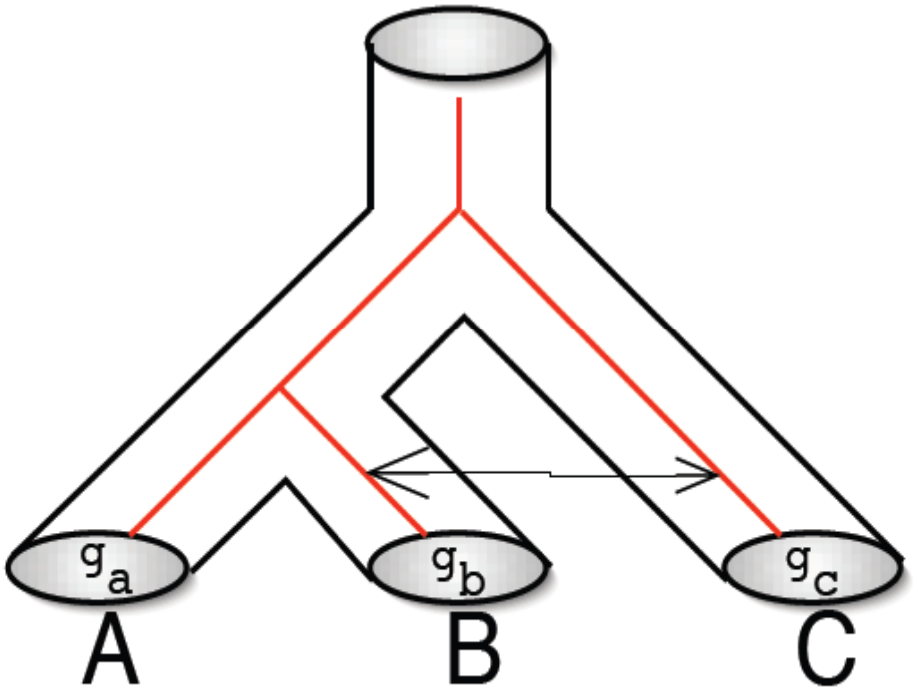
Topological Concordance of Gene Trees and Species Tree

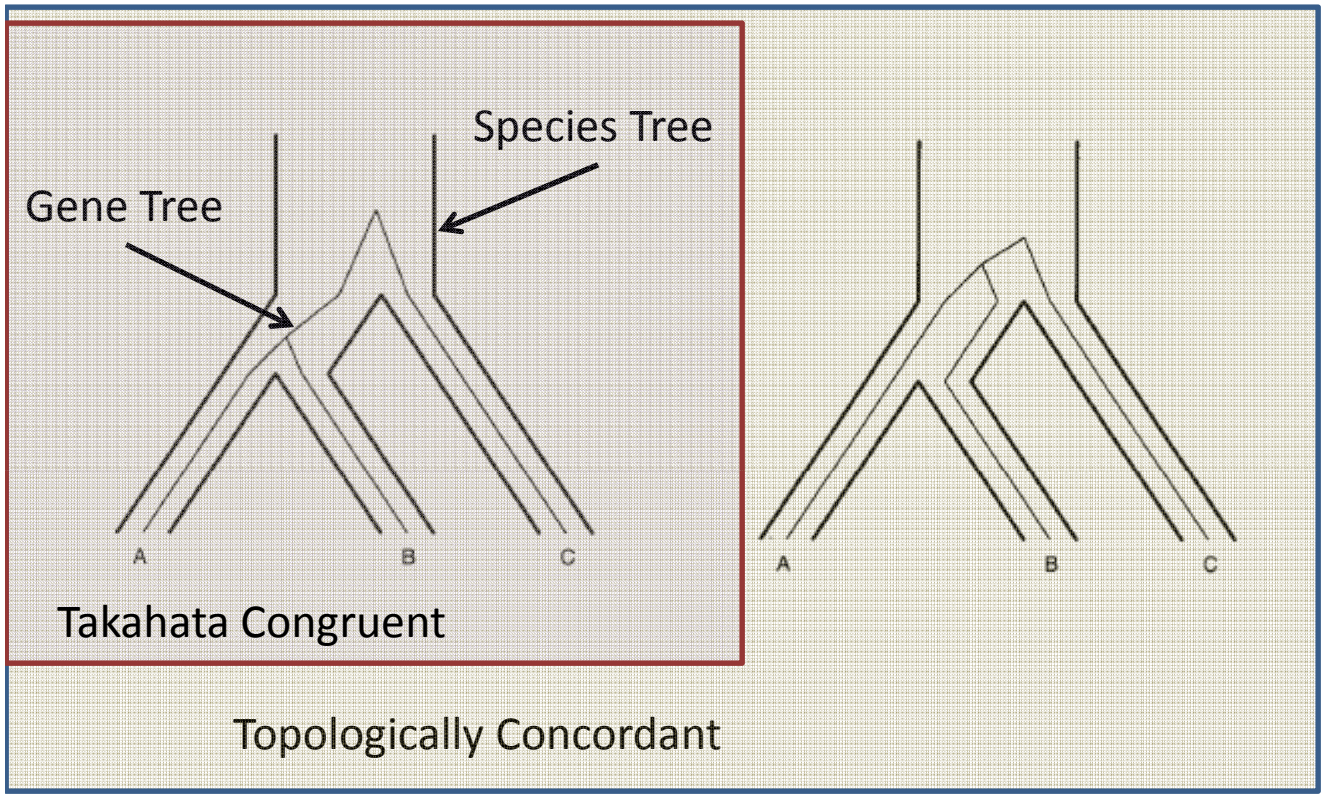
jatin narula

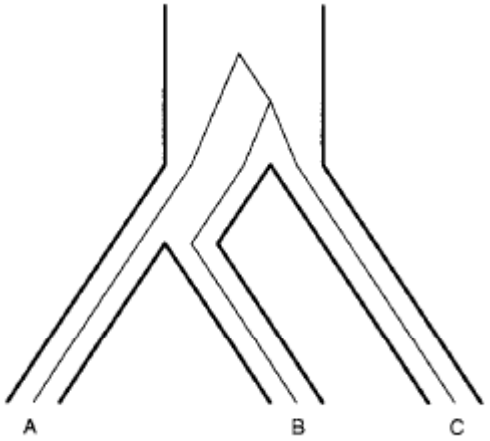
Gene Trees and Species Tree



- Gene Loss and Duplication
- Horizontal Gene Transfer and Recombination
- Stochastic Factors

