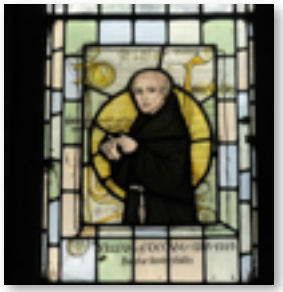
Basically, we have three different means to reconstruct phylogenetic trees from sequence data



Find tree that requires the least number of changes

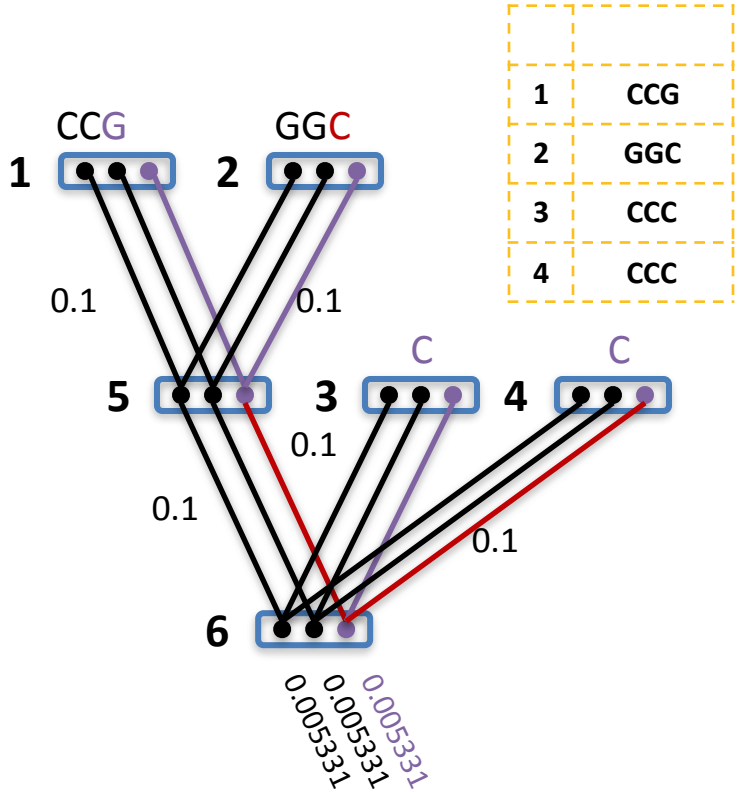
Data	Method	Evaluation Criterion
Characters (Alignment)	Maximum Parsimony	Parsimony
	Statistical Approaches: Likelihood, Bayesian	Evolutionary Models
Distances	Distance Methods	

Find the tree that most likely gave rise to the data

Reconstruct the best fitting tree from a pair-wise distance matrix[1]

[1] see Grundlagen der Bioinformatik, Lecture 12

Calculating tree likelihoods



1	CCG
2	GGC
3	CCC
4	CCC

For an alignment of four sequences and length m=3 the likelihood is then

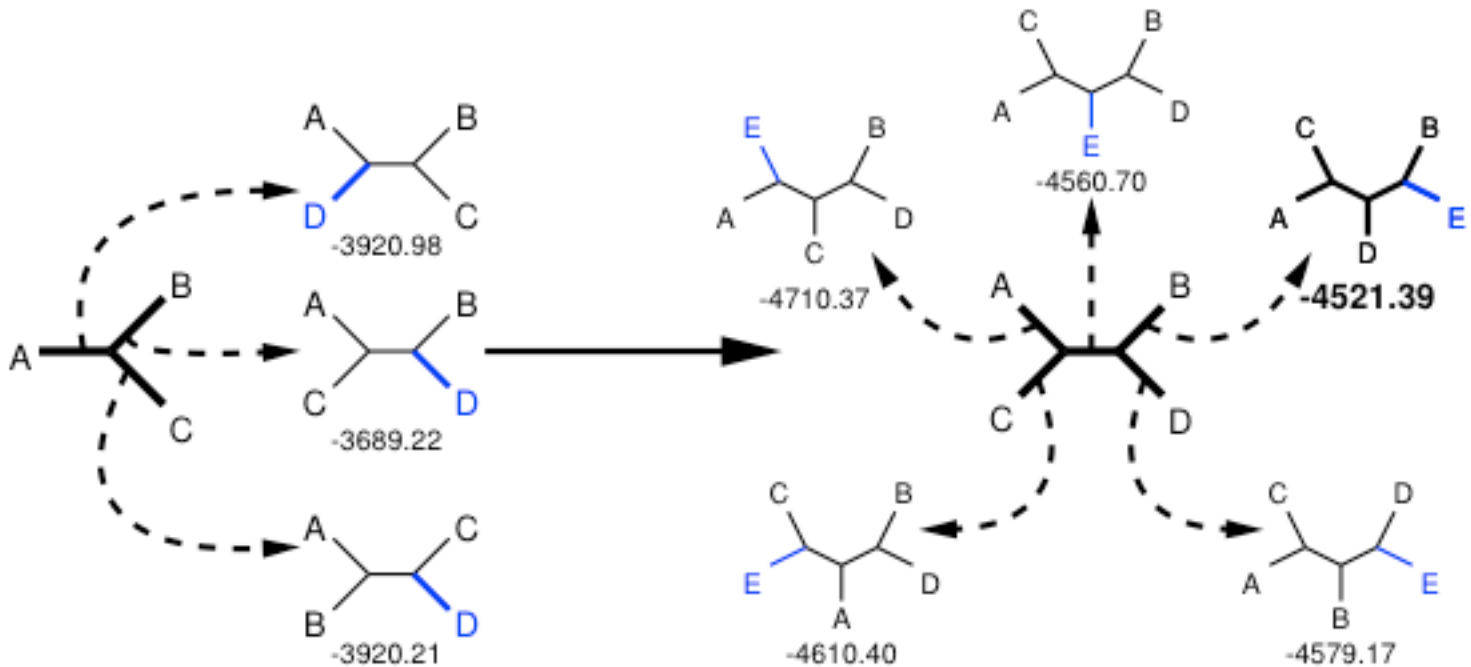
$$L(T) = \prod_{k=1}^{m} L^{(k)} = 0.005331^2 \times 0.005331$$
$$= 0.000000152$$

or the log-likelihood is

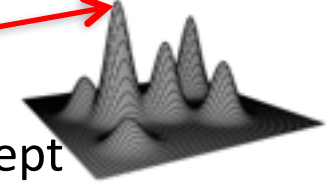
$$\ln L(T) = \sum_{k=1}^{m} \ln L^{(k)} = -15.7$$

Now that we know how to evaluate the likelihood of any given tree, we need to ask how to find the ML tree

Heuristic tree search begins with an initial sub-optimal solution (starting tree) obtained either via step-wise addition (or using a distance tree)

