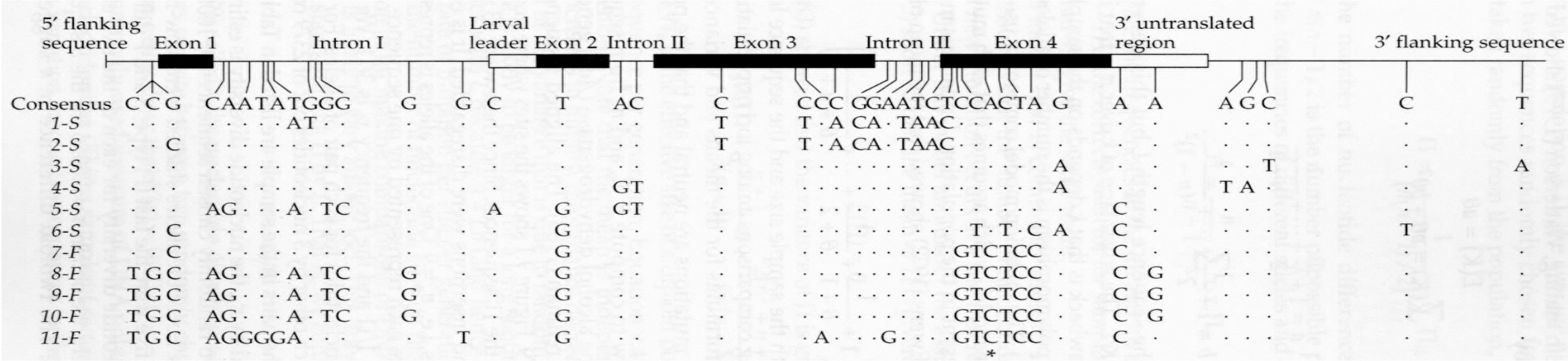
Copied from Hudson 1991



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

The 1983 Kreitman Data & the infinite site assumption

(M. Kreitman 1983 Nature from Hartl & Clark 1999)



Infinite Site Assumption (Otha & Kimura, 1971)

Each position is at most hit by one mutation

Recoded Kreitman data

i. (0,1) ancestor state known.

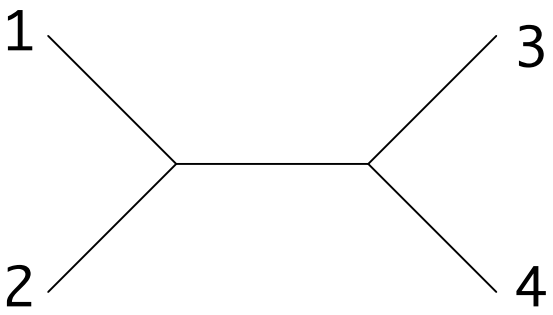
ii. Multiple copies represented by 1 sequences

iii. Non-informative sites could be removed

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

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.

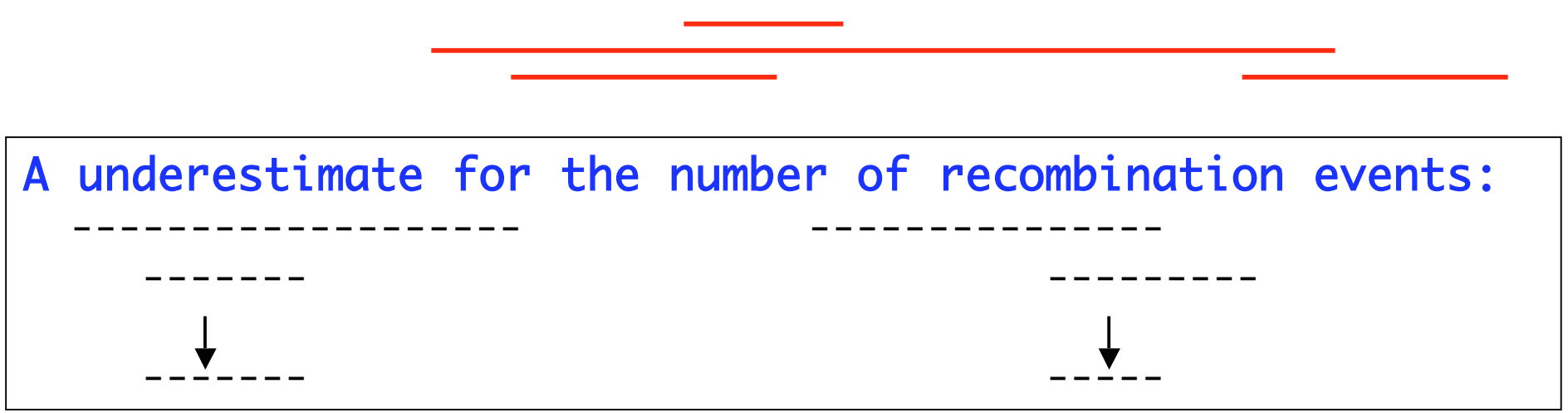
Compatibility can be determined without reference to a specific tree!!

