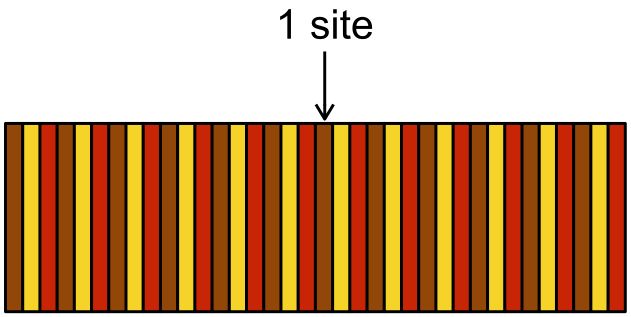
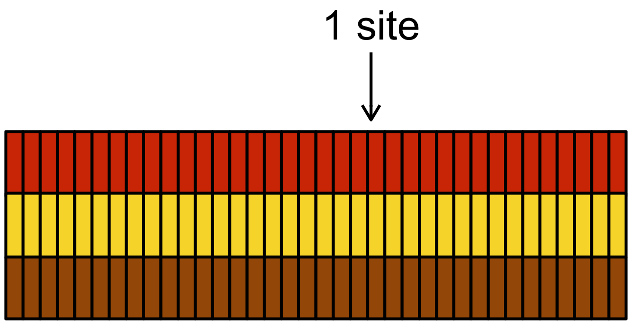
Two ways to model rate heterogeneity



site-specific rates

each site assigned to 1
of 3 rate categories

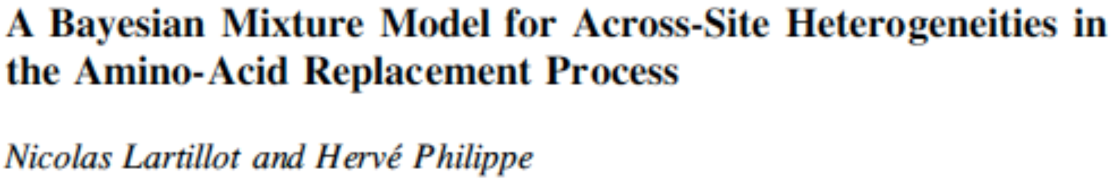


mixture model

each site has probability
1/3 of being in each of
the 3 rate categories

Dirichlet process priors provide a third option...

Dirichlet Process Priors

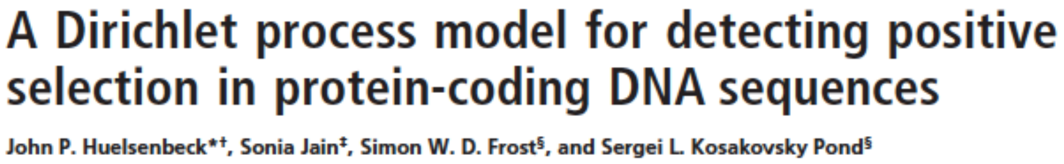


A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process

Nicolas Lartillot and Hervé Philippe

Lartillot and Philippe (2004)

PhyloBayes
http://www.atgc-montpellier.fr/phylobayes

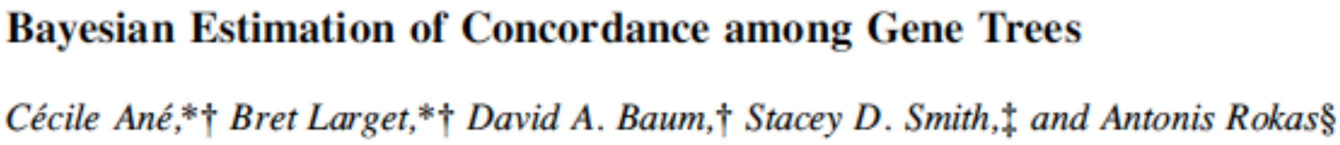


A Dirichlet process model for detecting positive selection in protein-coding DNA sequences

John P. Huelsenbeck[*†], Sonia Jain[‡], Simon W. D. Frost[§], and Sergei L. Kosakovsky Pond[§]

Huelsenbeck et al. (2006)

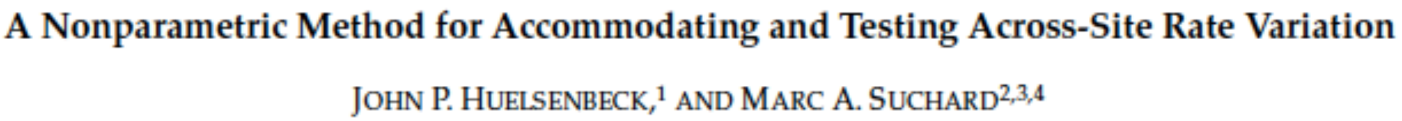
BUCKy
http://www.stat.wisc.edu/~ane/bucky/



Bayesian Estimation of Concordance among Gene Trees

Cécile Ané,[†] Bret Larget,[*†] David A. Baum,[†] Stacey D. Smith,[‡] and Antonis Rokas[§]*

Ané et al (2007)