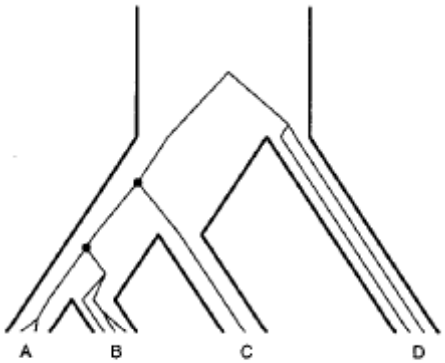
Topological Concordance

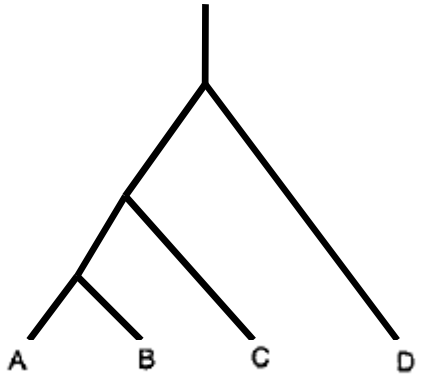




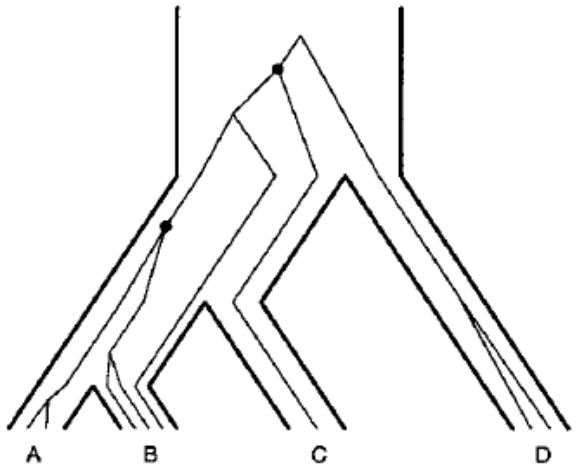
Takahata Congruence => Topological Concordance

