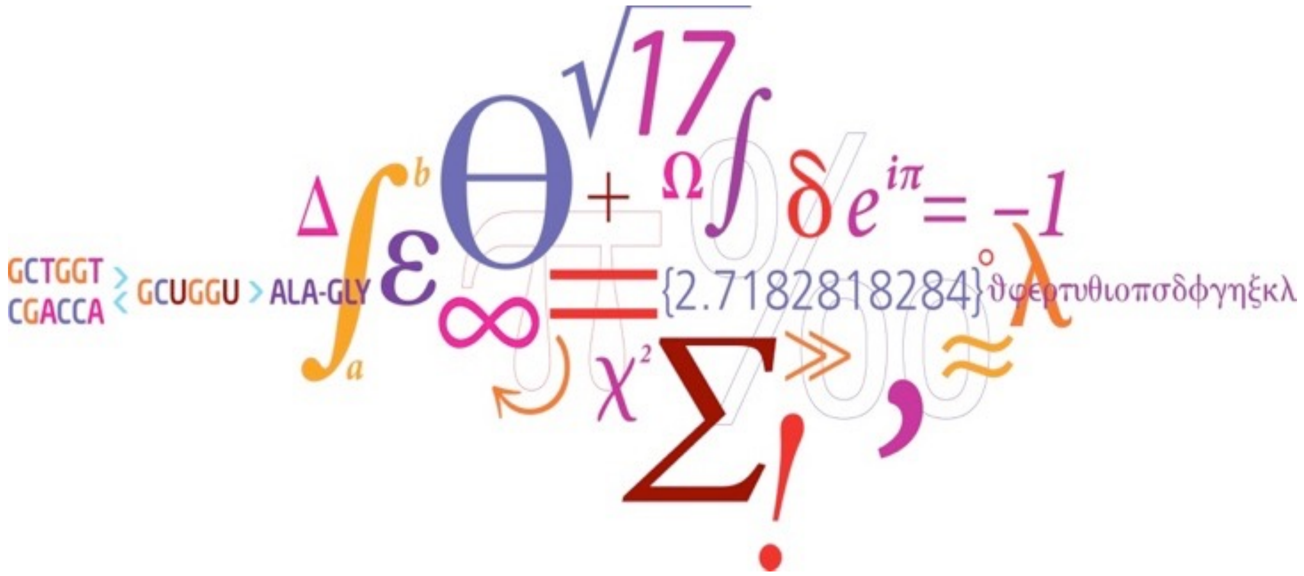
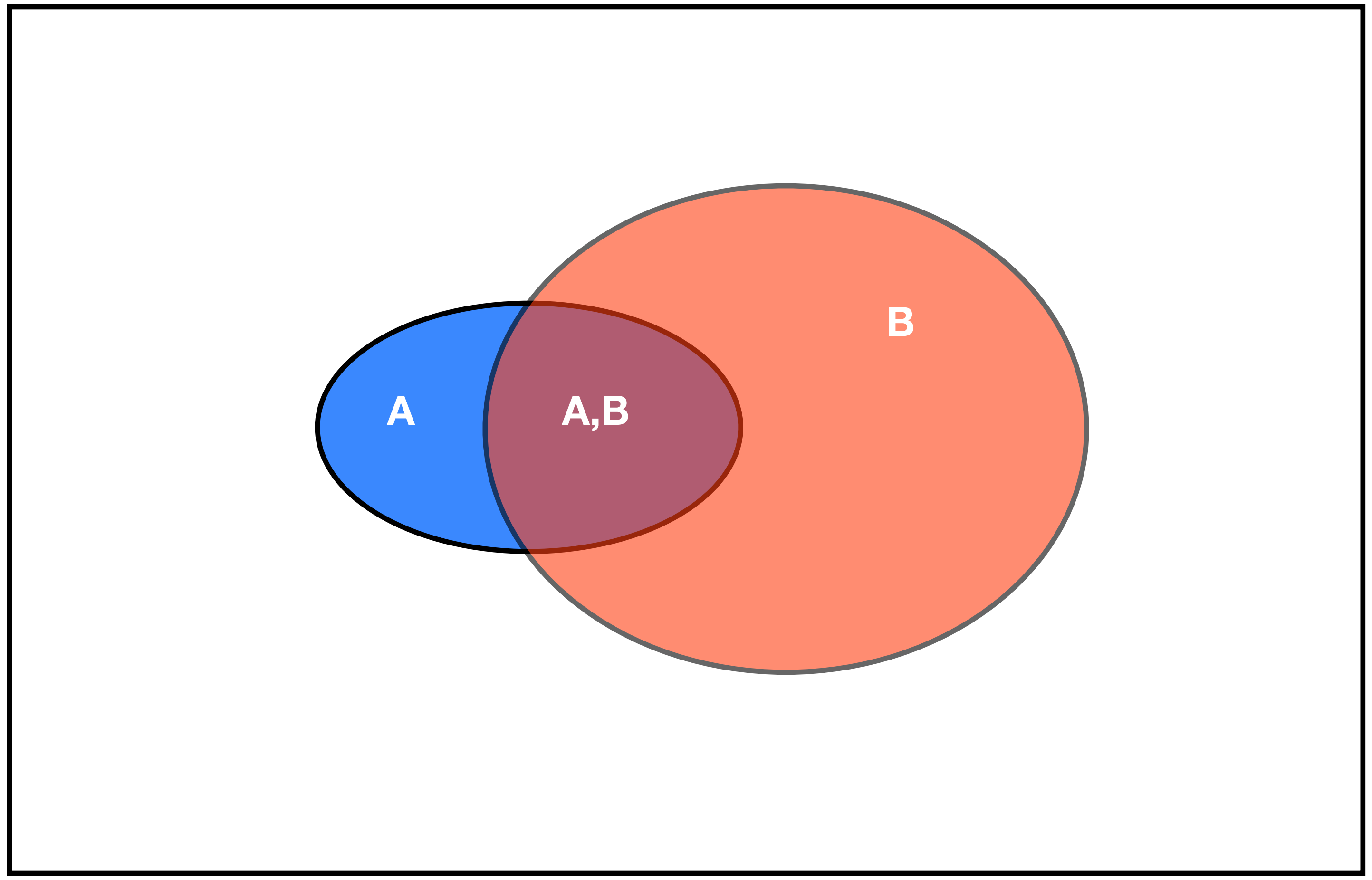
Bayesian Inference