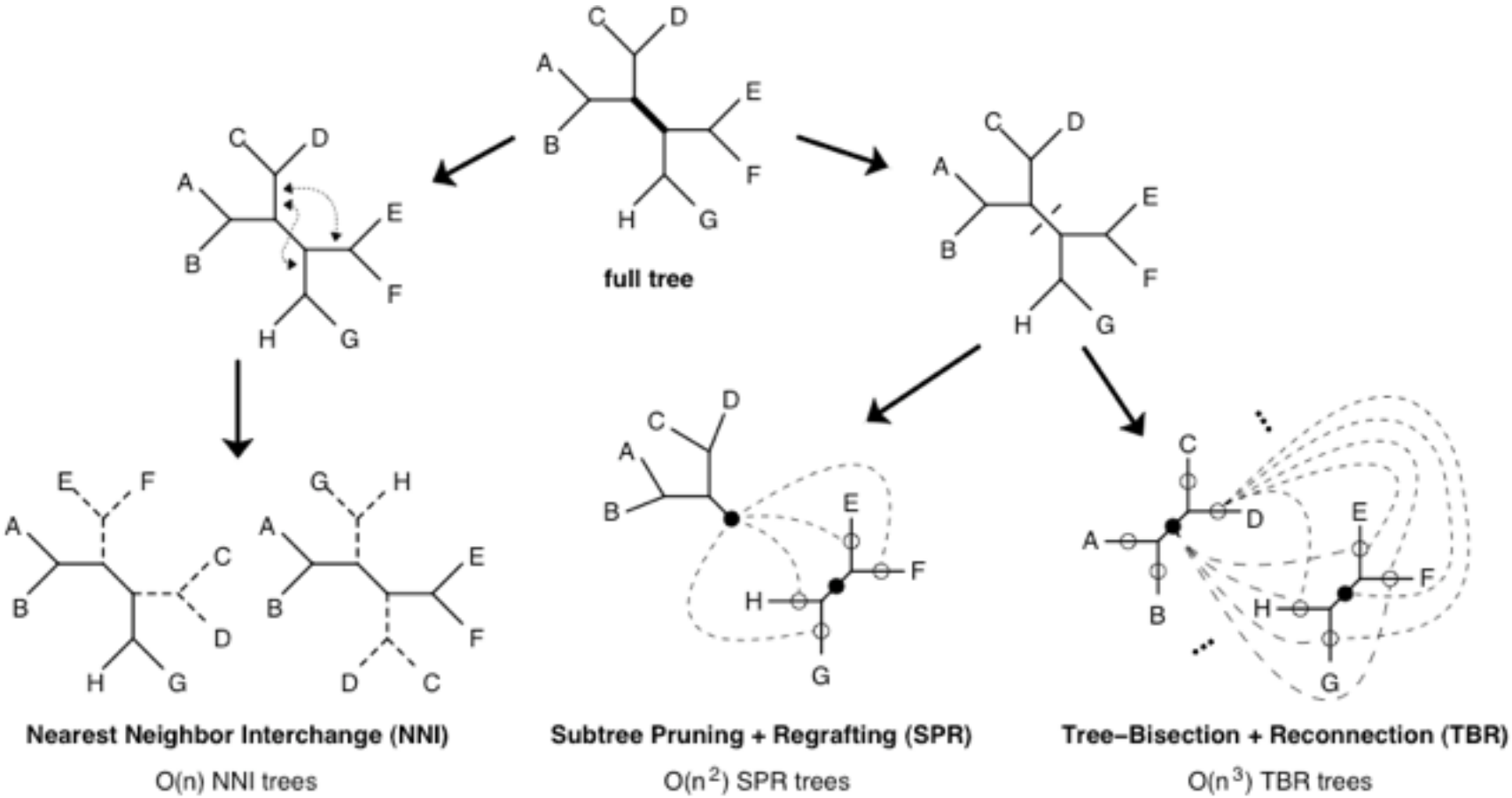


our goal!



Finding the best tree

Evaluate random rearrangements of the starting tree and accept new tree if it improves P(D|M,T). Continue until convergence.

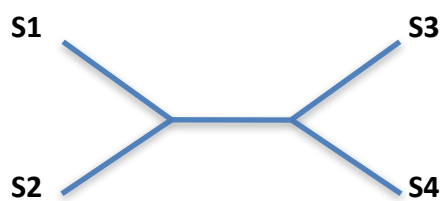


Again we have an iterative stochastic process as we have seen in the alignment case

Resampling methods for assessing the support of a (ML[1]) tree given the data

Rationale: All positions in a sequence, and hence all alignment columns, should have the same evolutionary history. Thus, we can summarize the phylogenetic information in a single tree.

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T

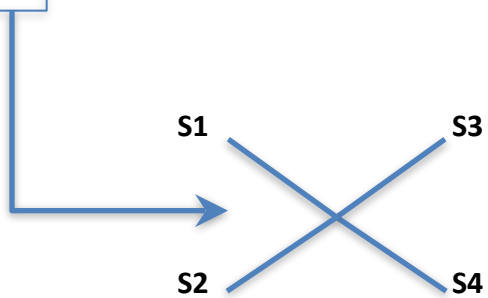


[1] works of course for maximum parsimony and distance trees as well.

Resampling methods for assessing the support of a tree given the data

Rationale: All positions in a sequence, and hence all alignment columns, should have the same evolutionary history. Thus, it should in principle not matter which subset of the data I am using for tree reconstruction if the phylogenetic signal is sufficiently strong <u>and</u> indeed consistent.

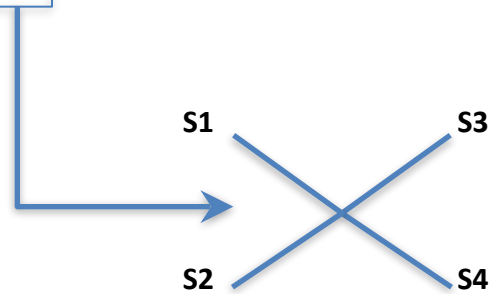
Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T



Resampling methods for assessing the support of a tree given the data

Rationale: All positions in a sequence, and hence all alignment columns, should have the same evolutionary history. Thus, it should in principle not matter which subset of the data I am using for tree reconstruction if the phylogenetic signal is sufficiently strong <u>and</u> indeed consistent.

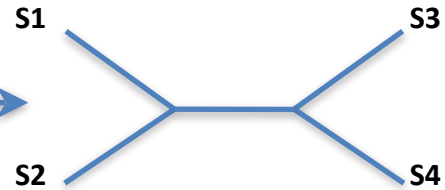
Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T



Resampling methods for assessing the support of a tree given the data

Rationale: All positions in a sequence, and hence all alignment columns, should have the same evolutionary history. Thus, it should in principle not matter which subset of the data I am using for tree reconstruction if the phylogenetic signal is sufficiently strong <u>and</u> indeed consistent.

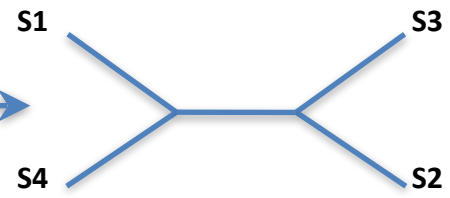
Taxon	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T



Resampling methods for assessing the support of a tree given the data

Rationale: All positions in a sequence, and hence all alignment columns, should have the same evolutionary history. Thus, it should in principle not matter which subset of the data I am using for tree reconstruction if the phylogenetic signal is sufficiently strong <u>and</u> indeed consistent.

Taxon	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T



Resampling methods for assessing the support of a tree given the data

Rationale: All positions in a sequence, and hence all alignment columns, should have the same evolutionary history. Thus, it should in principle not matter which subset of the data I am using for tree reconstruction if the phylogenetic signal is sufficiently strong <u>and</u> indeed consistent.