Topological Concordance for Multiple Lineages

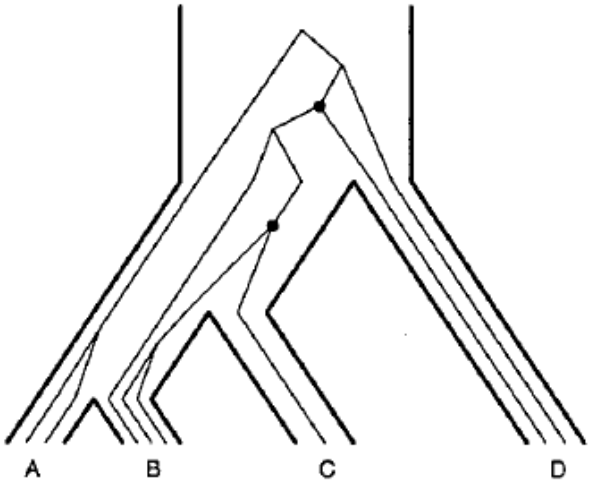




Collapsed gene tree is both topologically concordant and Takahata congruent

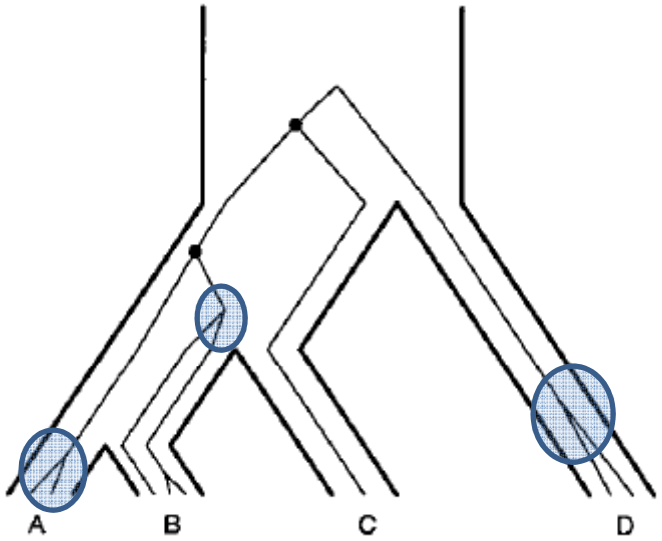


Topologically Concordant

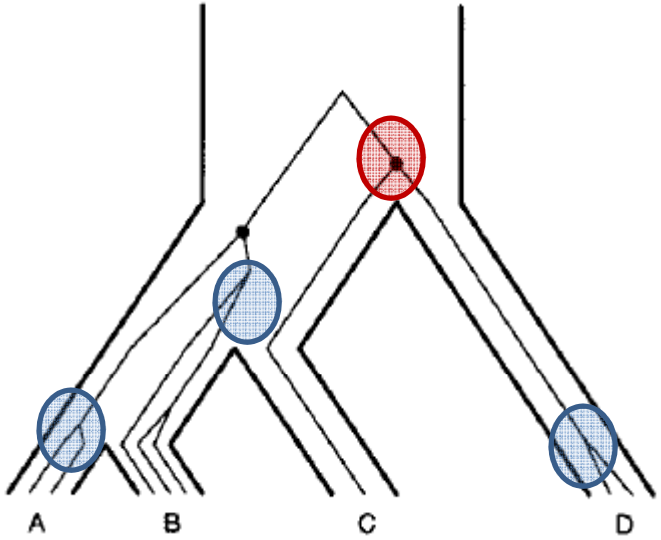


Neither

Monophyletically Concordant



Monophyletically Concordant



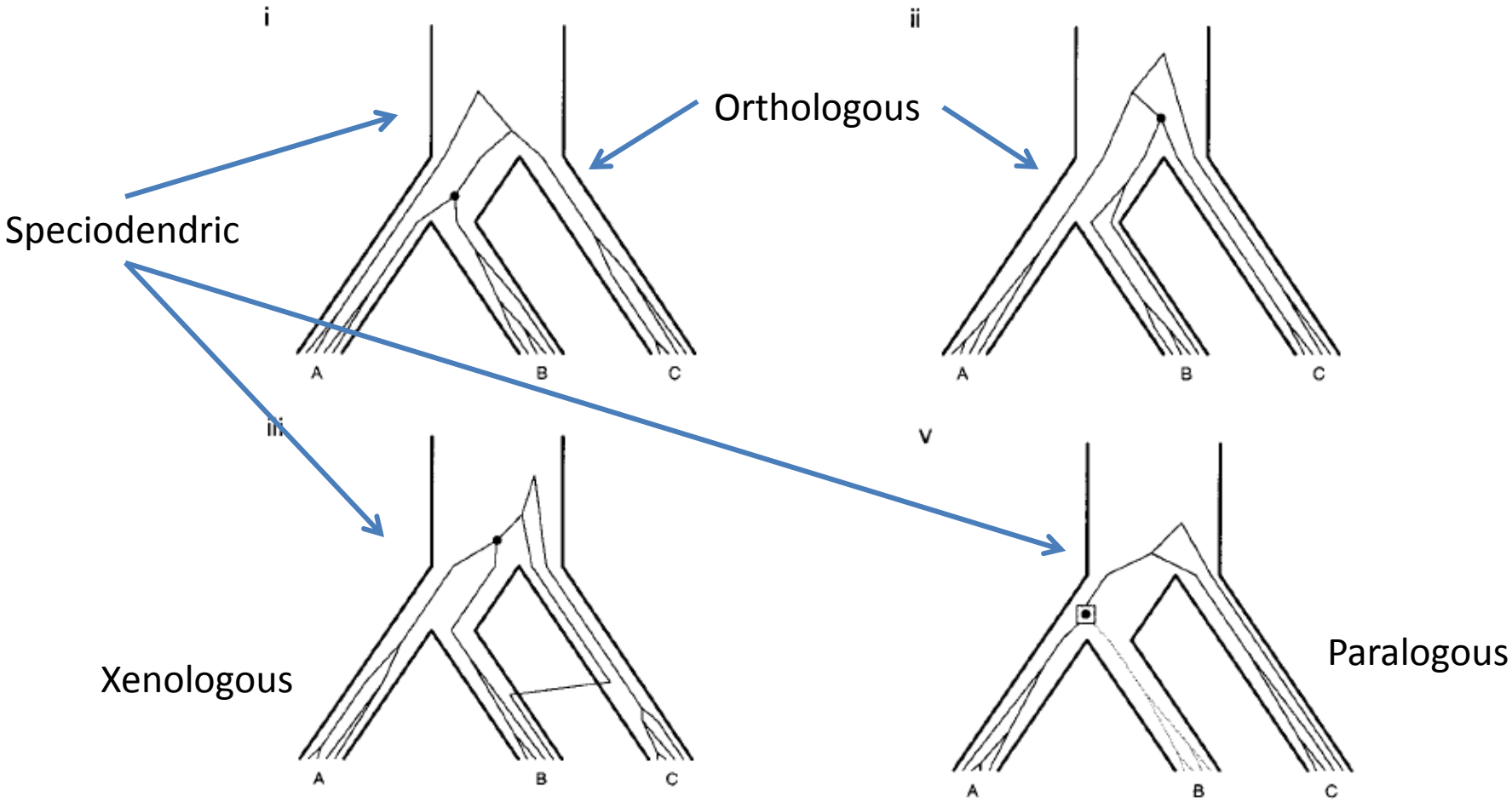
All species are Monophyletic but gene tree is not topologically concordant

Monophyletic + Topologically Concordant = Monophyletically Concordant

Speciodendric Genes

Orthology : Genes whose homology was the result of speciation and subsequent descent, with no duplication

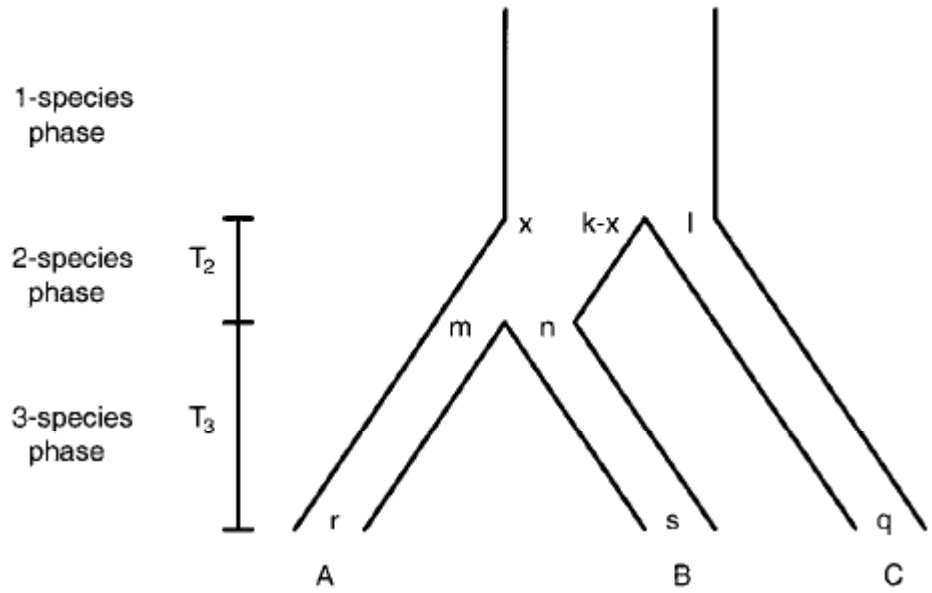
Speciodendricity : Gene Tree constructed from all copies of the gene in all species is topologically concordant



The Problem

‘‘conditioned on the species tree topology and assuming no gene exchange between species, what is the probability that a tree of orthologous genes is topologically concordant with a species tree?’’

Takahata Concordance Probability



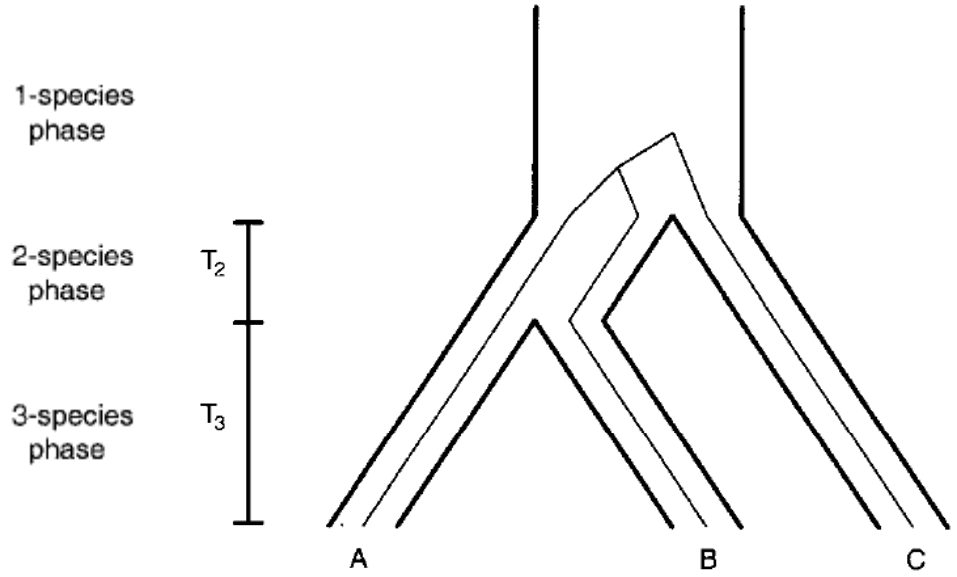
P(Takahata Congruence) = $\sum_{m=1}^{m} \sum_{n=1}^{s} \sum_{k=1}^{m+n} g_{rm}(T_3)*g_{sn}(T_3)*g_{m+n,k}(T_2) \; *F_k^{A,B}(m,n,0)$