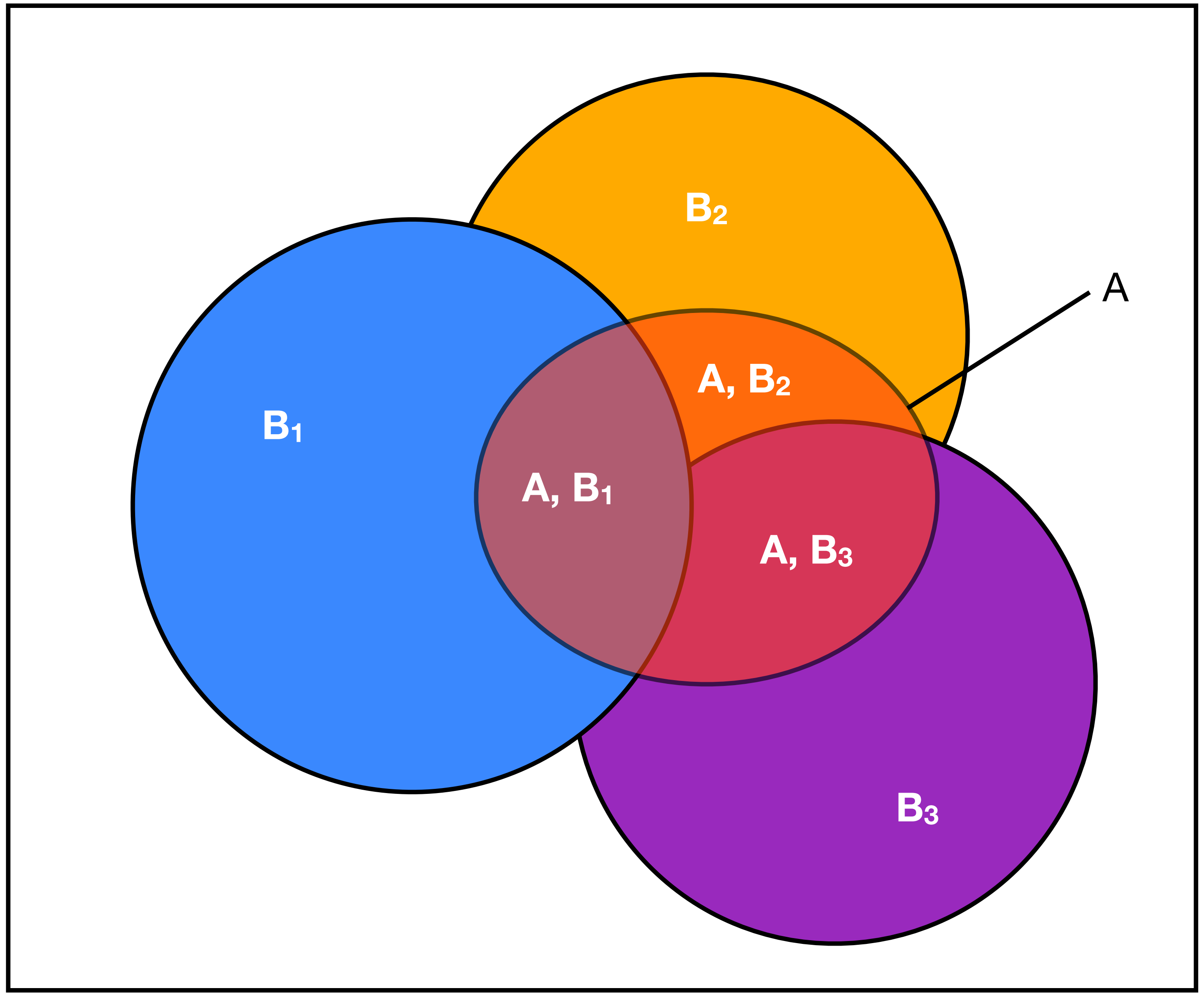
Conditional probability



$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$P(A,B) = P(A|B)P(B)$$

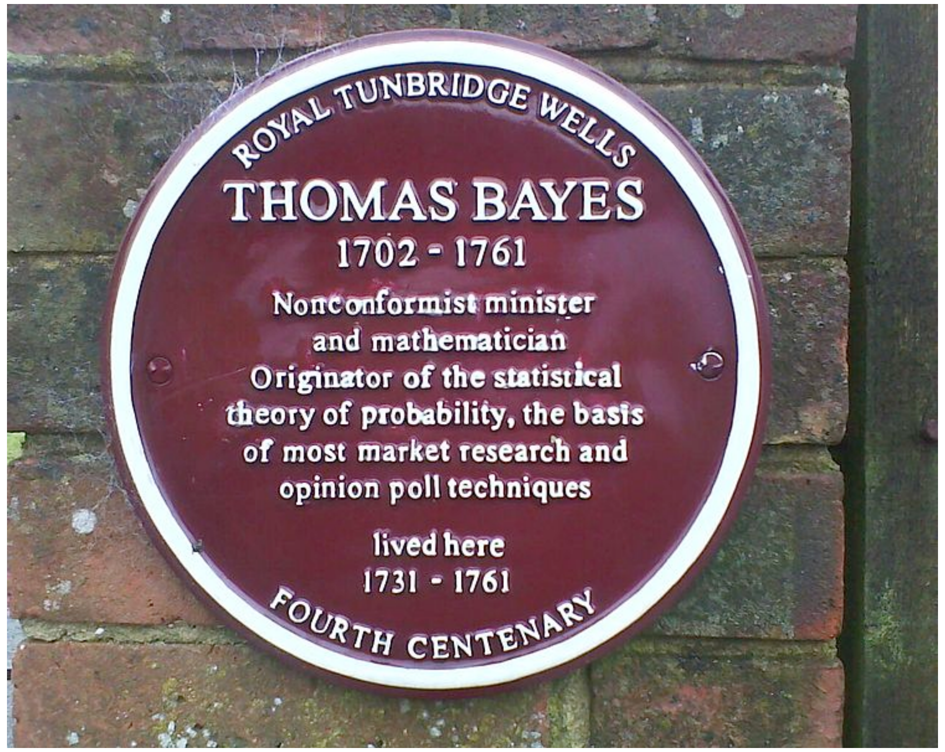
The law of total probability



$$
\begin{aligned}
P(A) &= P(A, B_1) + P(A, B_2) + P(A, B_3) \\
&= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)
\end{aligned}
$$

Bayes' Theorem

$$\boxed{P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}}$$