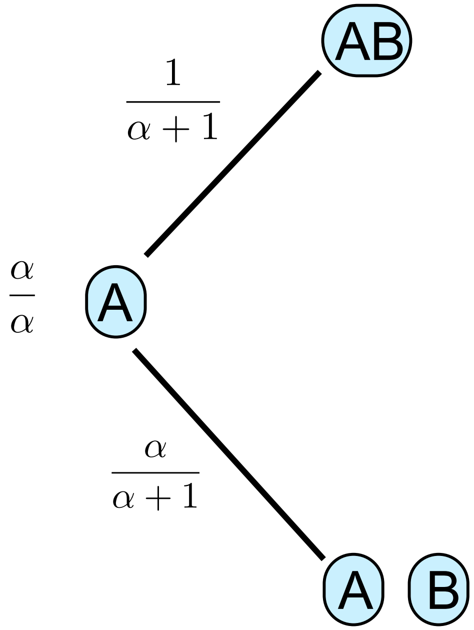


A Nonparametric Method for Accommodating and Testing Across-Site Rate Variation

JOHN P. HUELSENBECK,[1] AND MARC A. SUCHARD[2,3,4]

Huelsenbeck and Suchard (2007)

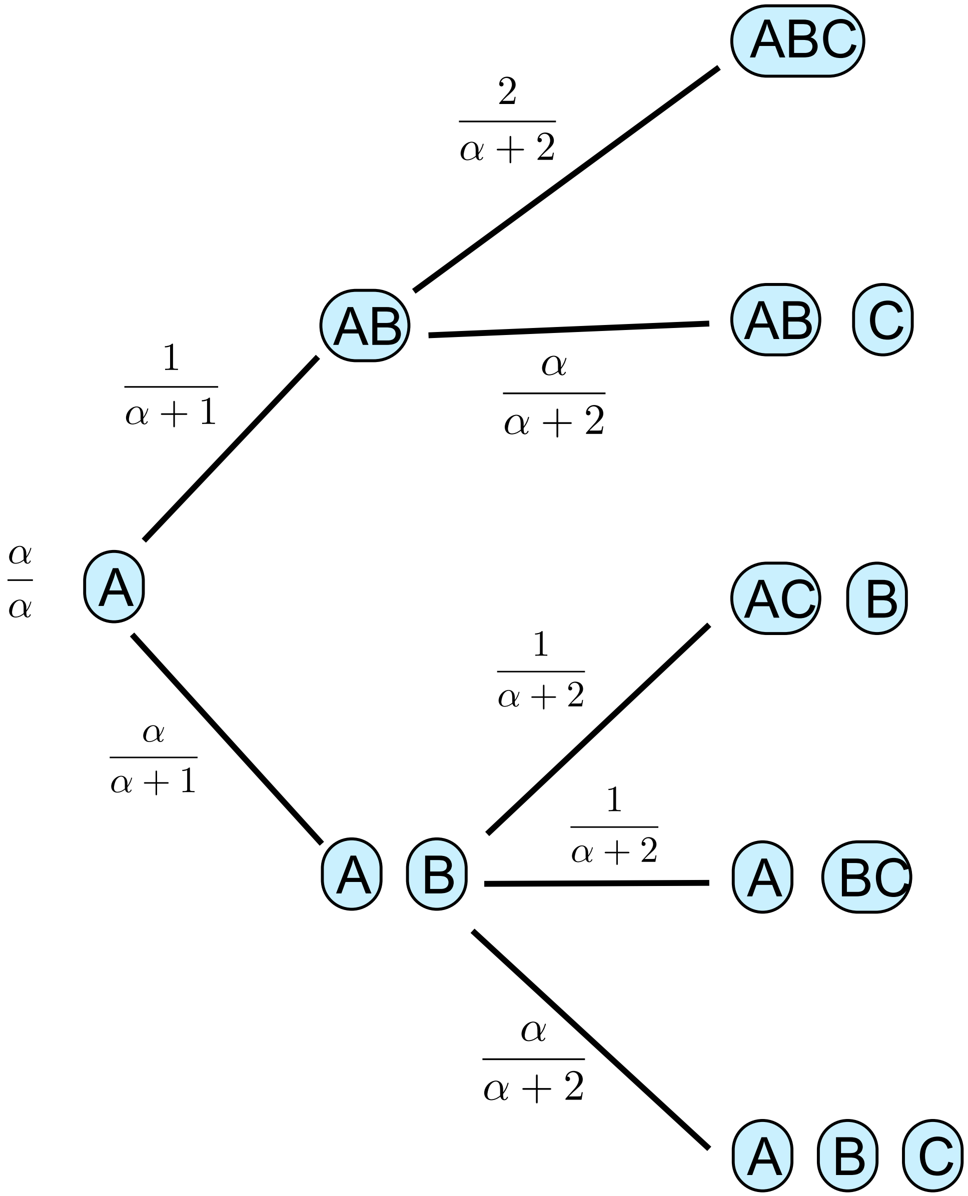
Dirichlet Process Prior



Imagine you have a collection of objects (e.g. sites) labeled A, B, C, ...