P(r lineages derive from m lineages at time T_3)*P(s lineages derive from n lineages at time T_3)

*P(m+n lineages at T3 derive from k lineages at time T_3+T_2)

*P(at least one interspecific coalescence occurs during this process, *and that the most recent interspecific coalescence joins a lineage from species A and a lineage from species B)*

Topological Concordance Probability



P(Topological Concordance) =

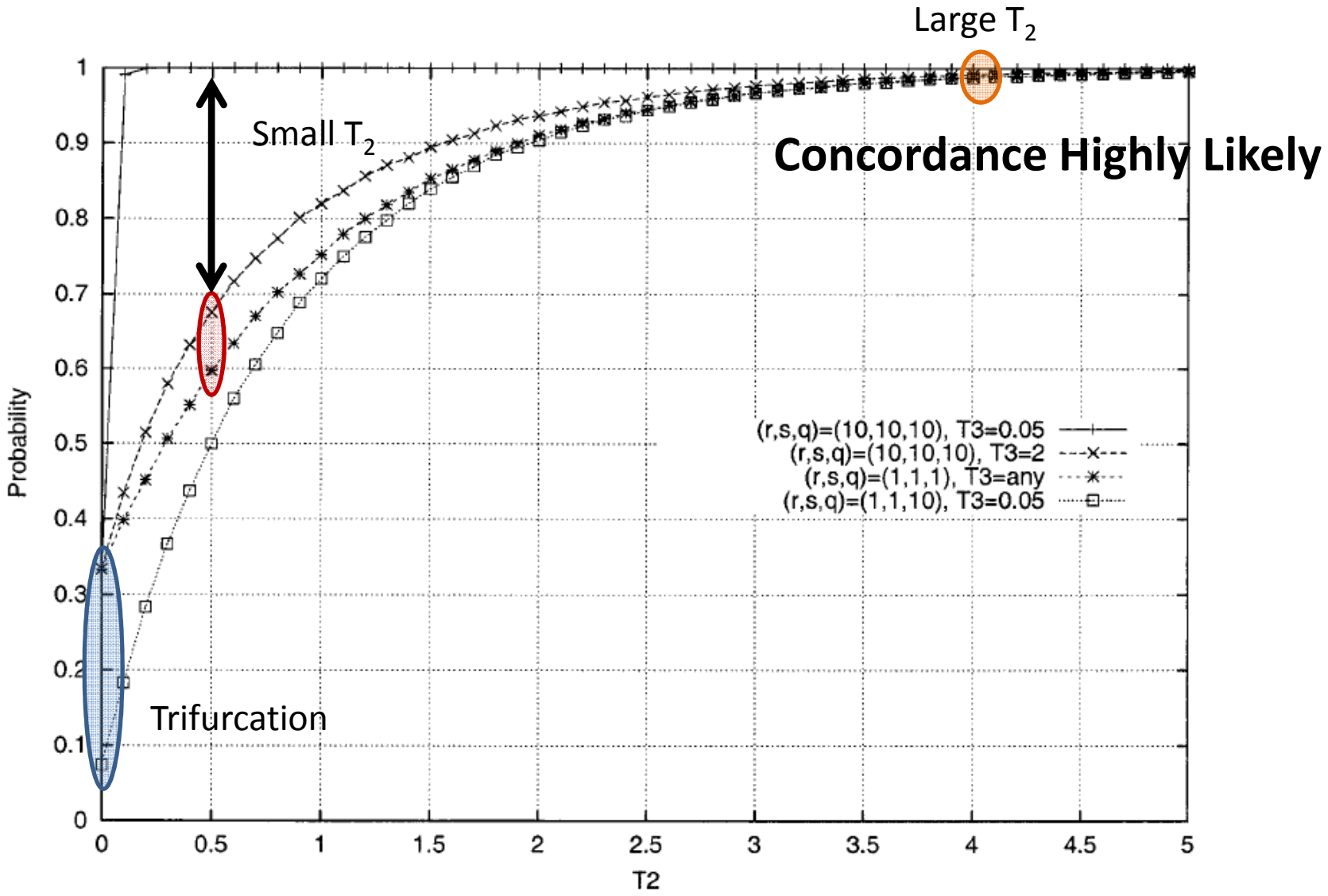
P(Tokabata Concordance) + P(no interspecific coalescences happen in two species

$$\sum_{m=1}^{r}\sum_{n=1}^{s}\sum_{k=1}^{m+n}\left[g_{rm}(T_3)\,g_{sn}(T_3)\,g_{m+n,k}(T_2)\times\left[F_k^{A,B}(m,n,0)+[1-F_k^{A,B}(m,n,0)]\times\sum_{x=1}^{k-1}\left[W_{(m,n),(x,k-x)}(T_2)\right.\right.\right.$$