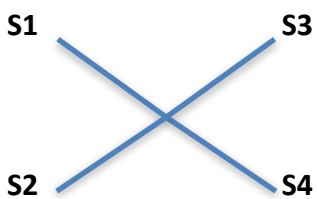
Taxon	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T

Resampling methods for assessing the support of a tree given the data

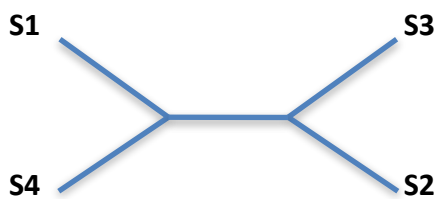
Observation: The phylogenetic signal in the data is apparently not entirely consistent and we would like to have a method to assess the extent of variability.

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T

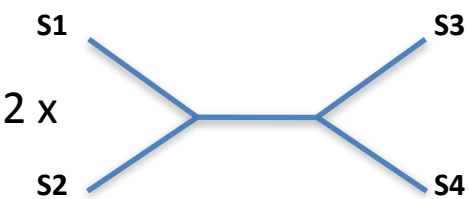
6 x



1 x



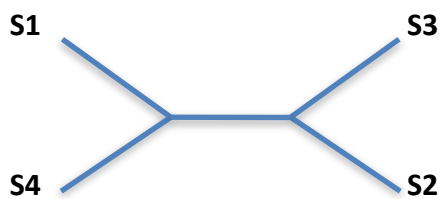
2 x



Resampling methods for assessing the support of a tree given the data

Approach 1 – Jackknife: Remove a random subset of alignment columns and re-compute the tree. Typically a 50% Jackknife analysis is performed.

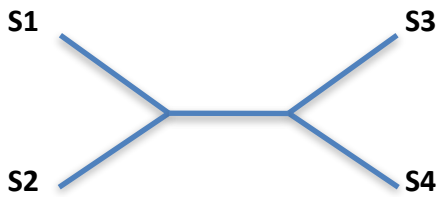
Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T



Resampling methods for assessing the support of a tree given the data

Approach 1 – Jackknife: Remove a random subset of alignment columns and re-compute the tree. Typically a 50% Jackknife analysis is performed.

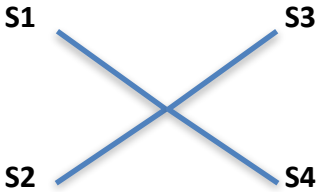
Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T



Resampling methods for assessing the support of a tree given the data

Approach 1 – Jackknife: Remove a random subset of alignment columns and re-compute the tree. Typically a 50% Jackknife analysis is performed.

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T

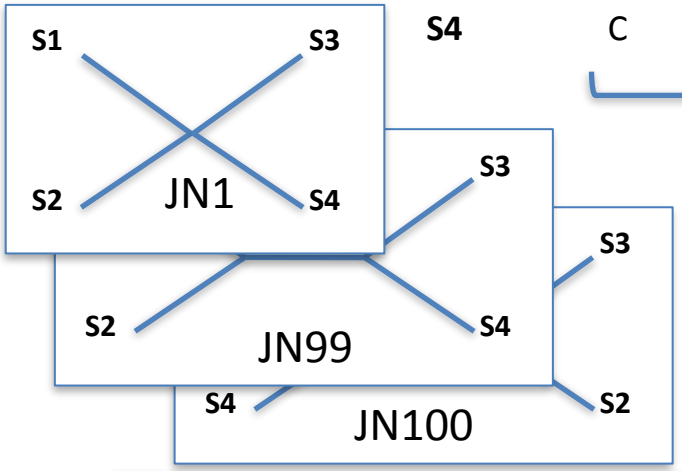


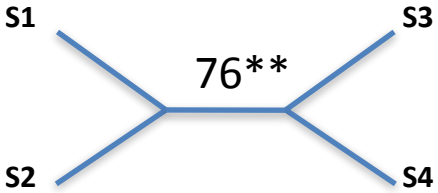
Resampling methods for assessing the support of a tree given the data

Approach 1 – Jackknife: Remove a random subset of alignment columns and re-compute the tree. Typically a 50% Jackknife analysis is performed.

repeat n* times

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T





*n is typically 100 or 1000 **value is typically given in percent

Resampling methods for assessing the support of a tree given the data

Approach 2 – Bootstrap: Resample randomly chosen columns from the original alignment (with replacement) to obtain a new alignment with the <u>same length</u> as the original alignment.

repeat n* times

Taxon	7	7	9	8	5	6	7	1	2
S1	G	G	T	T	C	T	G	C	G
S2	C	C	T	T	C	T	C	C	G
S3	G	G	T	C	C	T	G	T	G
S4	G	G	T	C	C	T	G	C	G

Taxon	1	1	4	4	7	7	1	5	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T

Taxon	4	4	4	4	4	4	4	4	4
S1	G	G	G	G	G	G	G	G	G
S2	A	A	A	A	A	A	A	A	A
S3	A	A	A	A	A	A	A	A	A
S4	G	G	G	G	G	G	G	G	G

Taxon	6	5	2	9	6	1	6	8	9
S1	T	C	G	T	T	C	T	T	T
S2	T	C	G	T	T	C	T	T	T
S3	T	C	G	T	T	T	T	C	T
S4	T	C	G	T	T	C	T	C	T

*n is typically 100 or 1000

Resampling methods for assessing the support of a tree given the data

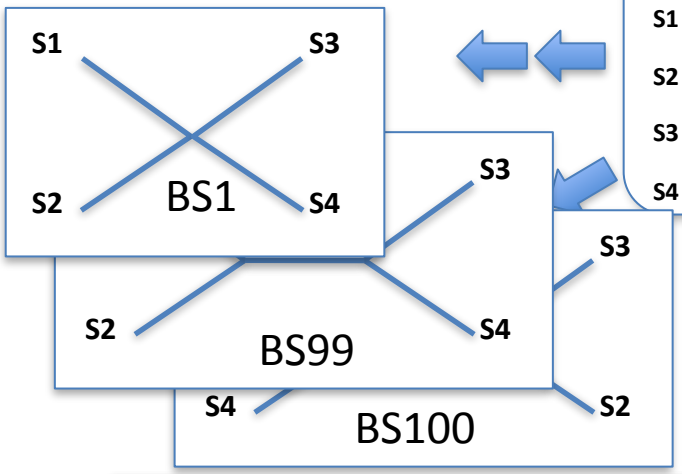
Approach 2 – Bootstrap: Resample randomly chosen columns from the original alignment (with replacement) to obtain a new alignment with the <u>same length</u> as the original alignment.

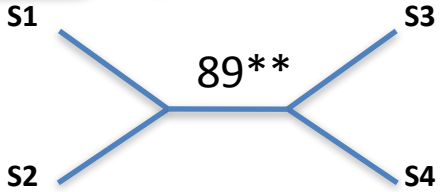
repeat n* times

Taxon	7	7	9	8	5	6	7	1	2
S1	G	G	T	T	C	T	G	C	G
S2	C	C	T	T	C	T	C	C	G
S3	G	G	T	C	C	T	G	T	G
S4	G	G	T	C	C	T	G	C	G

Taxon	1	1	4	4	7	7	1	5	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T

Taxon	1	2	3	4	5	6	7	8	9
S1	C	G	C	G	C	T	G	T	T
S2	C	G	C	A	C	T	C	T	T
S3	T	G	A	A	C	T	G	C	T
S4	C	G	A	G	C	T	G	C	T





*n is typically 100 or 1000 **value is typically given in percent

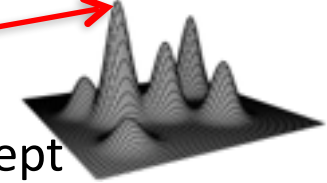
Maximum Parsimony and Maximum Likelihood only evaluate trees and do not reconstruct them!
Finding the best tree is highly problematic!