Hudson & Kaplan's R_M

1985

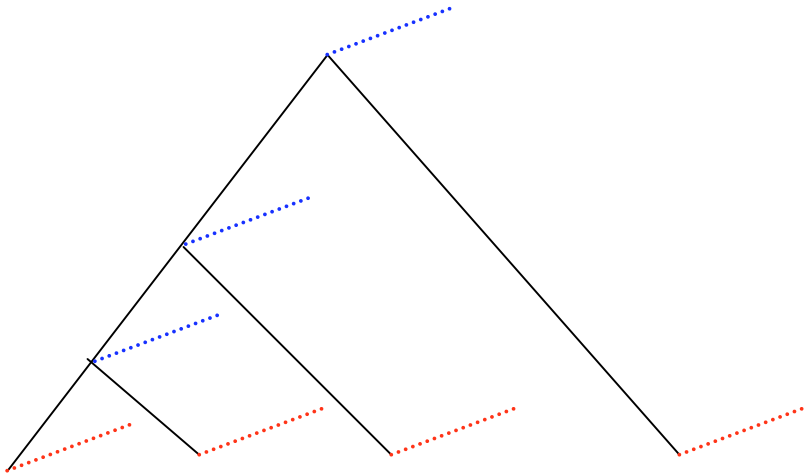
(k positions can have at most (k+1) types without recombination)

ex. Data set:



A underestimate for the number of recombination events:

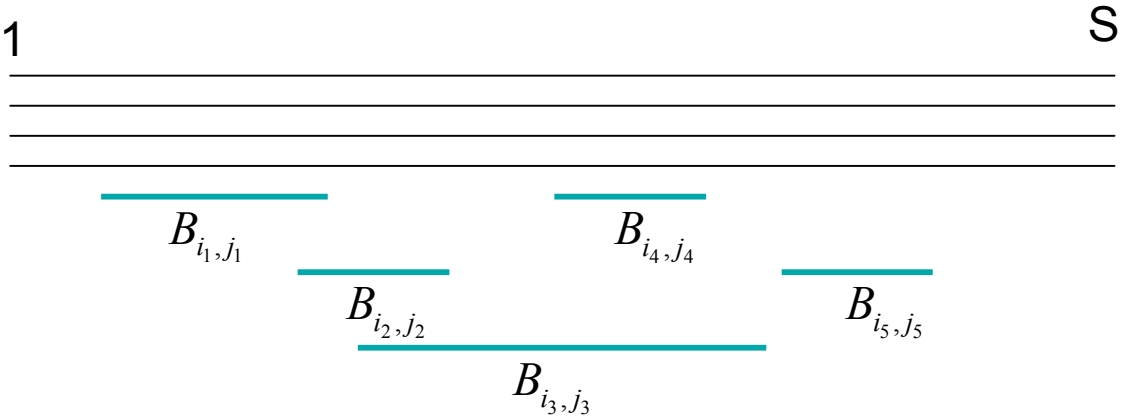
If you equate R_M with expected number of recombinations, this could be used as an estimator. Unfortunately, R_M is a gross underestimate of the real number of recombinations.



Myers-Griffiths' R_M

(2002)

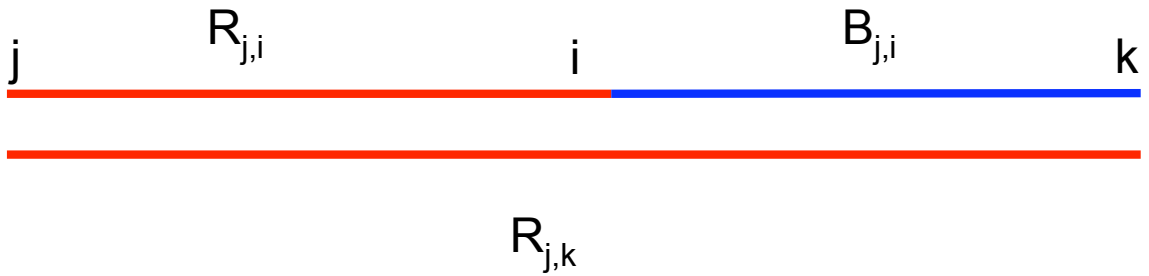
Basic Idea:



$$\text{Minimize } \sum_{l=1}^{S-1} r_l \text{ so } \sum_{l=i}^{j-1} r_l \geq B_{i,j} \text{ for all } B_{i,j}\text{'s and } r_l\text{'s positive}$$