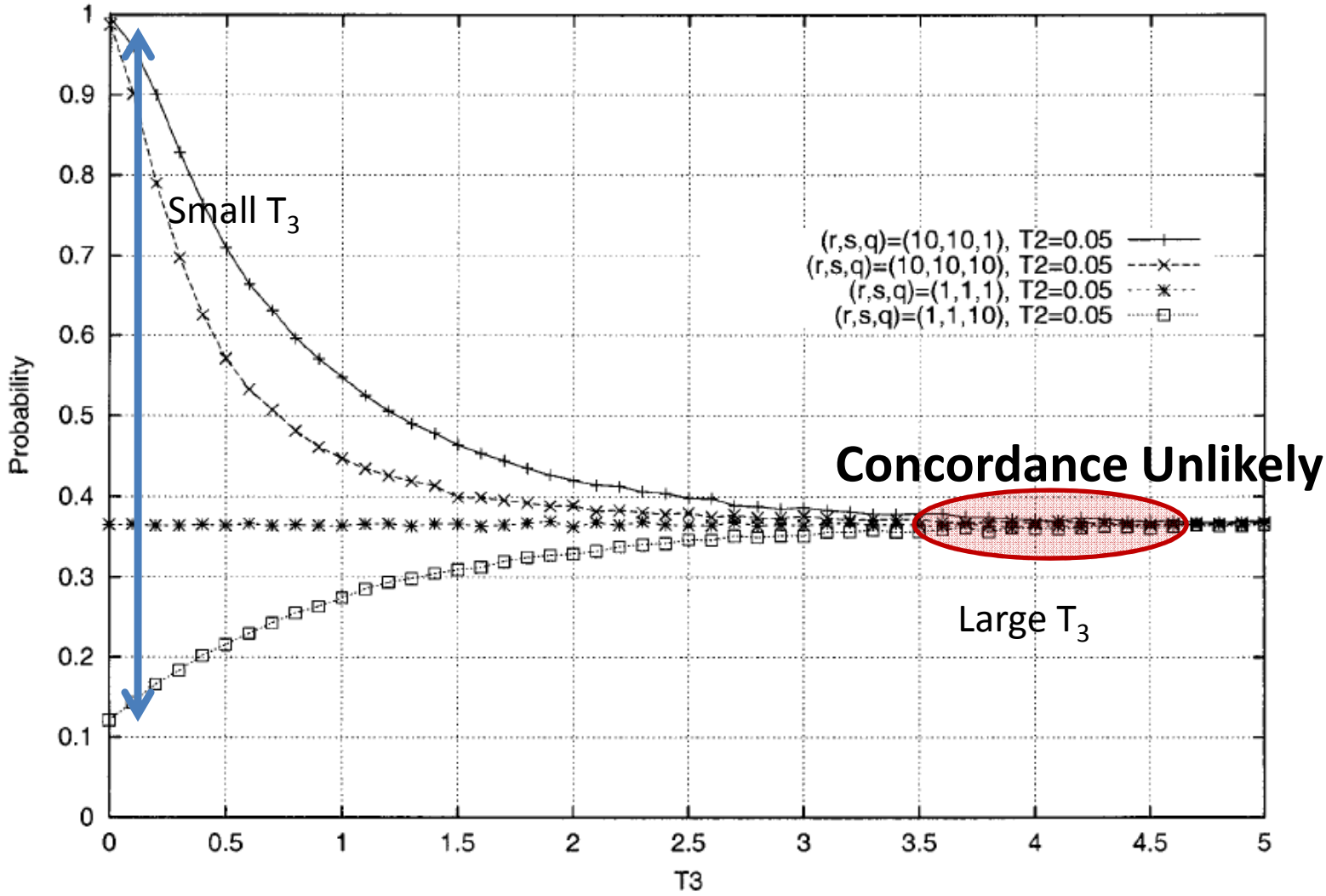
his coalescence joins the ancestral lineages of A and B)

$$\left.\left.\left.\times\sum_{l=1}^{q}[g_{ql}(T_3+T_2)\,F_1^{A,B}(x,k-x,l)]\right]\right]\right].$$

Key Determinants of Topological Concordance : T_2

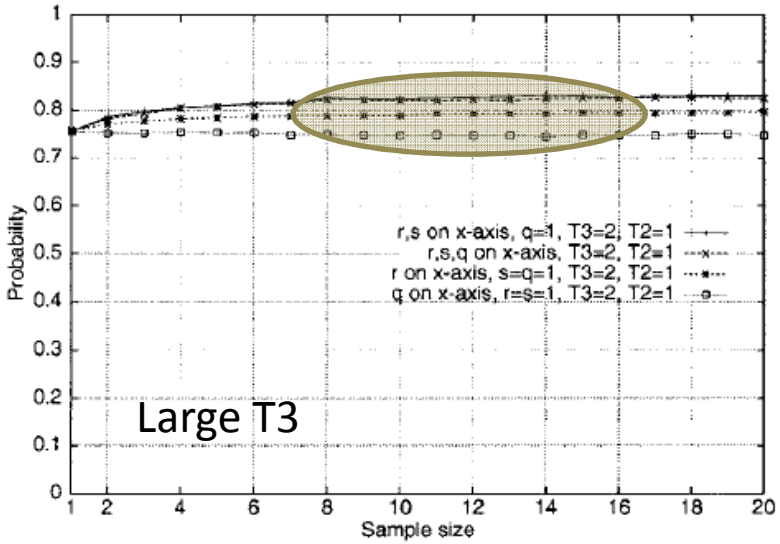


Key Determinants of Topological Concordance : T_3

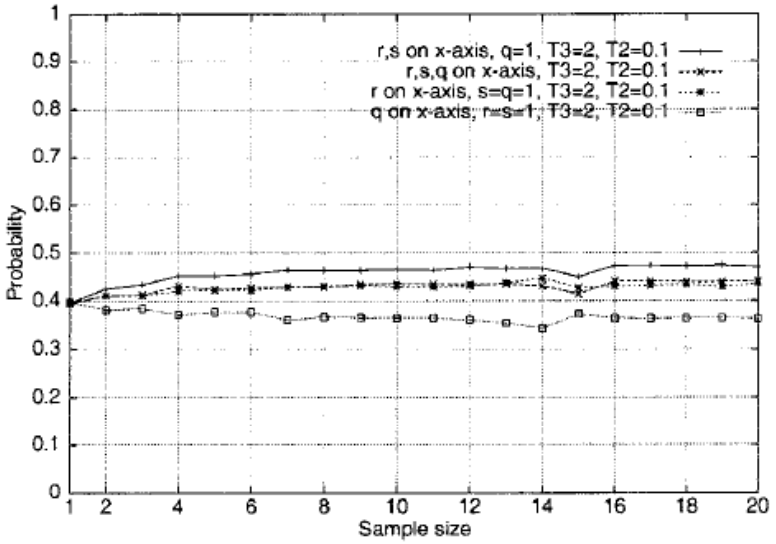


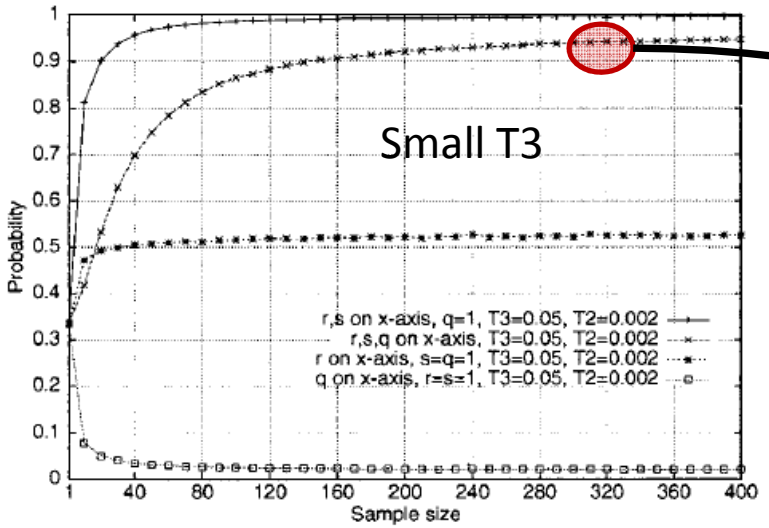
Key Determinants of Topological Concordance : Sample Sizes