1. **Exhaustive Search:** evaluates every possible tree and hence an optimal solution is guaranteed. Limit: 10-12 taxa

2. **Branch and Bound:** excludes parts from the tree space from the search where the optimal tree cannot be found. Guarantees to find the optimal tree.

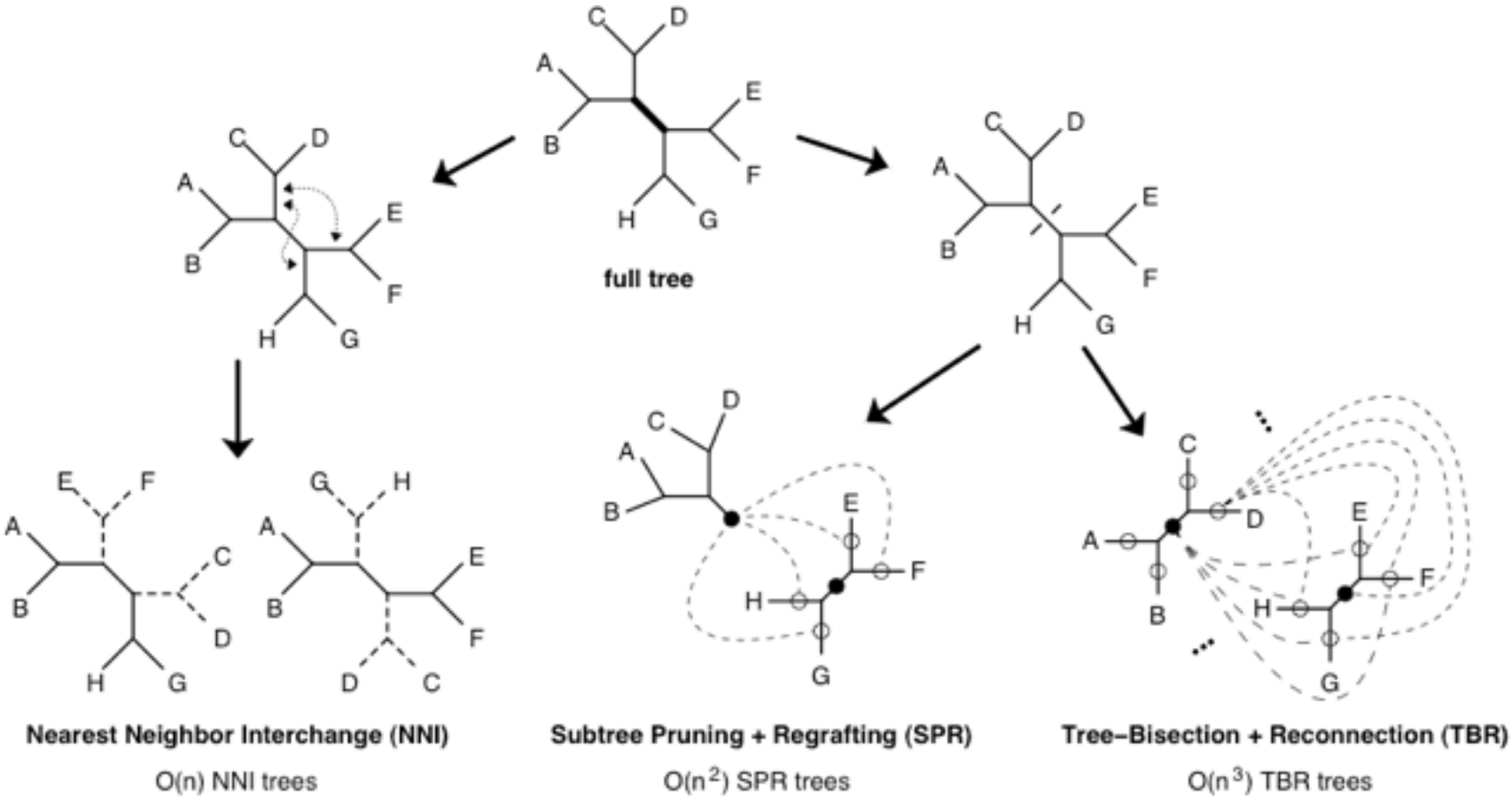
3. **Heuristics:** Can be applied to large taxon sets but does not guarantee an optimal solution

our goal!



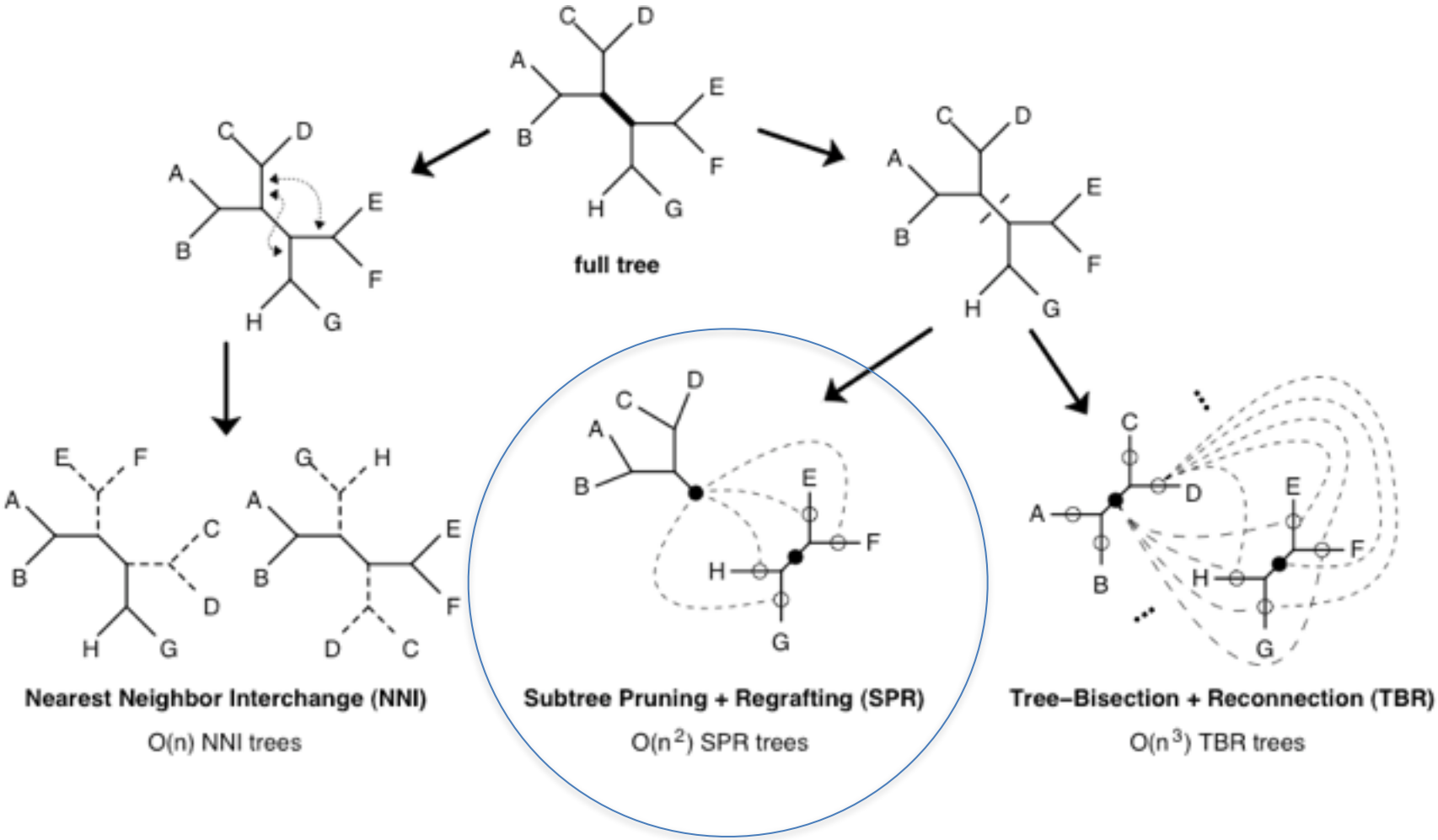
Finding the best tree

Evaluate random rearrangements of the starting tree and accept new tree if it improves P(D|M,T). Continue until convergence.