B can either be added to A’s group or form its own group

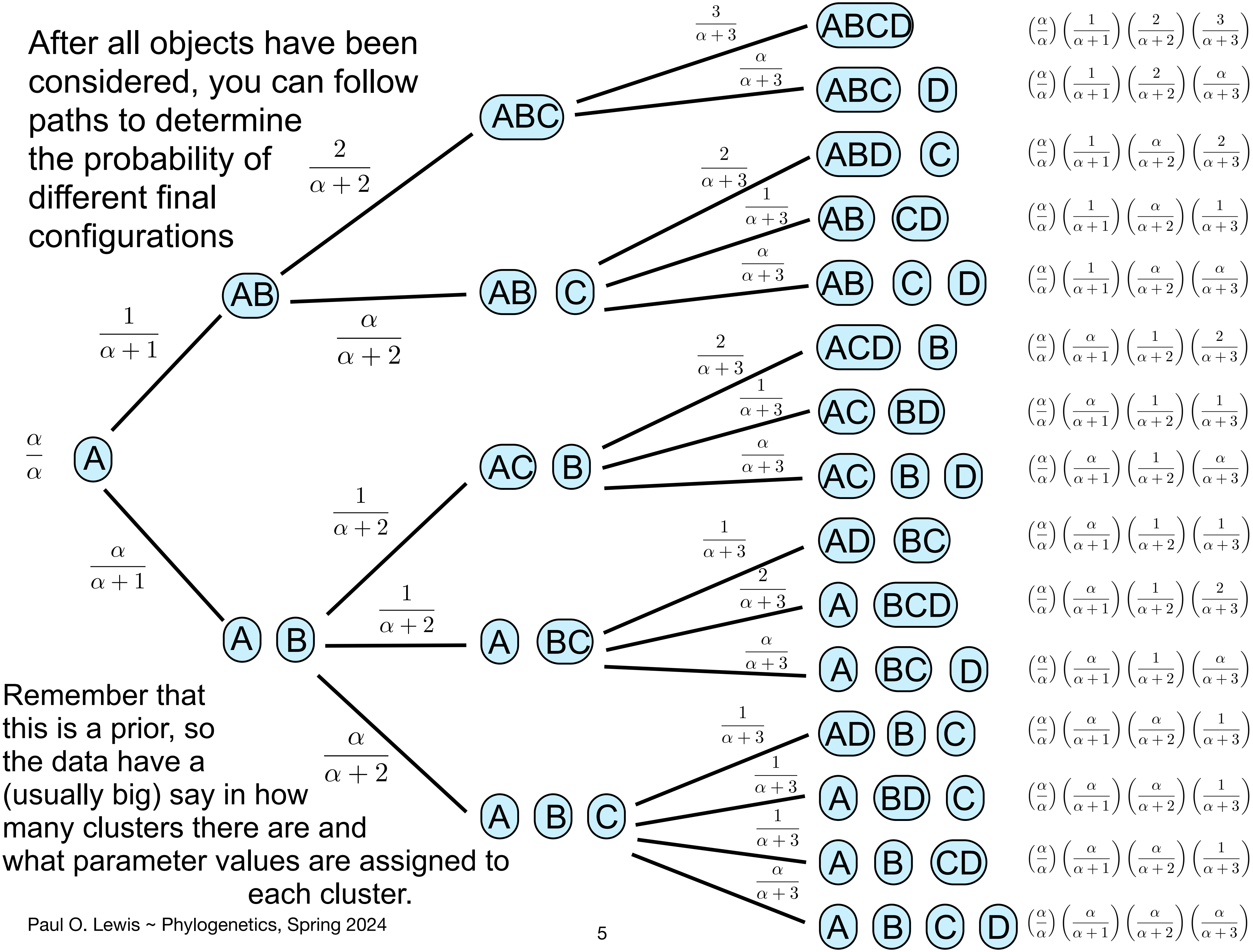
The parameter α determines the propensity for forming a new group



The third object C can either be added to an existing group...

...or form its own group

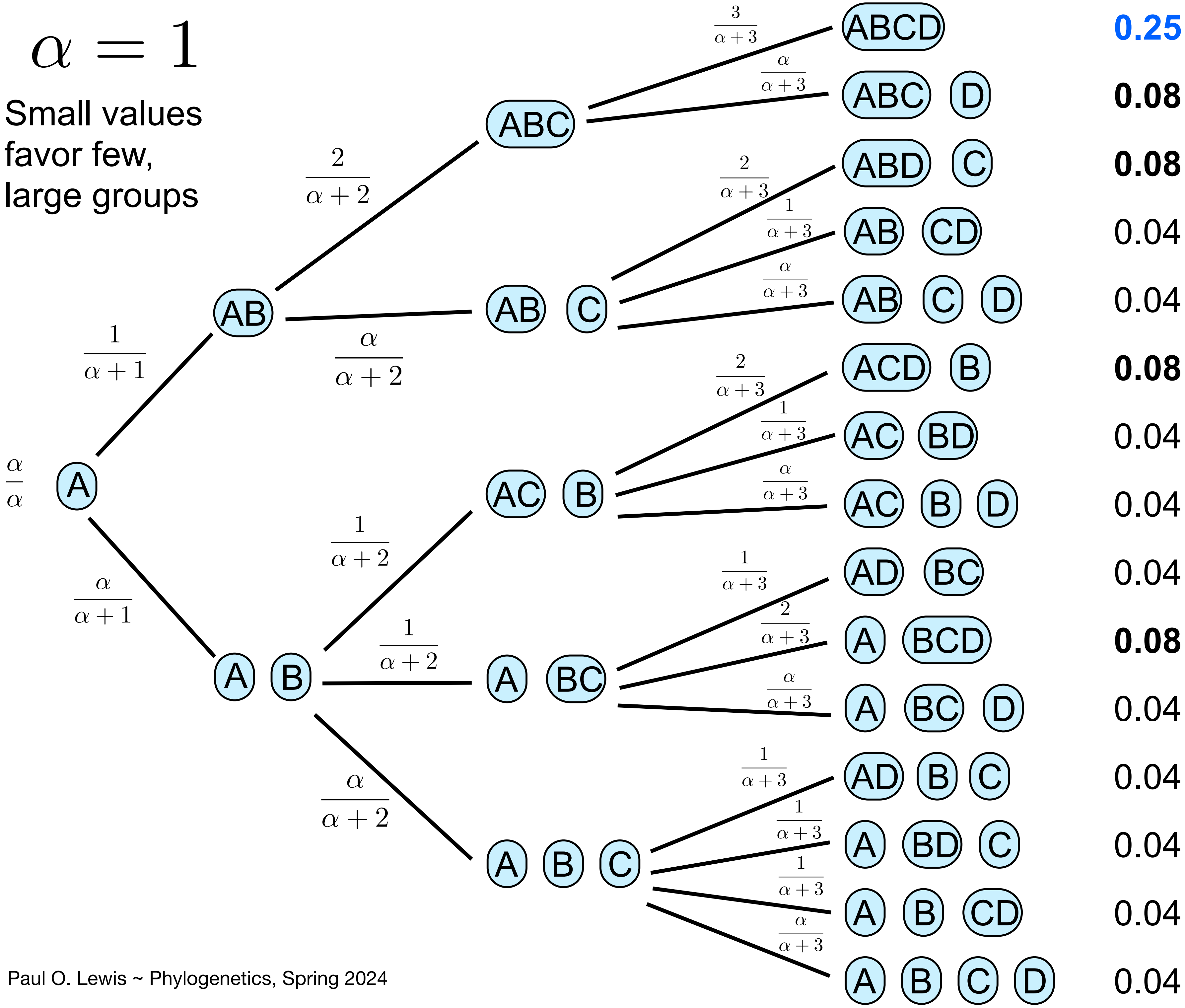
After all objects have been considered, you can follow paths to determine the probability of different final configurations