Reverend Thomas Bayes
(1702-1761)
Image Source: Wikimedia

Bayes' Theorem

$$P(A|B) = \frac{P(A,B)}{P(B)} \Rightarrow P(A,B) = P(A|B)P(B)$$

$$P(B|A) = \frac{P(A,B)}{P(A)} \Rightarrow P(A,B) = P(B|A)P(A)$$

$$\Downarrow$$

$$P(B|A)P(A) = P(A|B)P(B)$$

$$\Downarrow$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \longleftarrow \text{Bayes' theorem}$$

Bayesians vs. Frequentists: Meaning of Probability

- **Frequentist**: long-run frequency of event in repeatable experiment

- **Bayesian**: degree of belief, way of quantifying uncertainty

Probabilities as extended logic

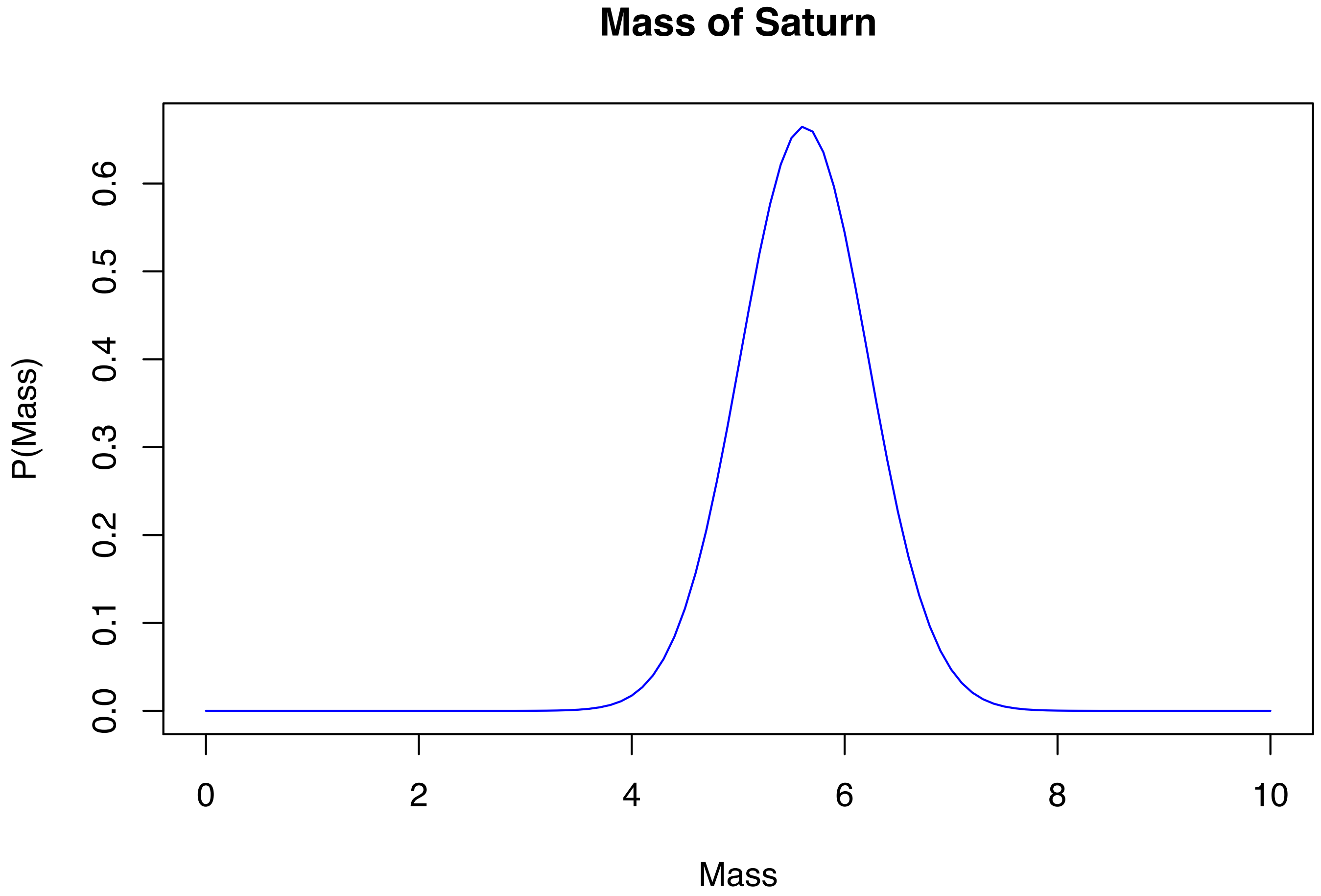
- Polya, Cox, Jeffreys, Jaynes: probabilities are the only consistent basis for plausible reasoning (reasoning when there is insufficient information for deductive reasoning).