Again we have an iterative stochastic process as we have seen in the alignment case

Tree rearrangements in RAxML*

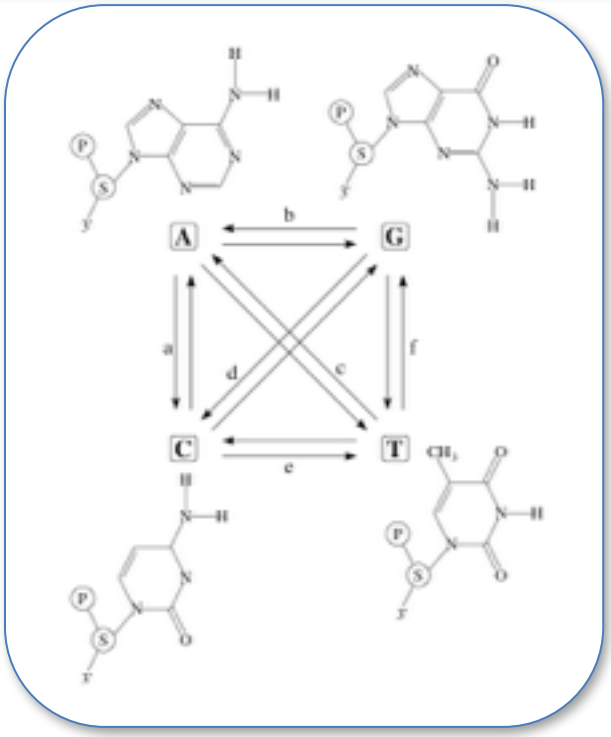


Stamatakis et al. 2008 Syst Biol 57:758-771

Modeling rate across sites
(Substitution rate heterogeneity across sites)

```
SRYC_DROME/358-380          YQCD...ICG...QKFVQKINLTHHARI...H
INSM1_HUMAN/441-464         HLCP...VCG...ESFASKGAQERHLRL..LH
XFIN_XENLA/1276-1298        YGCN...CCD...RSFSTHSASVRHQRM...C
XFIN_XENLA/1044-1066        YKCG...LCE...RSFVEKSALSRHQRV...H
ZNF76_HUMAN/285-309         YTCPE.PHCG...RGFTSATNYKNHVRI...H
CF2_DROME/401-423           YTCS...YCG...KSFTQSNTLKQHTRI...H
IKZF1_MOUSE/144-166         FQCN...QCG...ASFTQKGNLLRHIKL...H
EVI1_HUMAN/131-154          YECE...NCA...KVFTDPSNLQRHIRS..QH
TRA1_CAEEL/337-362          YSCQI.PQCT...KSYTDPSSLRKHIKA..VH
SUHW_DROAN/349-373          YACK...ICG...KDFTRSYHLKRHQKYS.SC
EGR1_HUMAN/396-418          FACD...ICG...RKFARSDERKRHTKI...H
ADR1_YEAST/104-126          FVCE...VCT...RAFARQEHLKRHYRS...H
SDC1_CAEEL/268-290          YFCH...ICG...TVFIEQDNLFKHWRL...H
SDC1_CAEEL/145-168          YMCQ...VCL...TLFGHTYNLFMHWRT..SC
KRUH_DROME/299-321          FECE...FCH...KLFSVKENLQVHRRI...H
TTKB_DROME/538-561          YPCP...FCF...KEFTRKDNMTAHVKI..IH
KRUP_DROME/222-244          FTCK...ICS...RSFGYKHVLQNHERT...H
BNC1_HUMAN/928-951          ITCH...LCQ...KTYSNKGTFRAHYKT..VH
ESCA_DROME/370-392          CKCN...LCG...KAFSRPWLLQGHIRT...H
ADR1_YEAST/132-155          YPCG...LCN...RCFTRRDLLIRHAQK..IH
CF2_DROME/429-451           FRCG...YCG...RAFTVKDYLNKHLTT...H
ZG28_XENLA/174-196          FTCT...ECG...KCLTRQYQLTEHSYL...H
ZG3_XENLA/6-28              FMCT...KCG...KCLSTKQKLNLHHMT...H
YL57_CAEEL/26-49            YLCY...YCG...KTLSDRLEYQQHMLK..VH
ZG5A_XENLA/90-112           FSCT...VCG...EMFTYRAQFSKHMLK...H
ZG52_XENLA/6-27             FTCP...ECG...KRF.SQKSNCWHTED...H
P43_XENBO/45-69             WKCGK.KDCG...KMFARKRQIQKHMKR...H
ZO2_XENLA/34-59             YSCA...DCG...KHFSEKMYLQFHQKNPSEC
ZG8_XENLA/146-168           FTCT...ECG...EHFANKVSLLGHLKM...H
SDC1_CAEEL/652-674          VVCF...HCG...TRC.HYTLLHDHLDY..CH
ZO61_XENLA/62-84            FTCF...ECG...TCFVNYSWLMLHIRM...H
ZG44_XENLA/5-27             FACT...KCK...RRFCSNKELFSHKRI...H
```

Modeling rate across sites
Revisiting substitution models



$$Q = \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{C} & \mathbf{G} & \mathbf{T} \end{array} \\ \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} \end{array}$$

It is a convention to set the diagonal entries q_{ii} such that the rows sum up to 0. Thus,

$$q_{ii} = -\sum_{j \neq i} q_{ij}$$

However, this model assumes that all sites in a sequence, or all columns in an alignment evolve with the same relative rate. Note, that we can rewrite the total rate for a given position as