$\frac{\alpha}{\alpha}$
A
$\frac{1}{\alpha+1}$
AB
$\frac{\alpha}{\alpha+1}$
A B
$\frac{2}{\alpha+2}$
ABC
$\frac{\alpha}{\alpha+2}$
AB C
$\frac{1}{\alpha+2}$
AC B
$\frac{1}{\alpha+2}$
A BC
$\frac{\alpha}{\alpha+2}$
A B C
$\frac{3}{\alpha+3}$
ABCD
$\frac{\alpha}{\alpha+3}$
ABC D
$\frac{2}{\alpha+3}$
ABD C
$\frac{1}{\alpha+3}$
AB CD
$\frac{\alpha}{\alpha+3}$
AB C D
$\frac{2}{\alpha+3}$
ACD B
$\frac{1}{\alpha+3}$
AC BD
$\frac{\alpha}{\alpha+3}$
AC B D
$\frac{1}{\alpha+3}$
AD BC
$\frac{2}{\alpha+3}$
A BCD
$\frac{\alpha}{\alpha+3}$
A BC D
$\frac{1}{\alpha+3}$
AD B C
$\frac{1}{\alpha+3}$
A BD C
$\frac{1}{\alpha+3}$
A B CD
$\frac{\alpha}{\alpha+3}$
A B C D
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right)\left(\frac{3}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{1}{\alpha+1}\right)\left(\frac{2}{\alpha+2}\right)\left(\frac{\alpha}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{2}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{1}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{1}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{\alpha}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)\left(\frac{2}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)\left(\frac{1}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)\left(\frac{\alpha}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)\left(\frac{1}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)\left(\frac{2}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{1}{\alpha+2}\right)\left(\frac{\alpha}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{1}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{1}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{1}{\alpha+3}\right)$
$\left(\frac{\alpha}{\alpha}\right)\left(\frac{\alpha}{\alpha+1}\right)\left(\frac{\alpha}{\alpha+2}\right)\left(\frac{\alpha}{\alpha+3}\right)$

Remember that this is a prior, so the data have a (usually big) say in how many clusters there are and what parameter values are assigned to each cluster.