Define R: $R_{j,k}$ is optimal solution to restricted interval., then:

$$R_{j,k} = \max\{R_{j,i} + B_{i,k} : i = j, j+1, ..k-1\}$$



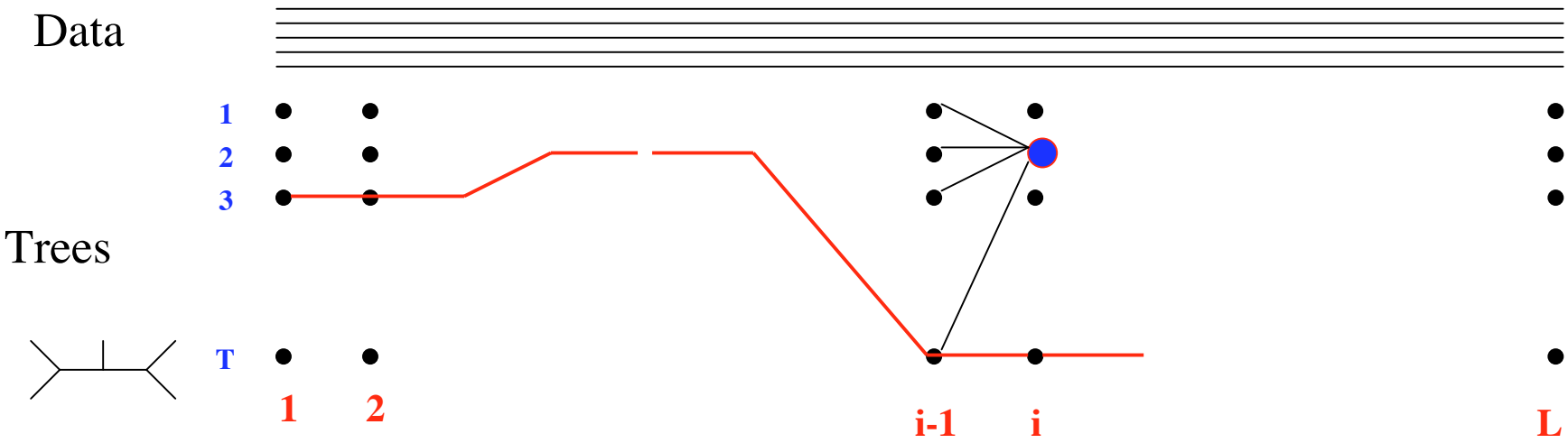
- 11 sequences of alcohol dehydrogenase gene in *Drosophila melanogaster.* Can be reduced to 9 sequences (3 of 11 are identical).
- 3200 bp long, 43 segregating sites.

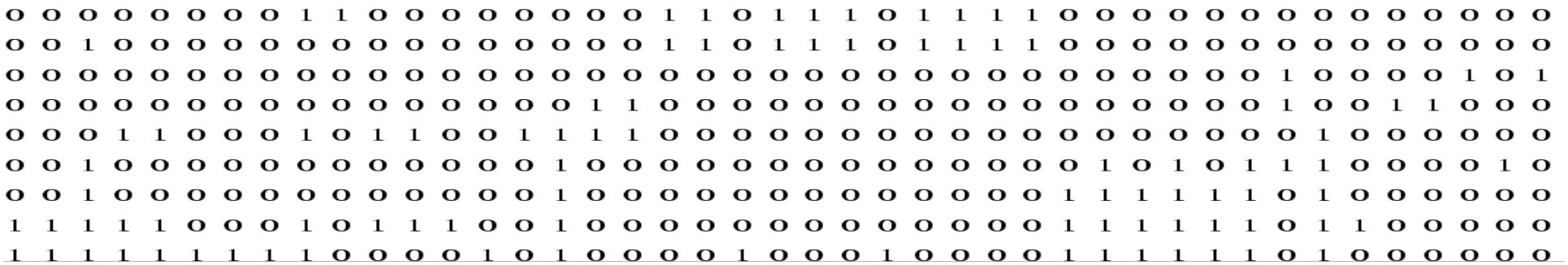
Methods	# of rec events obtained
Hudson & Kaplan (1985)	5
Myers & Griffiths (2002)	6
Song & Hein (2002). *Set theory based approach.*	7
Song & Hein (2003). *Current program using rooted trees.*	7

We have checked that it is possible to construct an ancestral recombination graph using only **7** recombination events.