- Probabilities should form basis of all scientific inference

- Evidence from different sources integrated by using simple laws of probability (multiplication and summation...)

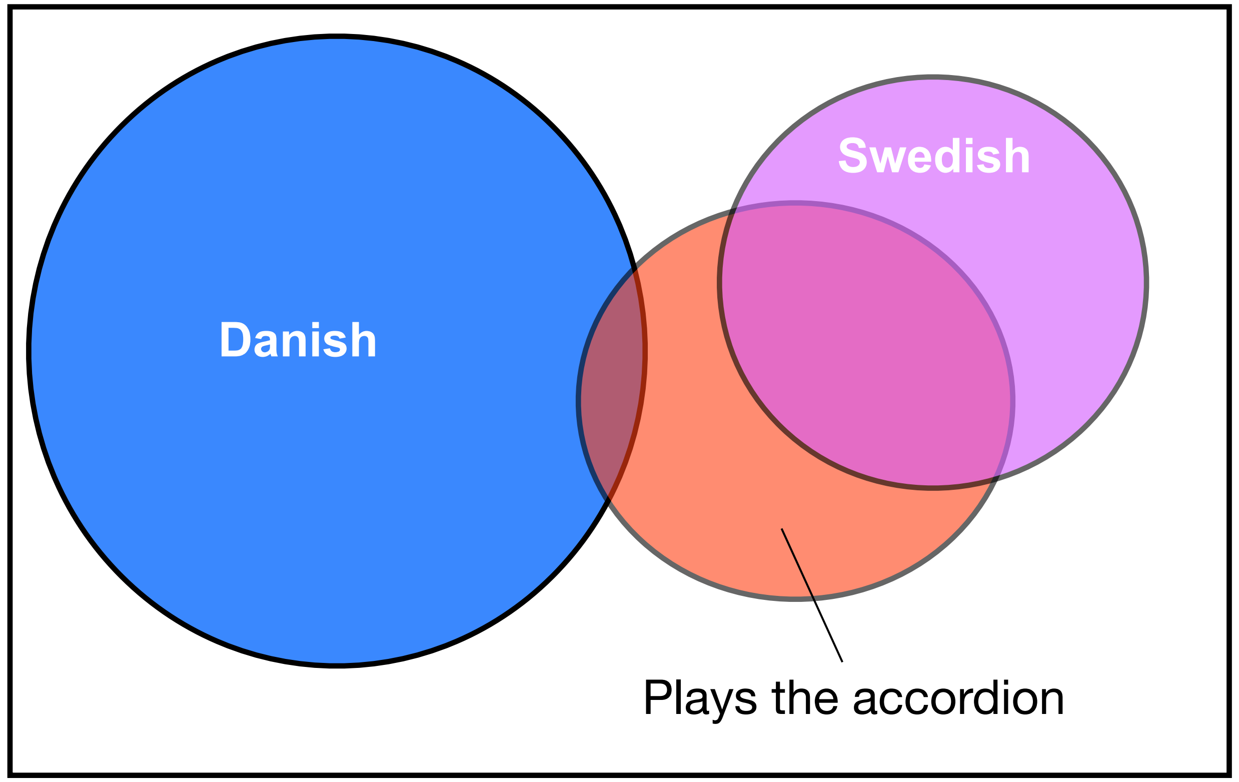
Bayes' Theorem: Probability distributions over possible parameter values as a way of expressing uncertainty