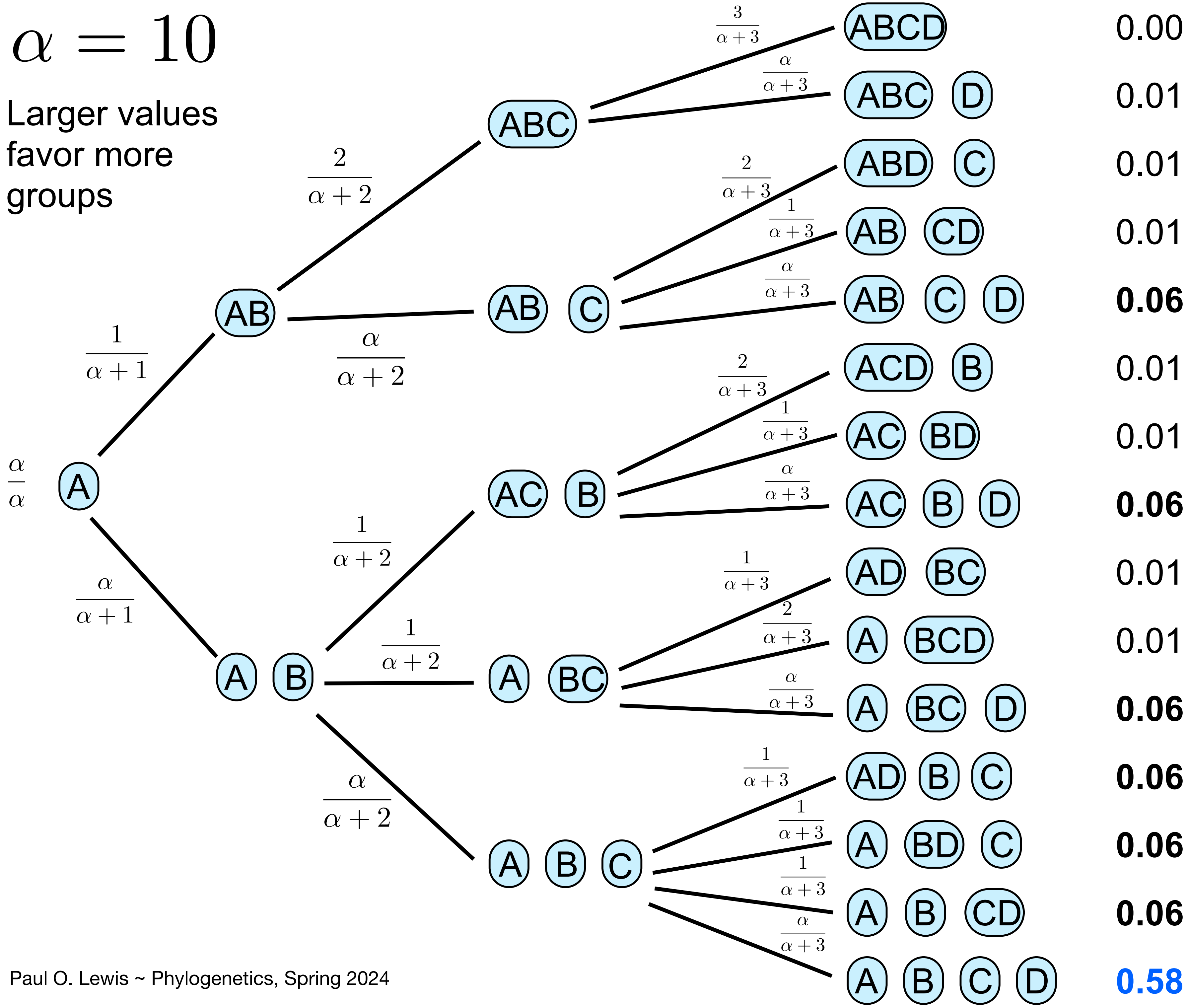
$$\alpha = 1$$

Small values favor few, large groups



$$\alpha = 10$$

Larger values favor more groups



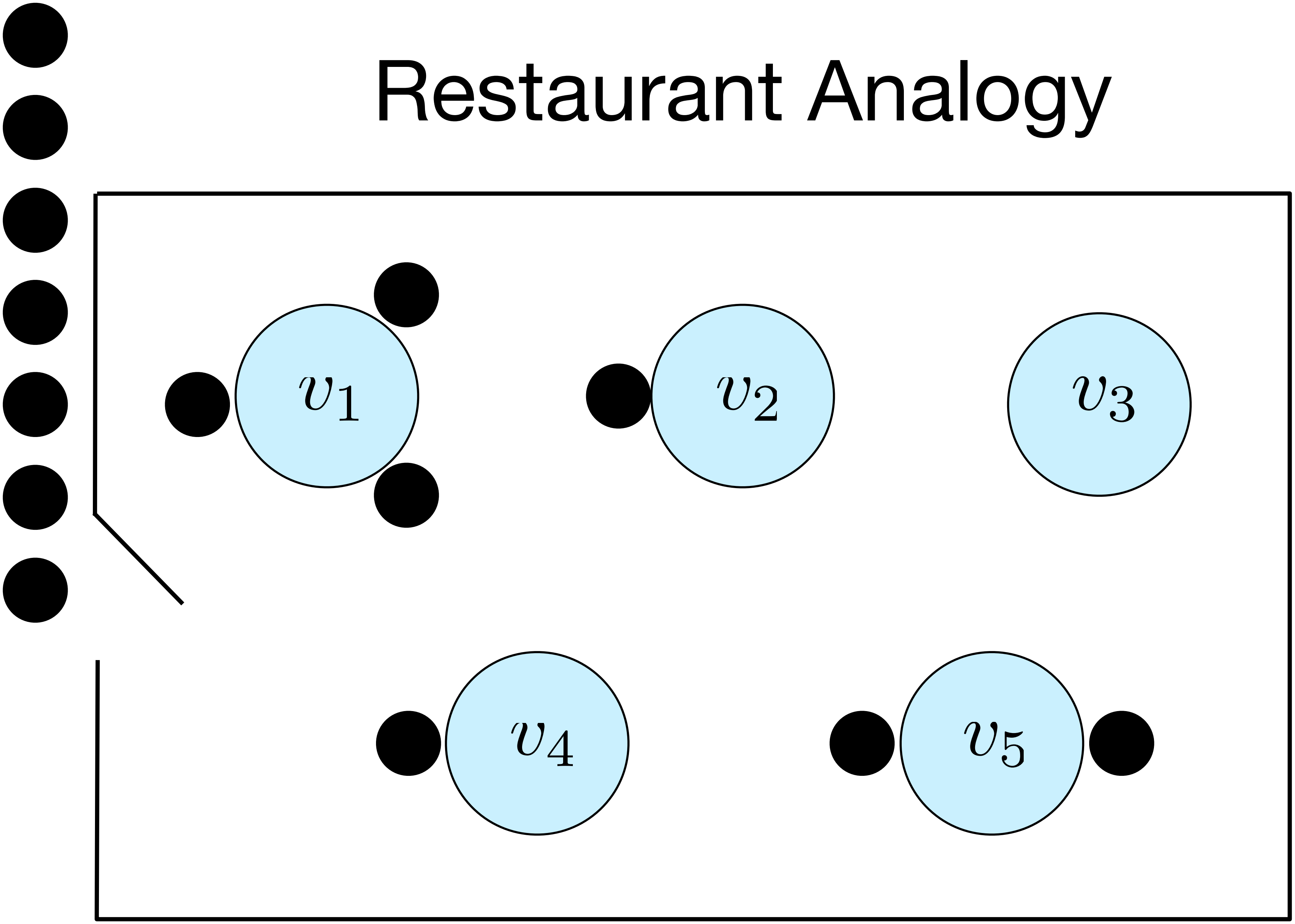
Expected Number of Groups

$$\sum_{i=0}^{n-1} \frac{\alpha}{\alpha + i}$$

For example, if *n* = 3 and $\boldsymbol{\alpha}$ = 1:

$$\frac{\alpha}{\alpha+0} + \frac{\alpha}{\alpha+1} + \frac{\alpha}{\alpha+2} = 1 + \frac{1}{2} + \frac{1}{3} = 1.83$$