Recombination Parsimony

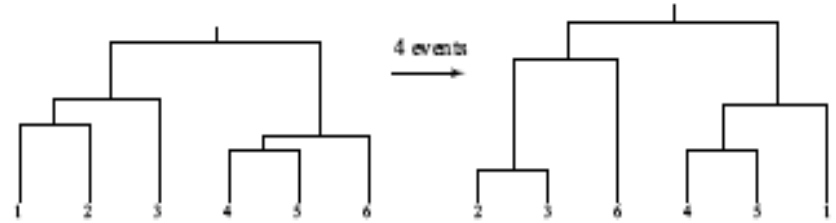
Hein, 1990,93 & Song & Hein, 2002+



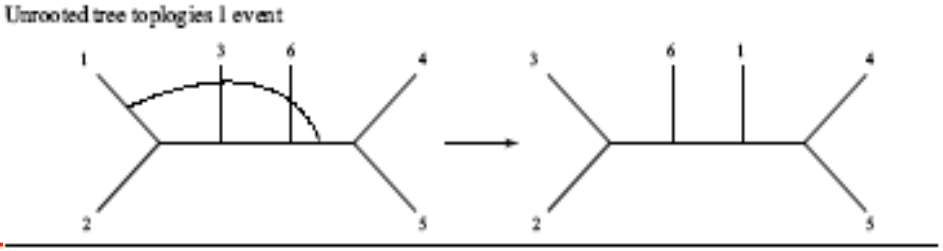


Metrics on Trees based on subtree transfers.

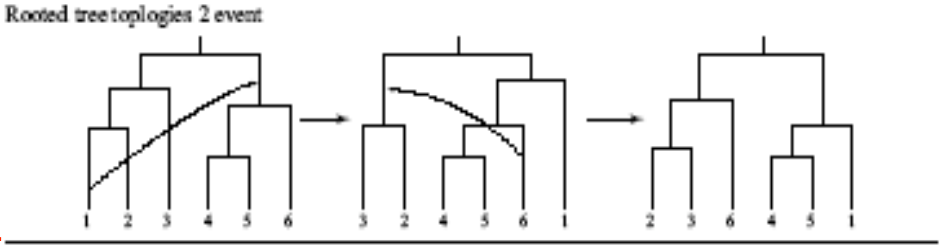
Trees including branch lengths



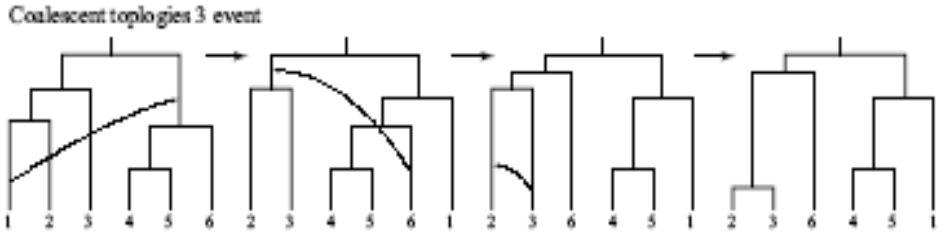
Unrooted tree topologies



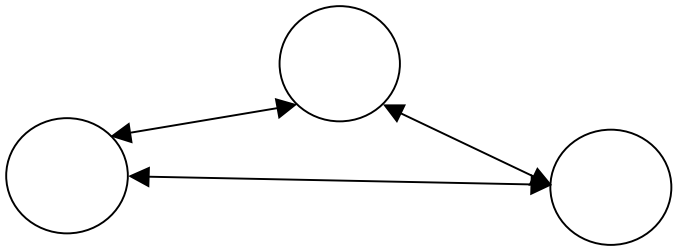
Rooted tree topologies



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	$\delta_{\max}$	$\delta_{\min}$	# of trees	$\delta_{\max}$	$\delta_{\min}$
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

$(2n-3)!! = \dfrac{(2n-2)!}{2^{n-1}(n-1)!}$

$2(n-3)(2n-7)$

Allen & Steel (2001)