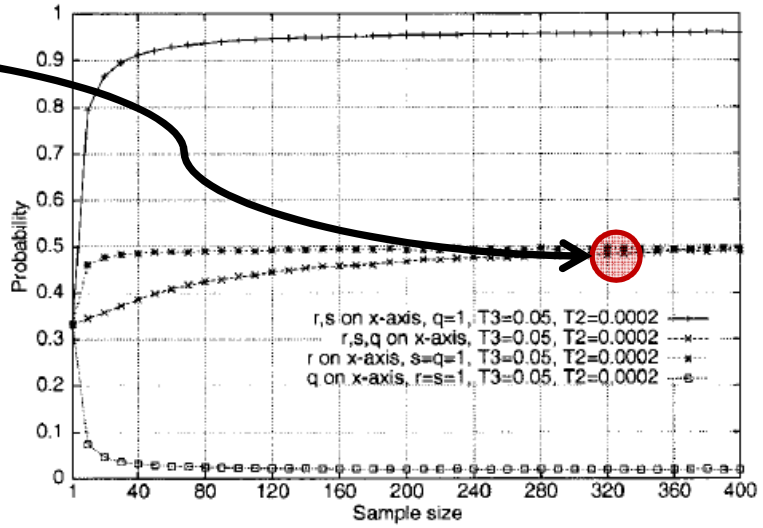
i



ii

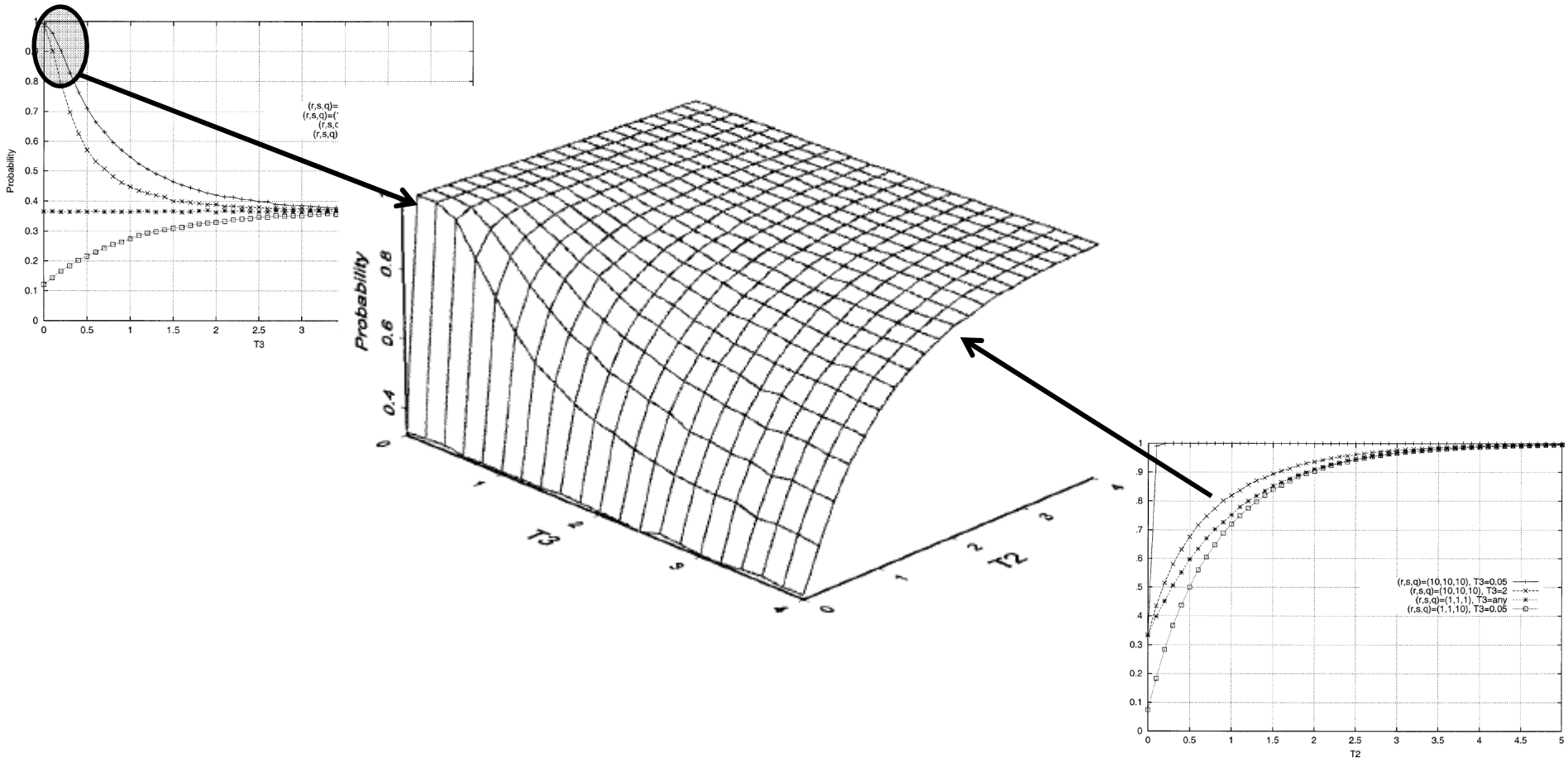






Probability of Speciodendricity

P(Speciodendricity) = P(Topological Concordance|Sample sizes = no. of copies of gene in respective species)
≈ P(Topological Concordance|Sample sizes = ∞)