Pierre-Simon, marquis de Laplace
(1745-1827)
Image Source: Wikimedia



- Extracting information about reality from empirical data:
  - **Frequentist**: parameters in model are fixed constants whose true values we are trying to find good (point) estimates for.
  - **Bayesian**: uncertainty concerning model parameters expressed by means of probability distribution over possible parameter values

Bayes’ Theorem: Updating degree of belief after seeing data



Among people working at DTU:

P(Swedish) = 0.16
P(plays accordion) = 0.1269
P(plays accordion | Swedish) = 0.5

$$
\begin{aligned}
P(\text{Swedish}|\text{plays the accordion}) &= \frac{P(\text{plays the accordion}|\text{Swedish})P(\text{Swedish})}{P(\text{plays the accordion})} \\
&= \frac{0.5 \times 0.16}{0.1269} = 0.6304
\end{aligned}
$$

Knowledge about reality updated by data via Bayes theorem:

Before data: P(Swedish) = 0.16
After data: P(Swedish | Data) = 0.63

Bayes' Theorem

$$\boxed{P(H|D) = \frac{P(D|H)P(H)}{P(D)}}$$

- P(H): Prior probability of hypothesis
- P(D|H): Probability of data given hypothesis = likelihood
- P(H|D): Posterior probability of hypothesis
- P(D): "Marginal probability" of observing data. Essentially a normalizing constant so posterior will sum to one (but useful for model comparison)

Bayes' Theorem

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Update knowledge about hypothesis H based on data D

$$P(w_i|D) = \frac{P(D|w_i)P(w_i)}{P(D)}$$

Focus is on learning about w, the parameters within some model

$$P(w_i|D, H_1) = \frac{P(D|w_i, H_1)P(w_i|H_1)}{P(D|H_1)}$$

H_1: We sometimes want to make it explicit that parameters w belong to (are conditioned upon) this particular model/hypothesis

$$P(w_i|D) = \frac{P(D|w_i)P(w_i)}{\sum_{j=1}^{N} P(D|w_j)P(w_j)}$$

P(D) can be found by summing over $P(D|w_j)P(w_j)$ for all the mutually exclusive, possible parameter values w_j (law of total probability)