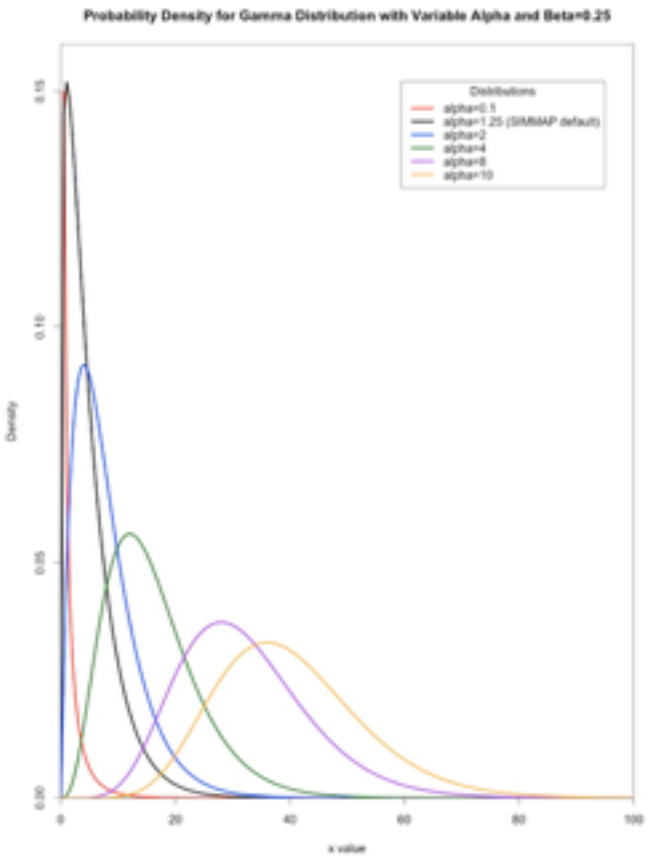
$$q_i = \sum_{j \neq i} q_{ij}$$

We can now introduce a neutral parameter $r=1$ such that can re-write q_i as $q_i * r$

For a sequence of L characters we have now the possibility to give the parameter r for $l=1 \ldots L$ a site specific value r_l.

We re-scale the substitution rate in a site-specific manner, i.e. the substitution rate at a position l is $q_i r_l$

Modeling rate across sites
Common approaches



Continuous Gamma distribution with a mean of 1*. Note that the parameter α determines the shape of the distribution.
(Problem of over-parameterization and over-fitting)

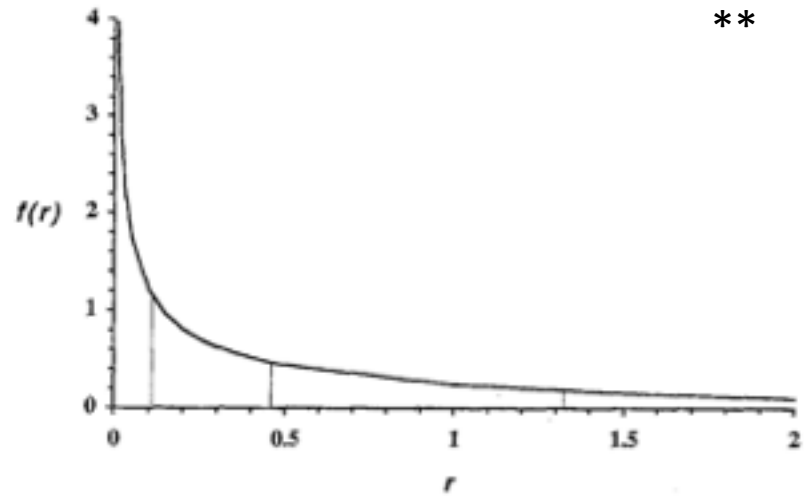
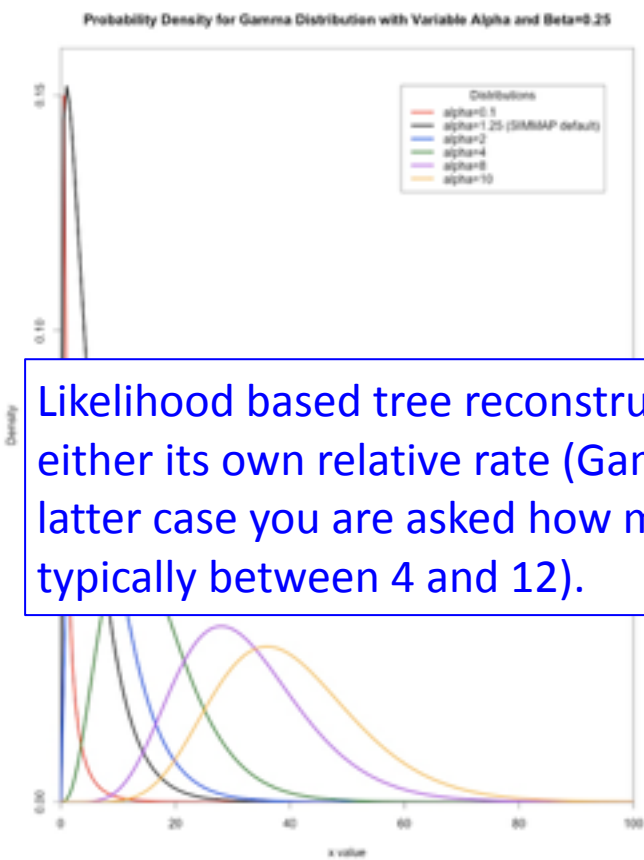


Fig. 1. Discrete approximation to the gamma distribution $G(\alpha,\beta)$, with $\alpha = \beta = 1/2$. Four categories are used to approximate the continuous distribution, with equal probability for each category. The three boundaries are 0.1015, 0.4549, and 1.3233, which are the percentage points corresponding to $p = 1/4, 2/4, 3/4$. The means of the four categories are 0.0334, 0.2519, 0.8203, 2.8944. The medians are 0.0247, 0.2389, 0.7870, 2.3535, and these are scaled to get 0.0291, 0.2807, 0.9248, and 2.7654, so that the mean of the discrete distribution is one.

*Thus, if I choose randomly from this distribution, the overall rate will not change!; **Z. Yang (1994) J Mol Evol 39:306

Modeling rate across sites
Common approaches



Continuous Gamma distribution with a mean of 1*. Note that the parameter α determines the shape of the distribution.
(Problem of over-parameterization and over-fitting)

Likelihood based tree reconstruction methods assign each position in the alignment either its own relative rate (Gamma model) or assigns it to a given rate category. In the latter case you are asked how many rate categories you want to use (values range typically between 4 and 12).

Fig. 1. Discrete approximation to the gamma distribution $G(\alpha,\beta)$, with $\alpha = \beta = 1/2$. Four categories are used to approximate the continuous distribution, with equal probability for each category. The three boundaries are 0.1015, 0.4549, and 1.3233, which are the percentage points corresponding to $p = 1/4, 2/4, 3/4$. The means of the four categories are 0.0334, 0.2519, 0.8203, 2.8944. The medians are 0.0247, 0.2389, 0.7870, 2.3535, and these are scaled to get 0.0291, 0.2807, 0.9248, and 2.7654, so that the mean of the discrete distribution is one.

*Thus, if I choose randomly from this distribution, the overall rate will not change!; **Z. Yang (1994) J Mol Evol 39:306

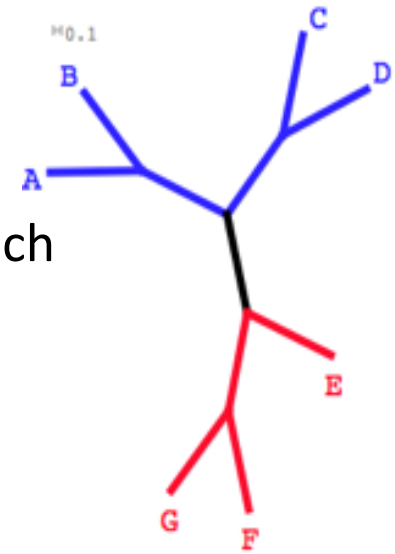
Looking at trees via their **splits**

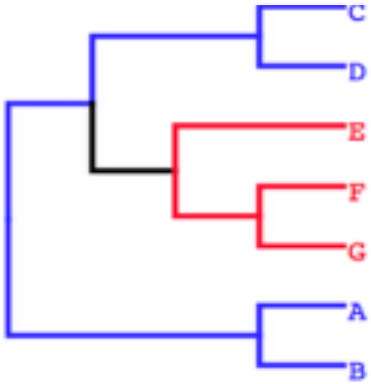
Each branch of a tree describes a **split** of OTUs into two sets