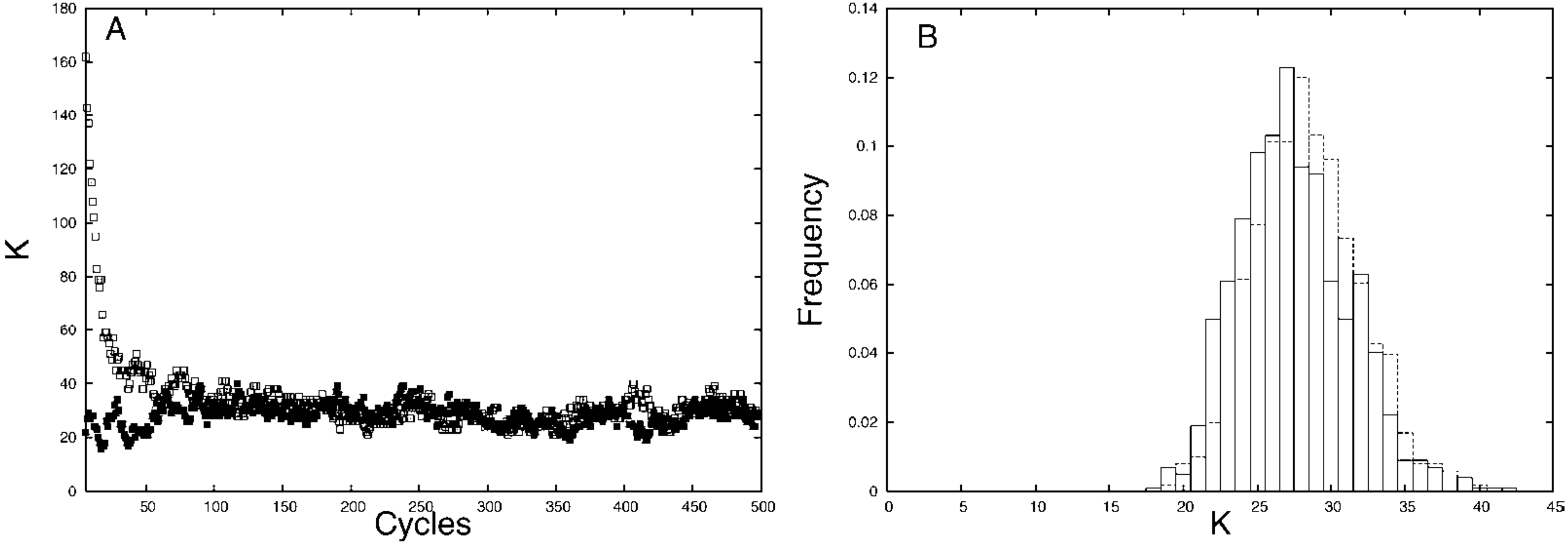
Restaurant Analogy



Dirichlet process prior applet

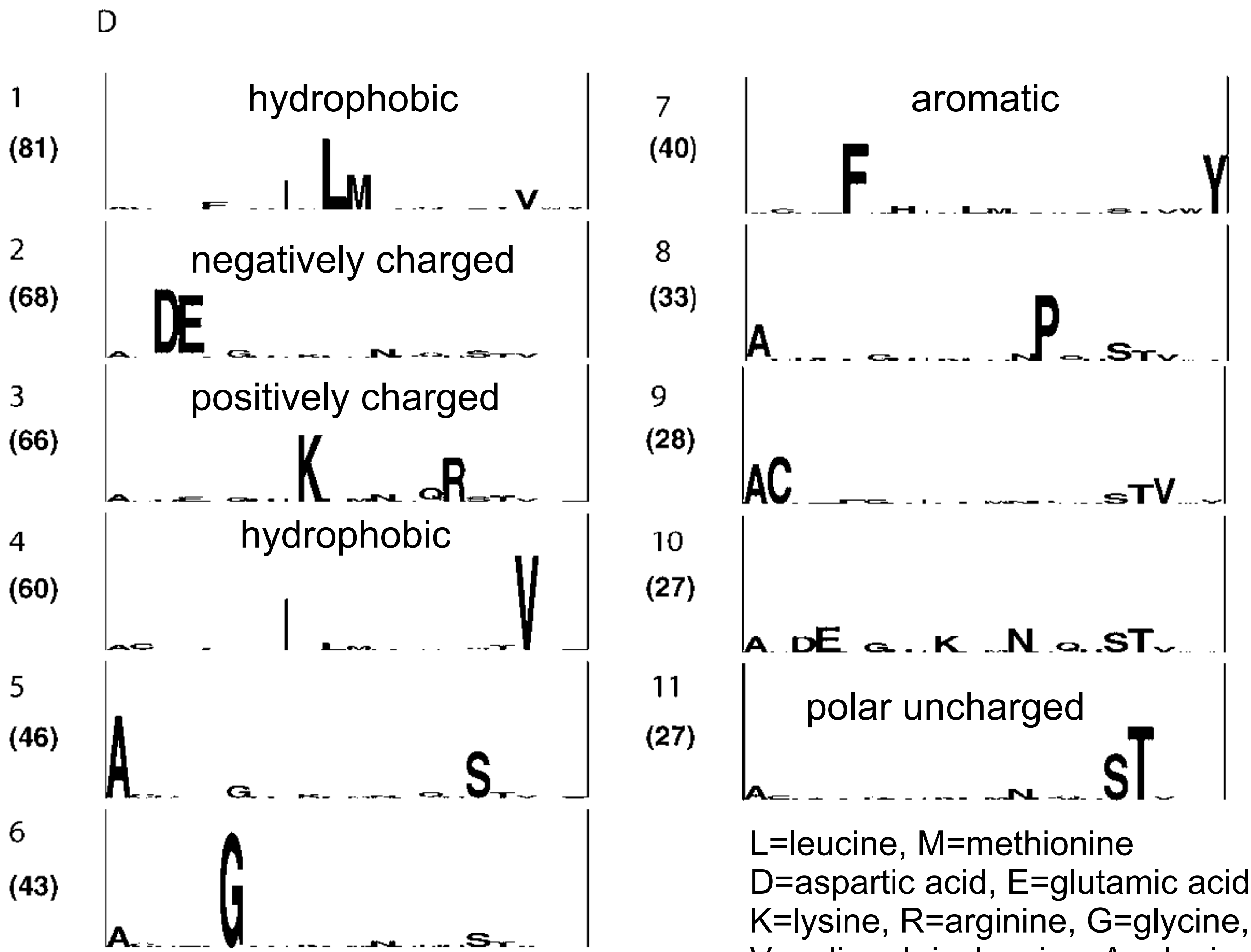
https://plewis.github.io/applets/dpp/

Example 1:Elongation factor 2 (software: PhyloBayes)



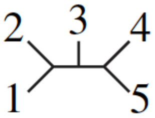
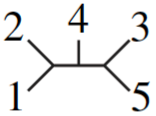
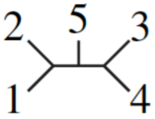
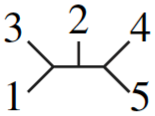
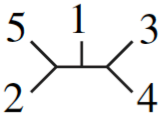
This model is implemented in PhyloBayes software: http://www.atgc-montpellier.fr/phylobayes/