Bayesian model comparison

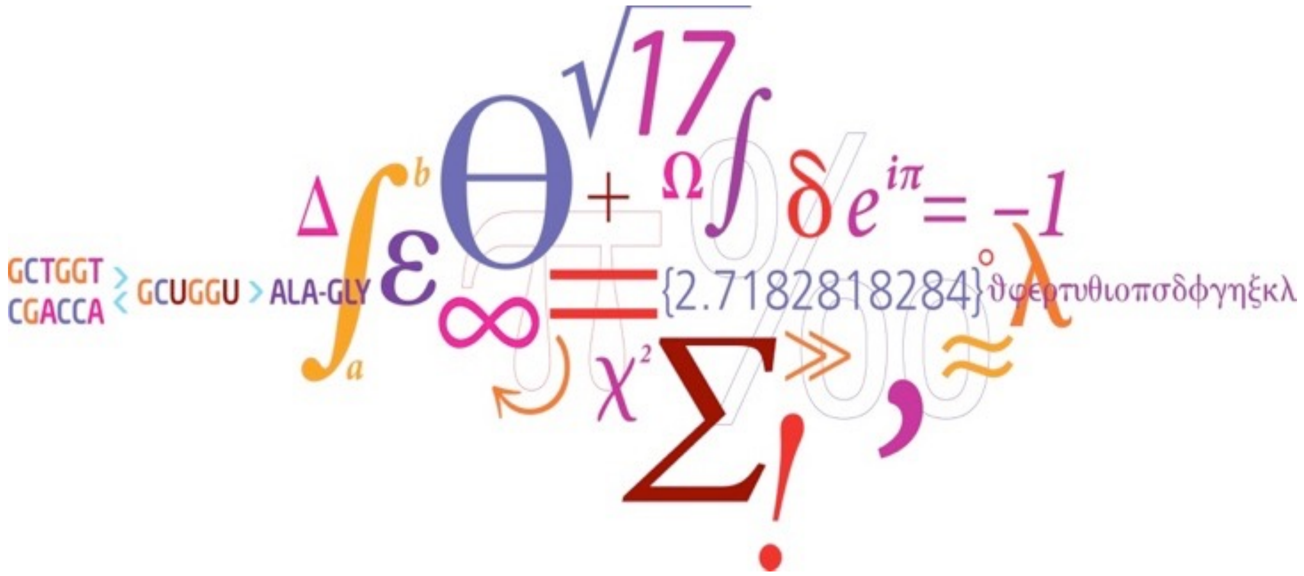
$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}$$

- It is possible to compute posterior probabilities for alternative hypotheses just like we can compute the posterior probabilities of different possible parameter values within a single hypothesis

- In this expression, P(DIH) is the denominator from the previous expression:

$$P(D|H_1) = \sum_{j=1}^{N} P(D|w_j)P(w_j)$$

Markov chain Monte Carlo



MCMC: Markov chain Monte Carlo

$$P(w_i|D) = \frac{P(D|w_i)P(w_i)}{\sum_{j=1}^{N} P(D|w_j)P(w_j)}$$

$$P(\tau_i, t_i, \alpha|data) = \frac{P(data|\tau_i, t_i, \alpha)P(\tau_i, t_i, \alpha)}{\sum_{j=1}^{C_s} \int_{t_j} \int_{\alpha'} P(data|\tau_i, t_j, \alpha')P(\tau_j, t_j, \alpha')dt_j d\alpha'}$$