$3n^2 - 13n + 14$

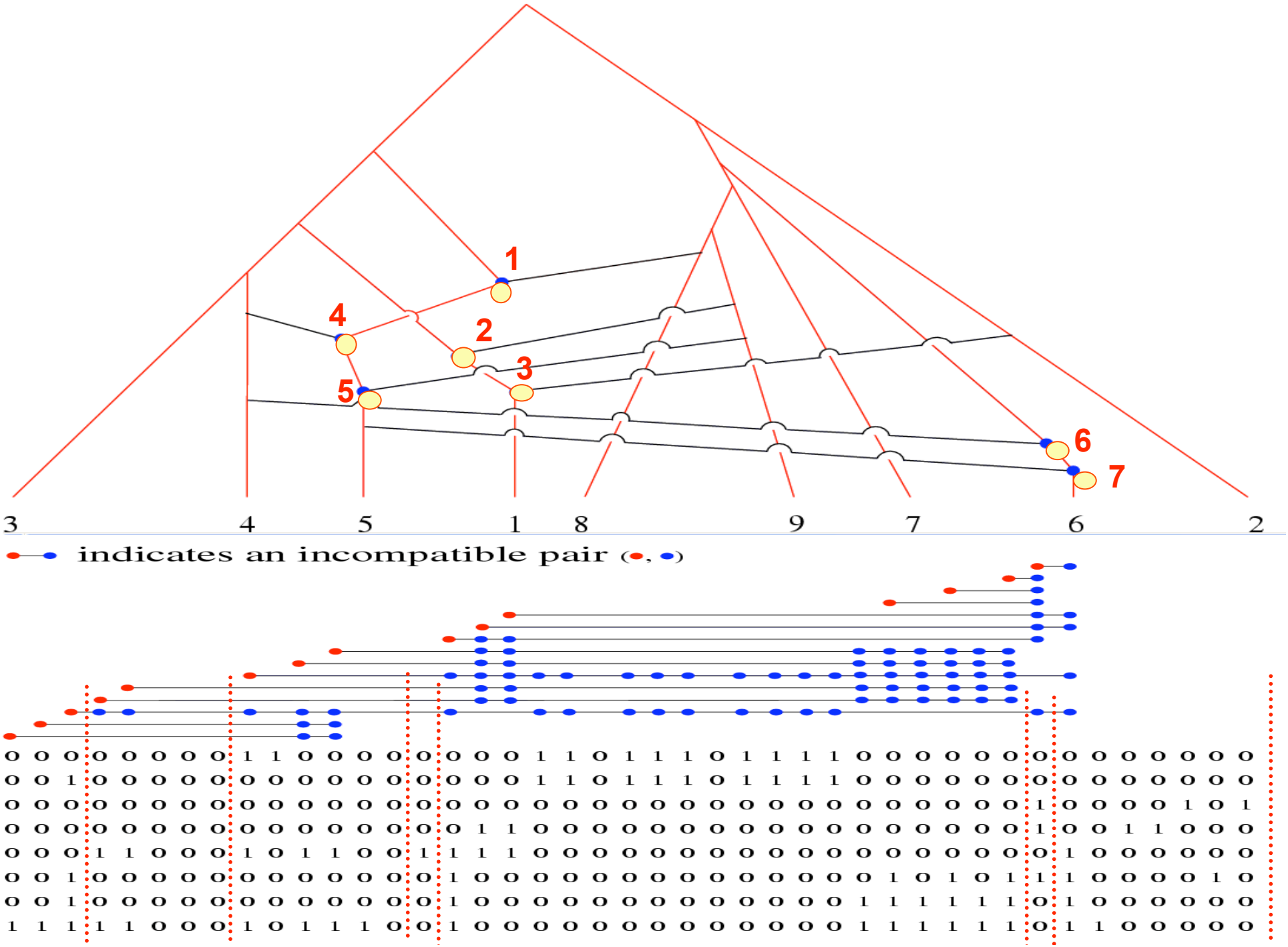
$4(n-2)^2 - 2\sum_{m=1}^{n-2} \lfloor \log_2(m+1) \rfloor$