Lartillot and Philippe (2004)



L=leucine, M=methionine
D=aspartic acid, E=glutamic acid
K=lysine, R=arginine, G=glycine,
V=valine, I=isoleucine, A=alanine,
F=phenylalanine, Y=tyrosine
S=serine, T=threonine

Example 2: Bayesian concordance analyses (software: BUCKy)

		m_1			m_2		
tree label	tree	g_1	g_2	g_3	g_1	g_2	g_3
1		0	0	0	0	0	0
2		**1**	**1**	**1**	**1**	**1**	0
3		0	0	0	0	0	**1**
4		0	0	0	0	0	0
⋮	⋮		⋮			⋮	
15		0	0	0	0	0	0
			(2,2,2)			(2,2,3)	

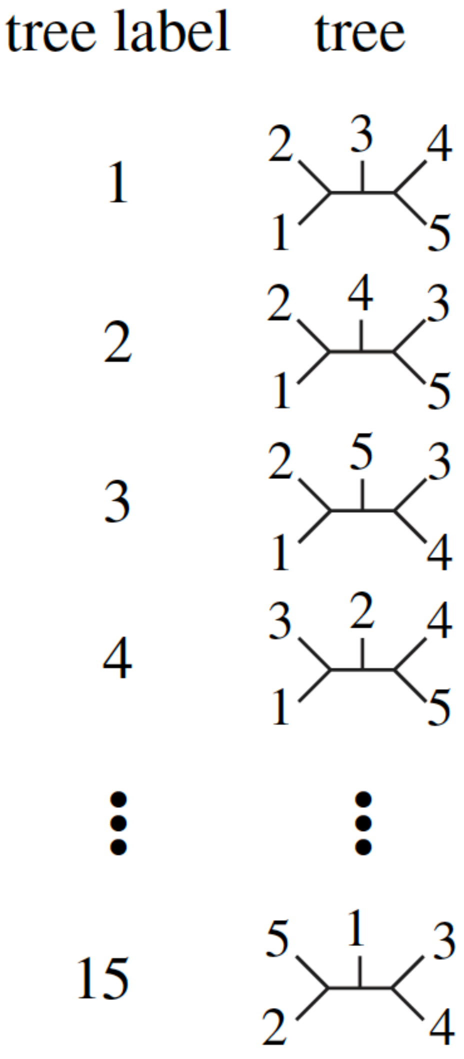
← Gene-to-Tree Mapping (GTM)