These sets correspond to the two clades associated with the branch

e.g. black branch of the tree specifies the split ABCD | EFG

- can also be written ADCB | GFE etc.
- i.e. the taxon lists in the two halves of the split are unordered





Looking at trees via their **splits**

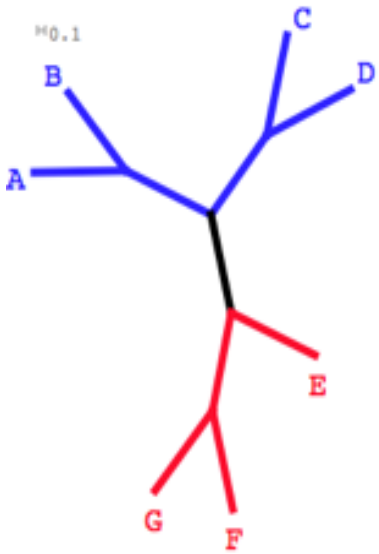
Splits are either

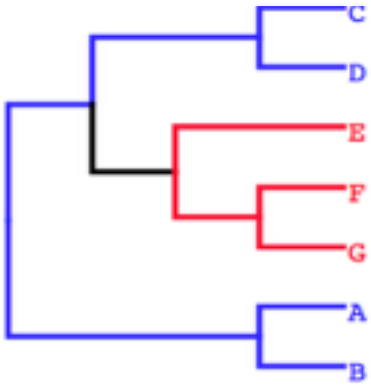
trivial

- example: F | ABCDEG
- associated with **terminal** branches
- provide **no** information about topology structure

non-trivial

- example: ABCD | EFG
- associated with **internal** branches
- provide information about the tree topology

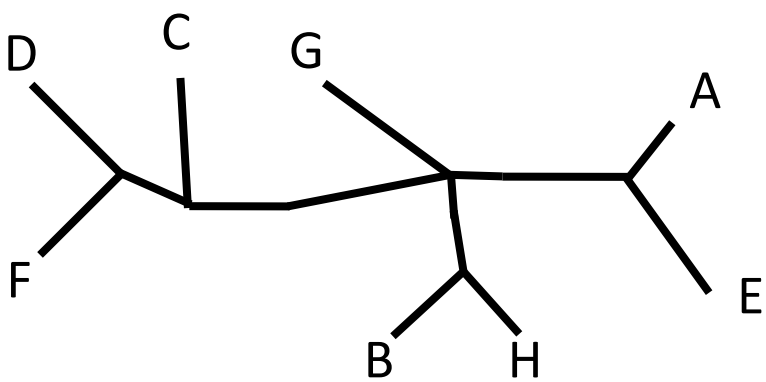




Looking at trees via their **splits**

Complete list of splits described by a tree allows reconstruction of that tree's topology

Helps to consider the sets of clades described by the splits



DF	ABCEGH
BCDFGH	AE
ABEGH	CDF
BH	ACDEFG

Split Compatibility

Sets (e.g. pairs) of splits are either:

compatible

- a tree **can** be drawn that contains all splits in the set

incompatible

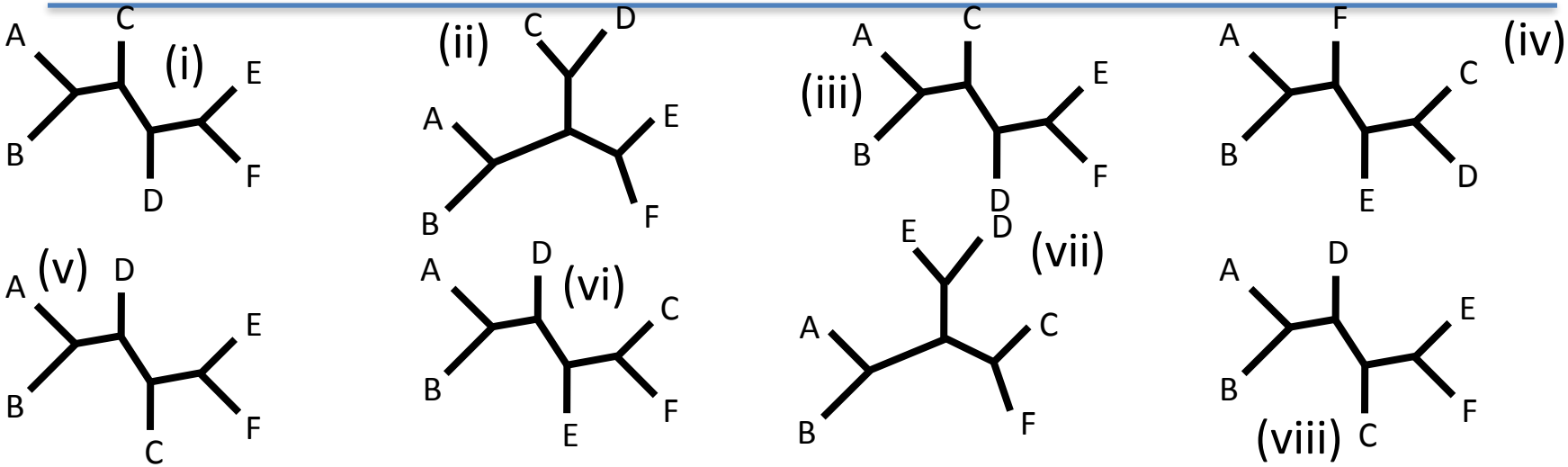
- a tree **cannot** be drawn that contains all splits in the set

Definition: Two splits W|X and Y|Z are compatible, i.e. not contradictory, if at least one intersection of $W \cap Y$,
$W \cap Z$, $X \cap Y$, $X \cap Z$ is empty.

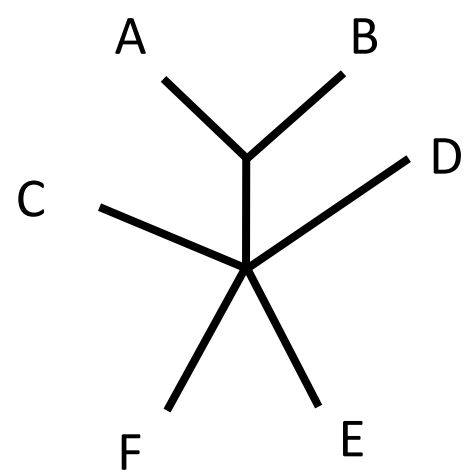
Which of these sets of splits is incompatible?