$\dfrac{n!\,(n-1)!}{2^{n-1}}$

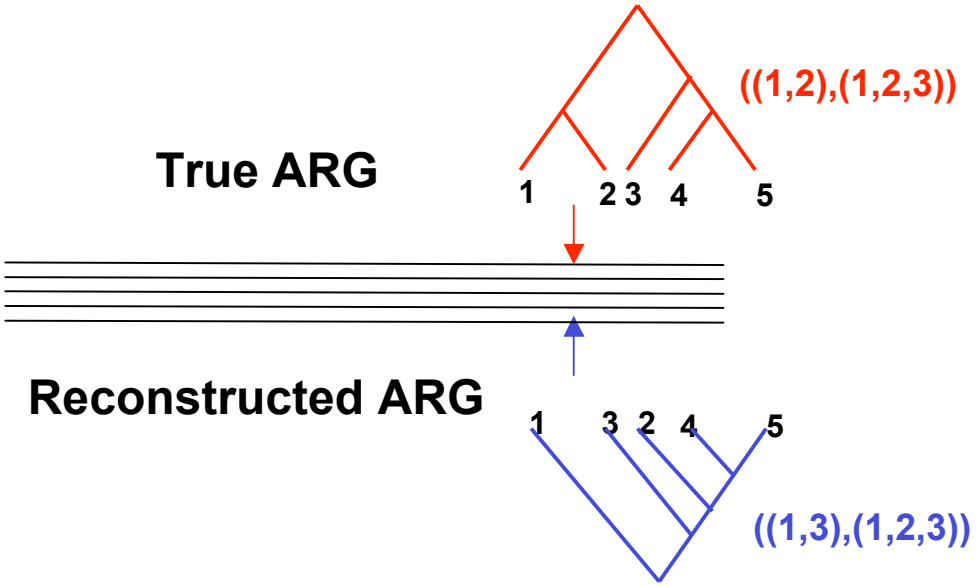
$\dfrac{1}{6}\left\{4n^3 - 9n^2 - 13n + 42 - 3(2n+3)\left\lfloor \dfrac{n-1}{2} \right\rfloor + 9\left(\left\lfloor \dfrac{n-1}{2} \right\rfloor\right)^2\right\}$

$\dfrac{1}{3}(2n^3 - 3n^2 - 20n + 39)$

Song (2003+)


1
4
2
3
5
6
7
3 4 5 1 8 9 7 6 2
indicates an incompatible pair (•, •)

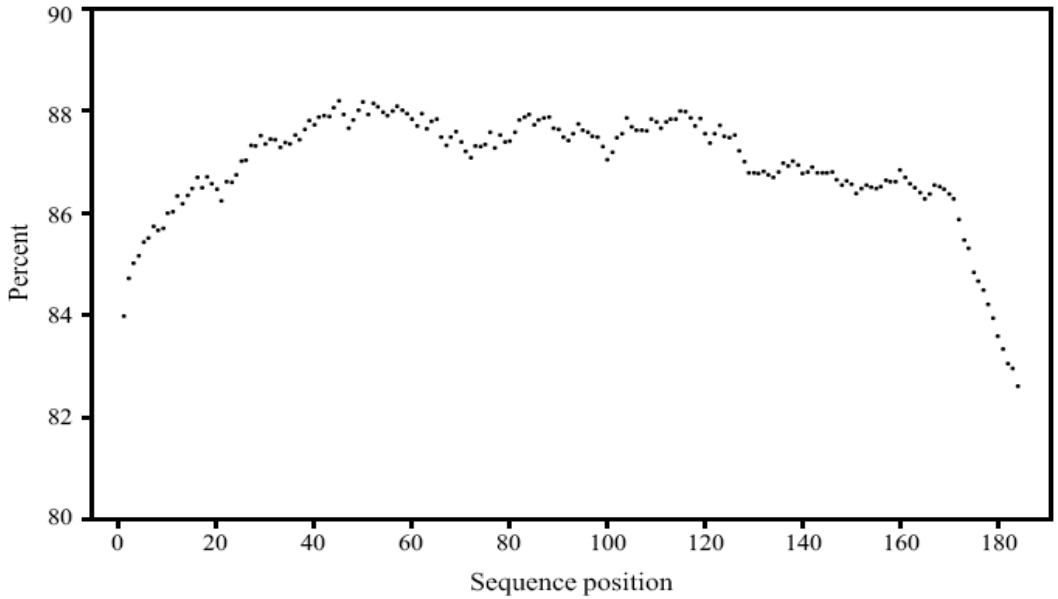
The Good News: Quality of the estimated local tree