Marginal posterior distributions for each gene separately

	gene		
tree	1	2	3
2	1	0	0
3	0	0.9	0.2
4	0	0.1	0.2
15	0	0	0.6

Ané et al. (2007)

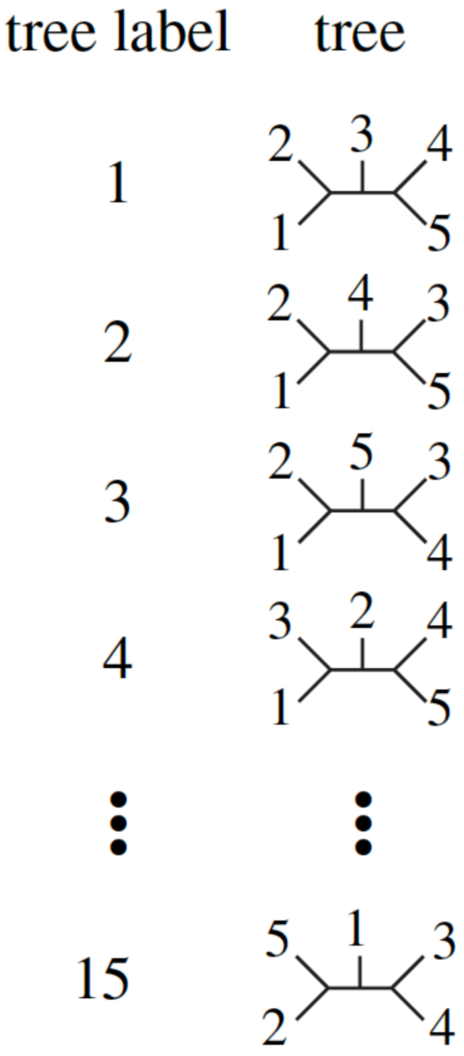
Concordance Factors (CF)



GTM	Posterior	K	CF 12|345
(2,3,3)	0.6600		
(2,3,4)	0.0600		
(2,3,15)	0.1800		
(2,4,3)	0.0067		
(2,4,4)	0.0733		
(2,4,15)	0.0200		

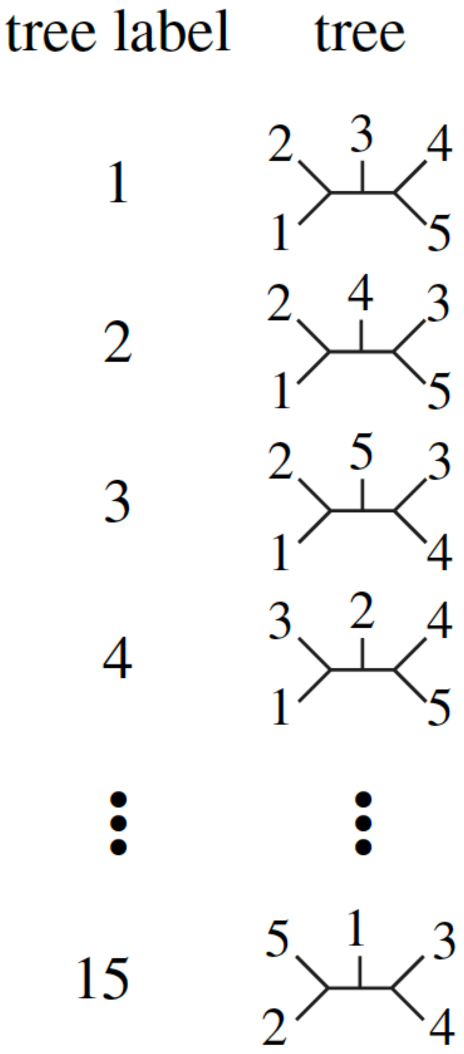
1.0000

Concordance Factors (CF)



GTM	Posterior	K	CF 12|345
(2,3,3)	0.6600	2	
(2,3,4)	0.0600	3	
(2,3,15)	0.1800	3	
(2,4,3)	0.0067	3	
(2,4,4)	0.0733	2	
(2,4,15)	0.0200	3	
	1.0000	2.27	

Concordance Factors (CF)



GTM	Posterior	K	CF 12|345
(2,3,3)	0.6600	2	1
(2,3,4)	0.0600	3	2/3
(2,3,15)	0.1800	3	2/3
(2,4,3)	0.0067	3	2/3
(2,4,4)	0.0733	2	1/3
(2,4,15)	0.0200	3	1/3
	1.0000	2.27	0.86

Dirichlet Process Priors

- To encourage **few, large** groups, use a **small** alpha value
- To encourage **lots of small** groups, use a **large** alpha value
- In practice, **hierarchical models** are used (i.e. alpha is a hyperparameter that can be estimated, so you need not worry about choosing the appropriate value for alpha)

- Bottom line: DP models are very nice for automatically grouping sites into clusters that have some property in common