Maximal Useful Sample Sizes

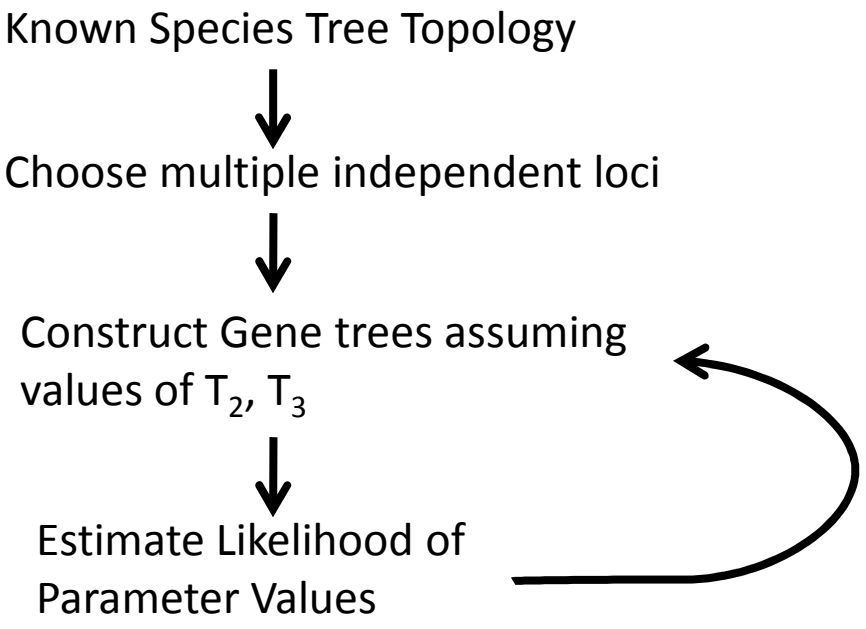
$$|E[A_{T_3} \mid A_0 = r] - E[A_{T_3} \mid A_0 = \infty]| < \varepsilon,$$

T3

Humans and Chimpanzees :
1.6 – 93.3
Humans and Neanderthals :
0.5 – 10
Modern Human Groups :
~0.05

T_3	Large-sample limiting mean number of ancestral lineages at time T_3	$\log_2(\varepsilon)$	Lower bound R	Mean number of ancestral lineages at time T_3 with a sample size of R
5	1.020	−3	1	1
2	1.418	−1	1	1
		−2	3	1.204
		−3	6	1.294
1	2.370	1	1	1
		0	3	1.577
		−1	6	1.879
		−2	14	2.126
		−3	30	2.248
0.05	40.335	4	61	24.474
		3	161	32.373
		2	361	36.345
		1	760	38.335
		0	1560	39.335
		−1	3160	39.835
		−2	6360	40.085
		−3	12759	40.210

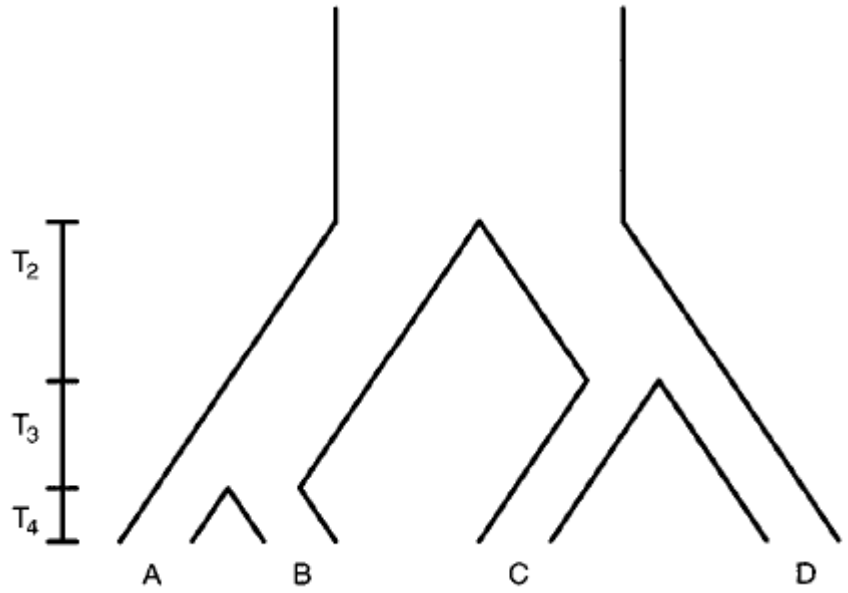
Estimation of Parameters T_2, T_3 and N



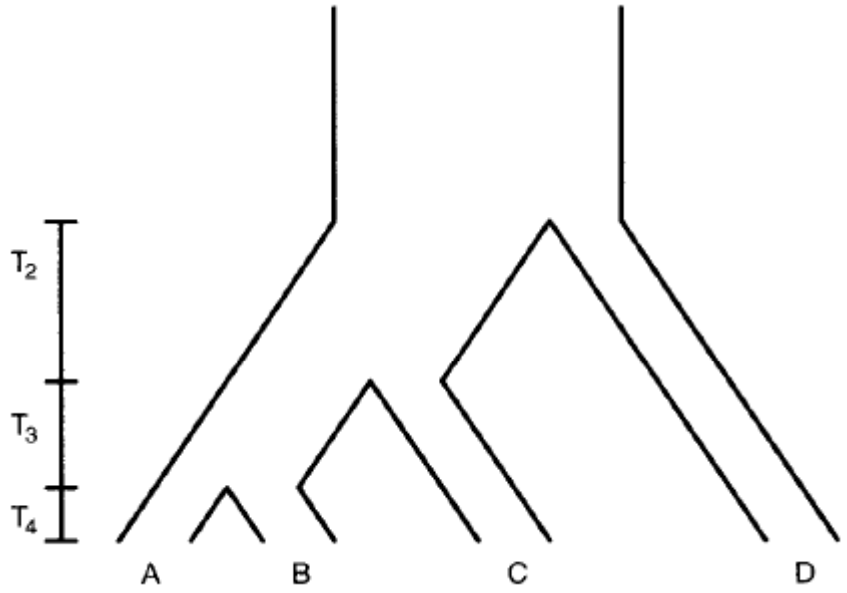
$$Lik(T_3, T_2) \propto P_C(r, s, q, T_3, T_2)^x \, Q_{((AC)\,B)}(r, s, q, T_3, T_2)^y \times Q_{((BC)\,A)}(r, s, q, T_3, T_2)^z. \quad (20)$$

Extension to Four or More Species

Balanced



Unbalanced



Topology Is ((*AB*)(*CD*))

Gene tree topology	Probability	Probability at $T_3 = T_2 = 1$
$((AB)(CD))$	$g_{21}(T_3+T_2)\,g_{21}(T_2)$ $+g_{21}(T_3+T_2)\,g_{22}(T_2)\,\frac{1}{3}$ $+g_{22}(T_3+T_2)\,g_{21}(T_2)\,\frac{1}{3}$ $+g_{22}(T_3+T_2)\,g_{22}(T_2)\,\frac{2}{6}\,\frac{1}{3}$	0.6867

Topology Is (((*AB*) *C*) *D*)