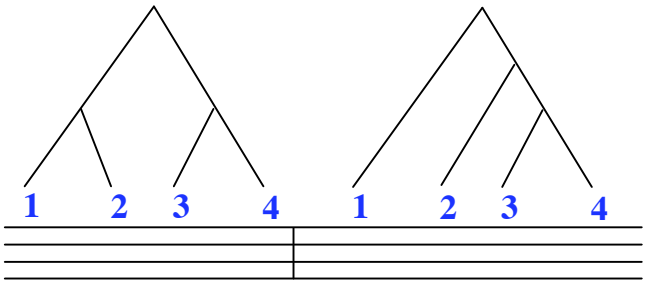
n=7

ρ=10

Θ=75



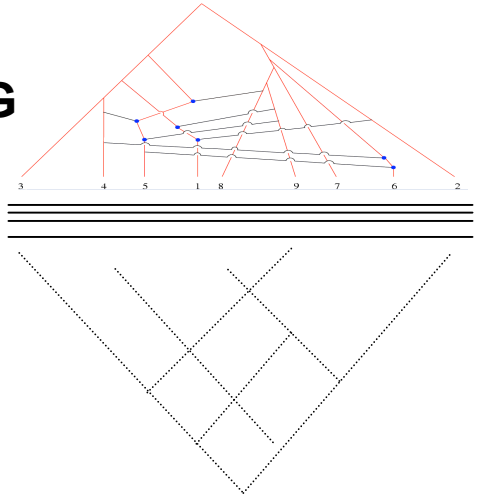
The Bad News: Actual, potentially detectable and detected recombinations

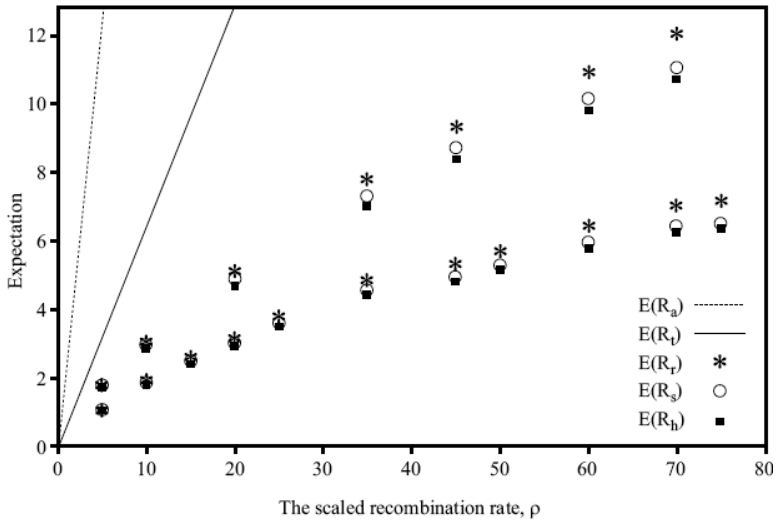


Leaves	Root	Edge-Length	Topo-Diff	Tree-Diff
2	1.0	2.0	0.0	.666
3	1.33	3.0	0.0	.694
4	1.50	3.66	0.073	.714
5	1.60	4.16	0.134	.728
6	1.66	4.57	0.183	.740
10	1.80	5.66	0.300	.769
15	1.87	6.50	0.374	.790
500	1.99		0.670	