(i)	(ii)	(iii)
AB | CDE	BCDFGH | AE	AB | CDE
DE | ABC	ABEGH | CDF	AC | BDE
	BG | ACDEFH	

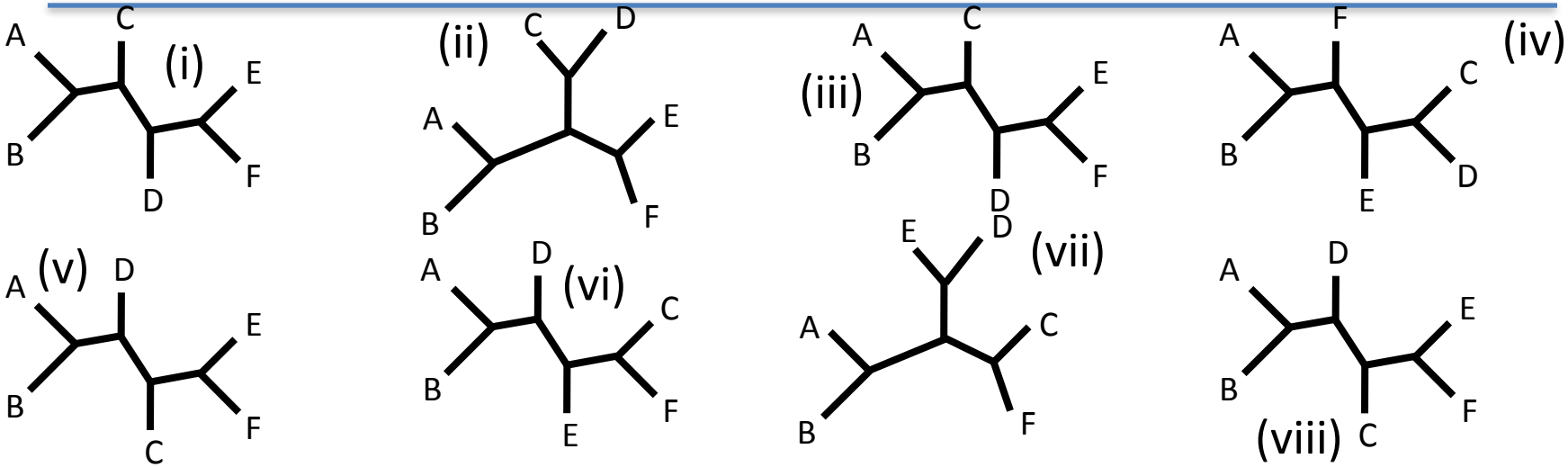
Sets of trees can be summarized by looking at their split sets: Strict Consensus Trees



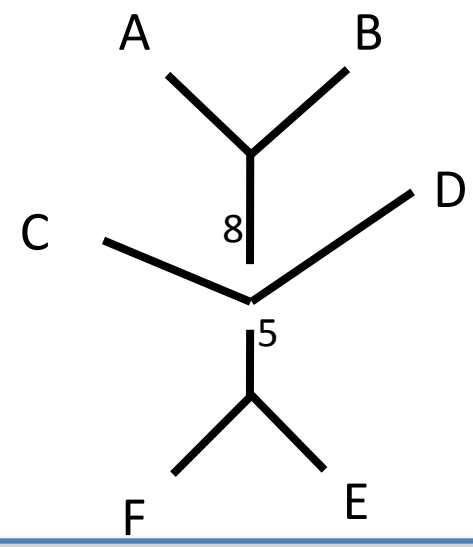
	i	ii	iii	iv	v	vi	vii	viii	
AB | CDEF	*	*	*	*	*	*	*	*	8
CD | ABEF		*		*					2
EF | ABCD	*	*	*		*			*	5
ABC | DEF	*		*						2
DE | ABCF							*		1
CF | ABED						*	*		2
ABD | ECF					*	*		*	3
ABE | CDF				*					1



Sets of trees can be summarized by looking at their split sets: 50% Majority Rule Consensus Trees

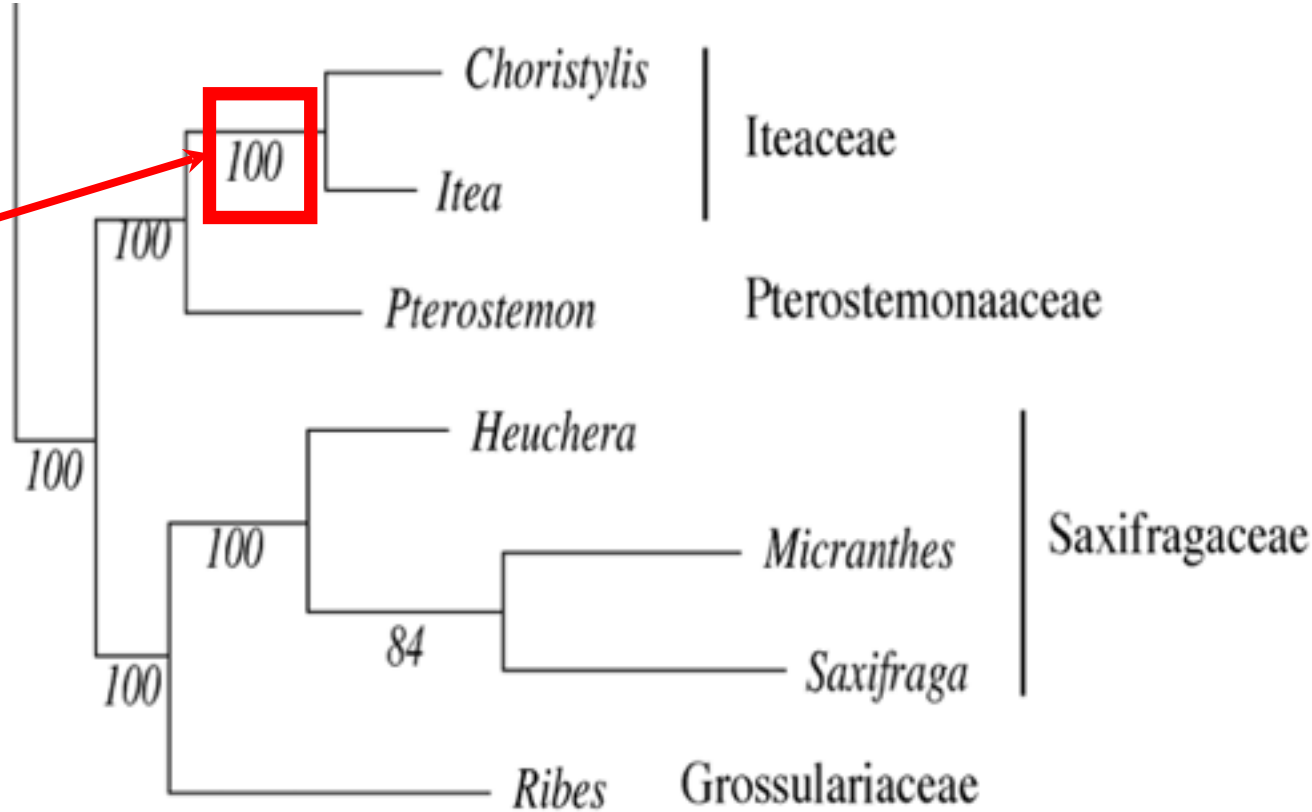


	i	ii	iii	iv	v	vi	vii	viii	
AB | CDEF	*	*	*	*	*	*	*	*	8
CD | ABEF		*		*					2
EF | ABCD	*	*	*		*			*	5
ABC | DEF	*		*						2
DE | ABCF							*		1
CF | ABED						*	*		2
ABD | ECF					*	*		*	3
ABE | CDF				*					1



Label the Branches!

Branches of consensus tree labeled to indicate proportion of trees containing that branch/split



Resolving an ancient, rapid radiation in Saxifragales.
Jian S, Soltis PS, Gitzendanner MA, Moore MJ, Li R, Hendry TA, Qiu YL, Dhingra A, Bell CD, Soltis DE.
Syst Biol. 2008 Feb;57(1):38-57.
PMID: 18275001