Gene tree topology	Probability	Probability at $T_3 = T_2 = 1$
$(((AB)\,C)\,D)$	$g_{21}(T_3)[g_{21}(T_2)+g_{22}(T_2)\,\frac{1}{3}]+g_{22}(T_3)$ $\times[g_{31}(T_2)\,\frac{1}{3}+g_{32}(T_2)\,\frac{1}{3}\,\frac{1}{3}+g_{33}(T_2)\,\frac{1}{6}\,\frac{1}{3}]$	0.5556

Summary

- Likelihood functions for observed gene tree conditioned on proposed species tree
- Inference of most likely Species History
- Estimation of optimal sample sizes that maximize concordance probability
- Estimation of divergence times and ancestral population sizes
- Identification of Speciation Genes
- Assumes equality and stability of population sizes
- Ignores gene exchange, mistaken orthology and other stochastic effects that cause discordance

References

- Rosenberg N A. *Theoretical Population Biology* **61**, 225-247 (2002).
- Takahata N. and Nei M. *Genetics* **110,** 325–344 (1985).
- Hudson R. R. *Evolution* **37,** 203–217 (1983).
- Takahata N. *Genetics* **122**, 957–966 (1989).

$g_{ij}(T)$

probability that two lineages coalesce in the immediately preceding generation = probability that they share a parent = 1/N

P_C(t generations) = $(1-1/N)^{t-1}(1/N) \approx \exp(-t/N)/N$ (for large N)

Let T = t/N

Probability that p lineages coalesce to p-1 at T

$f_{p-1}(T) = p*(p-1)*\exp(-T)/2$

T_{p-1} is the waiting time for p lineages to coalesce to p-1 with distribution f_{p-1}

Define $S_{ij} = \sum_{p=j}^{p=i-1} T_p$

Then P(i lineages converge to j in T) = $P(S_{ij} = T) = g_{ij}(T)$

(Hudson 1983, Takahata and Nei 1985)

$$g_{ij}(T) = \sum_{k=j}^{i} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!\,(k-j)!\, i_{(k)}},$$

where $a_{(k)} = a(a+1)\cdots(a+k-1)$ for $k \geqslant 1$ with $a_{(0)} = 1$;
and $a_{[k]} = a(a-1)\cdots(a-k+1)$ for $k \geqslant 1$ with $a_{[0]} = 1$.

$$g_{11}(T) = 1 \qquad g_{21}(T) = 1 - e^{-T} \qquad g_{22}(T) = e^{-T}$$

$$g_{31}(T) = 1 - \tfrac{3}{2}e^{-T} + \tfrac{1}{2}e^{-3T}$$
$$g_{32}(T) = \tfrac{3}{2}e^{-T} - \tfrac{3}{2}e^{-3T}$$
$$g_{33}(T) = e^{-3T}.$$

$F_k^{A,B}(m,n,l)$

probability that during the coalescence of m+n+l lineages from A, B and C, at least one interspecific coalescence occurs during this process, *and that the most recent interspecific coalescence joins a lineage from species A and a lineage from species B.*

$$F_k^{A,B}(a,b,c) = \frac{ab}{\binom{a+b+c}{2}} + F_k^{A,B}(a-1,b,c)\frac{\binom{a}{2}}{\binom{a+b+c}{2}} + F_k^{A,B}(a,b-1,c)\frac{\binom{b}{2}}{\binom{a+b+c}{2}} + F_k^{A,B}(a,b,c-1)\frac{\binom{c}{2}}{\binom{a+b+c}{2}}. \qquad (21)$$

$F_k^{A,B}(a,b,c)=0$ if $a+b+c \le k$.

Probability of Monophyletic Concordance

$$P_{M3}(r, s, q, T_3, T_2)$$

$$= \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{m+n} \sum_{l=1}^{q} \Bigg[g_{rm}(T_3)\, g_{sn}(T_3)\, g_{m+n,k}(T_2) \times g_{ql}(T_3+T_2) \Bigg[\delta_{k,1}[1-F_2^{A,B}(m,n,0)] \frac{2}{l(l+1)} + (1-\delta_{k,1})[1-F_k^{A,B}(m,n,0)] \times \sum_{x=1}^{k-1} \Bigg[W_{(m,n),(x,k-x)}(T_2)[1-F_3^{A,B}(x,k-x,l) - F_3^{A,C}(x,k-x,l) - F_3^{B,C}(x,k-x,l)] \frac{1}{3} \Bigg] \Bigg] \Bigg]. \quad (11)$$

$$P_{monophyly}(r, s, q, T_3, T_2)$$

$$= \sum_{m=1}^{r} \sum_{n=1}^{s} \sum_{k=1}^{m+n} \sum_{l=1}^{q} \Bigg[g_{rm}(T_3)\, g_{sn}(T_3)\, g_{m+n,k}(T_2) \times g_{ql}(T_3+T_2) \Bigg[\delta_{k,1}[1-F_2^{A,B}(m,n,0)] \frac{2}{l(l+1)} + (1-\delta_{k,1})[1-F_k^{A,B}(m,n,0)] \times \sum_{x=1}^{k-1} [W_{(m,n),(x,k-x)}(T_2)\, [1-F_3^{A,B}(x,k-x,l) - F_3^{A,C}(x,k-x,l) - F_3^{B,C}(x,k-x,l)]] \Bigg] \Bigg]. \quad (12)$$

(r, s, q)	T_3	T_2	Monophyly probability (all three species)	Monophyletic concordance probability (all three species)
(1,1,1)	Any value	0	1	0.333
		0.05	1	0.366
		0.5	1	0.596
		5	1	0.996
(2,2,1)	0.05	0	0.048	0.016
		0.05	0.056	0.019
		0.5	0.109	0.049
		5	0.134	0.133
	0.5	0	0.247	0.082
		0.05	0.260	0.092
		0.5	0.329	0.175
		5	0.355	0.353
	5	0	0.989	0.330
		0.05	0.989	0.362
		0.5	0.991	0.590
		5	0.991	0.987
(2,2,2)	0.05	0	0.011	0.004
		0.05	0.015	0.005
		0.5	0.059	0.028
		5	0.133	0.132
	0.5	0	0.124	0.041
		0.05	0.137	0.049
		0.5	0.234	0.127
		5	0.354	0.352
	5	0	0.983	0.328
		0.05	0.984	0.360
		0.5	0.987	0.589
		5	0.991	0.987
(5,5,1)	0.05	0	0.0002	0.00007
		0.05	0.0003	0.0001
		0.5	0.001	0.0005
		5	0.002	0.002
	0.5	0	0.031	0.010
		0.05	0.035	0.012
		0.5	0.066	0.030
		5	0.082	0.081
	5	0	0.978	0.326
		0.05	0.978	0.358
		0.5	0.981	0.584
		5	0.982	0.978