Minimal ARG

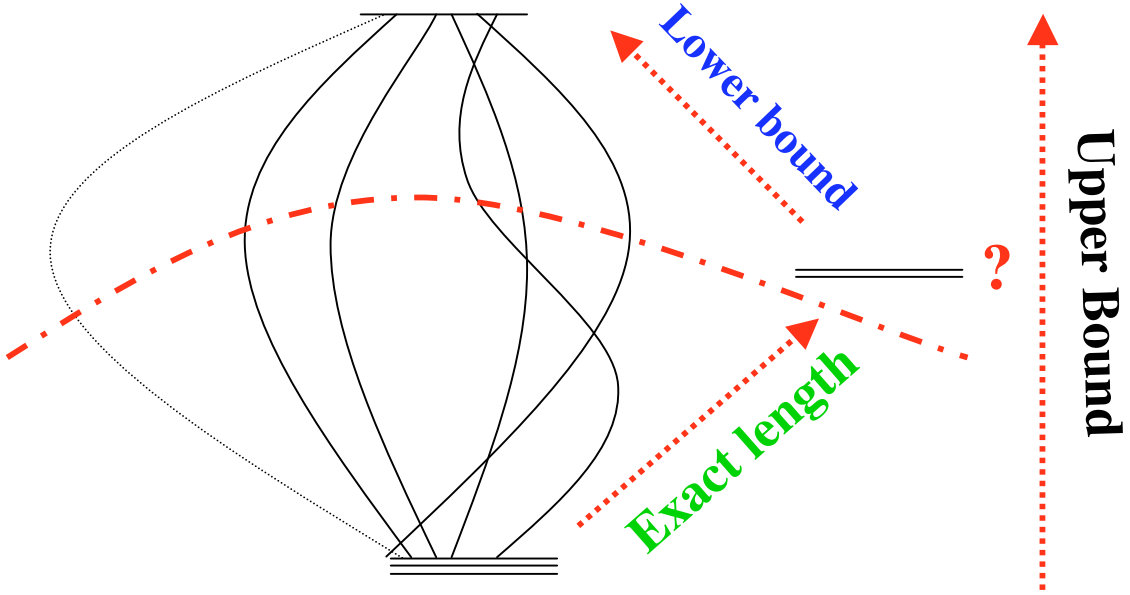
True ARG

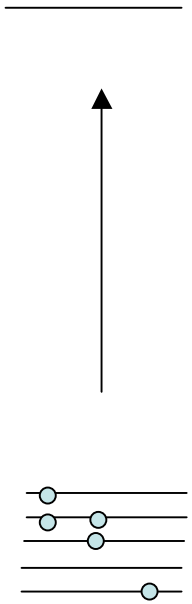




0 **4 Mb**

Branch and Bound Algorithm





k	k-recombination neighborhood	AC's encountered on k-recombi. ARG
0	3	0
1	91	94
2	1314	1312
3	8618	9618
4	30436	30436
5	62794	62794
6	78970	79970
7	63049	63049
8	32451	32451
9	10467	3467
10	1727	1727

1. The number of ancestral sequences in the ACs.
2. Number of ancestral sequences in the ACs for neighbor pairs
3. AC compatible with the minimal ARG.
4. AC compatible with close-to-minimal ARG.

Recombination, Phylogenies and Parsimony

Overview:

The History of a set of Sequences

The Ancestral Recombination Graph (ARG) & the minimal ARG

Dynamical programming algorithm for finding the minimal ARG

Branch and Bound algorithm for minimal ARGs

Domain of Application:

Sequence Variation

Fine scale mapping of disease genes

Pathogen Evolution

References

•Allen, B. and Steel, M., Subtree transfer operations and their induced metrics on evolutionary trees,*Annals of Combinatorics* 5, 1-13 (2001)

•Baroni, M., Grunewald, S., Moulton, V., and Semple, C. Bounding the number of hybridisation events for a consistent evolutionary history. Journal of Mathematical Biology 51 (2005), 171-182

•Bordewich, M. and Semple, C. On the computational complexity of the rooted subtree prune and regraft distance. Annals of Combintorics 8 (2004), 409-423

•Griffiths, R.C. (1981). Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.* **19**, 169-186.

•J.J.Hein: Reconstructing the history of sequences subject to Gene Conversion and Recombination. Mathematical Biosciences. (1990) 98.185-200.

•J.J.Hein: A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination. J.Mol.Evol. 20.402-411. 1993

•Hein,J.J., T.Jiang, L.Wang & K.Zhang (1996): "On the complexity of comparing evolutionary trees" Discrete Applied Mathematics 71.153-169.

•Hein, J., Schierup, M. & Wiuf, C. (2004) *Gene Genealogies, Variation and Evolution*, Oxford University Press

•Hudson, 1993 Properties of a neutral allele model with intragenic recombination.Theor Popul Biol. 1983 23(2):183-2

•Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster.Nature. 1983 304(5925):412-7.

•Lyngsø, R.B., Song, Y.S. & Hein, J. (2005) Minimum Recombination Histories by Branch and Bound. *Lecture Notes in Bioinformatics: Proceedings of* WABI 2005 *3692*: 239–250.

•Myers, S. R. and Griffiths, R. C. (2003). Bounds on the minimum number of recombination events in a sample history. *Genetics* **163**, 375-394.

•Song, Y.S. (2003) On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics,* 7:365–379

•Song, Y.S., Lyngsø, R.B. & Hein, J. (2005) Counting Ancestral States in Population Genetics. Submitted.

•Song, Y.S. & Hein, J. (2005) Constructing Minimal Ancestral Recombination Graphs. *J. Comp. Biol.*, 12:147–169

•Song, Y.S. & Hein, J. (2004) On the minimum number of recombination events in the evolutionary history of DNA sequences. *J. Math. Biol.*, 48:160–186.

•Song, Y.S. & Hein, J. (2003) Parsimonious reconstruction of sequence evolution and haplotype blocks: finding the minimum number of recombination events, *Lecture Notes in Bioinformatics, Proceedings of WABI'03*, 2812:287–302.

•Song YS, Wu Y, Gusfield D. Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution.Bioinformatics. 2005 Jun 1;21 Suppl 1:i413-i422.

•Wiuf, C. Inference on recombination and block structure using unphased data.Genetics. 2004 Jan;166(1):537-45.