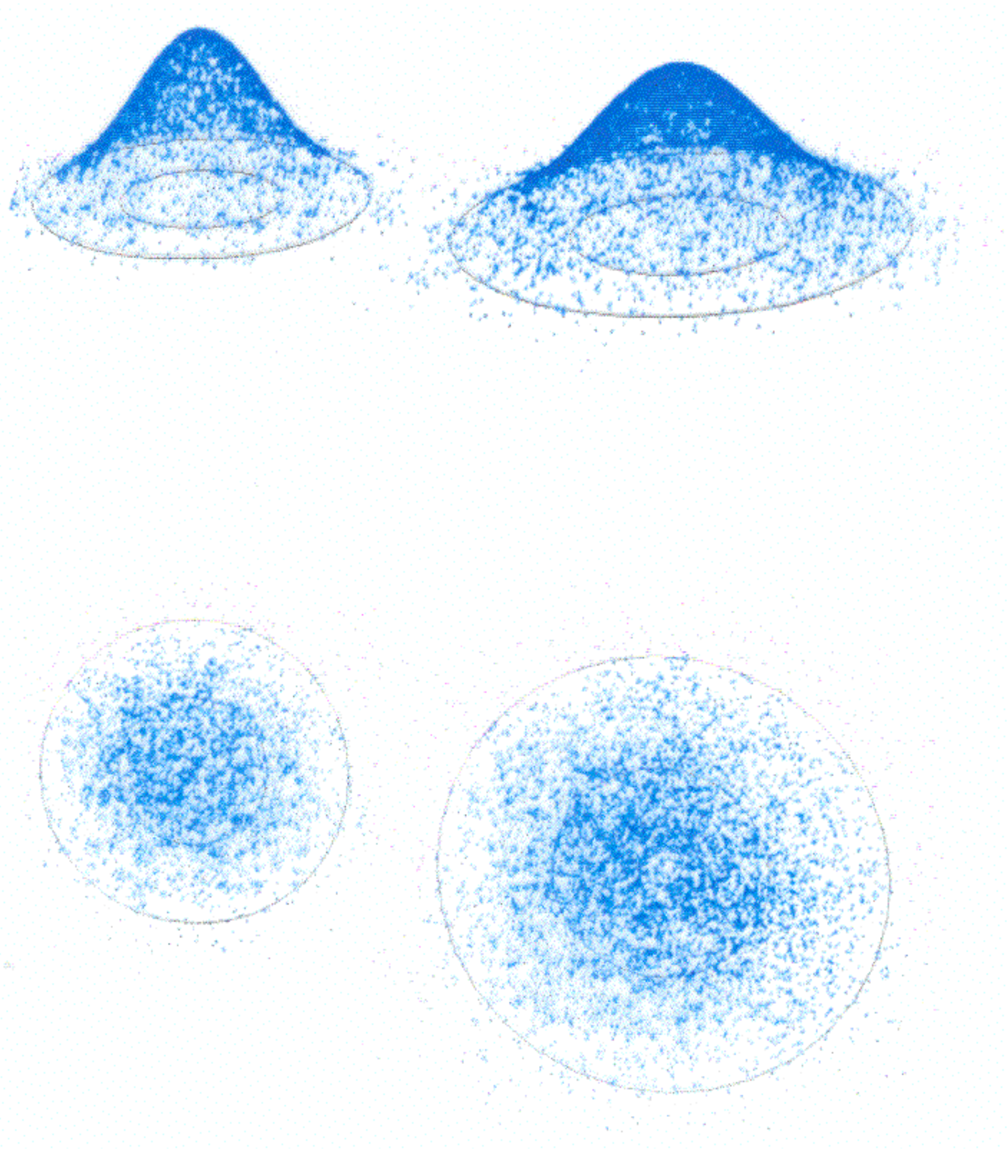
- Can be difficult or impossible to compute (either analytically or numerically)
- Solution: Markov chain Monte Carlo (MCMC)

MCMC: Markov chain Monte Carlo

Starting point:

- Parameter space (covering all possible parameter values for all parameters in model)
- For each possible parameter value we can compute the likelihood = P(D | parameter values)
- For each parameter value we know the prior probability = P(parameter values)
- We can therefore compute prior x likelihood for any given point in parameter space

MCMC: Markov chain Monte Carlo



- Start in random position on probability landscape (X). Compute prior x likelihood here. Let's call that P_X.
- Based on current position: attempt move to new position (Y) by randomly drawing from "proposal distribution": q(Y| X)
- (For example, the proposal distribution can be a normal distribution with mean X and standard deviation 1)
- Compute prior x likelihood at new position. We'll call that P_Y
  - (a) If move ends higher up, i.e. $P_Y > P_X$: accept move
  - (b) If move ends below: accept move with probability

$$\boxed{P(\text{accept}) = \frac{P_Y \times q(X|Y)}{P_X \times q(Y|X)}}$$

- If q(X|Y) = q(Y|X), i.e., q is symmetric, this becomes:

$$\boxed{P(\text{accept}) = \frac{P_Y}{P_X}}$$

- Write parameter values for accepted moves in file (if proposed move is not accepted: write previous values again).
- After many, many repetitions points will be sampled in proportion to the height of the probability landscape: We therefore have an empirical approximation of the distribution

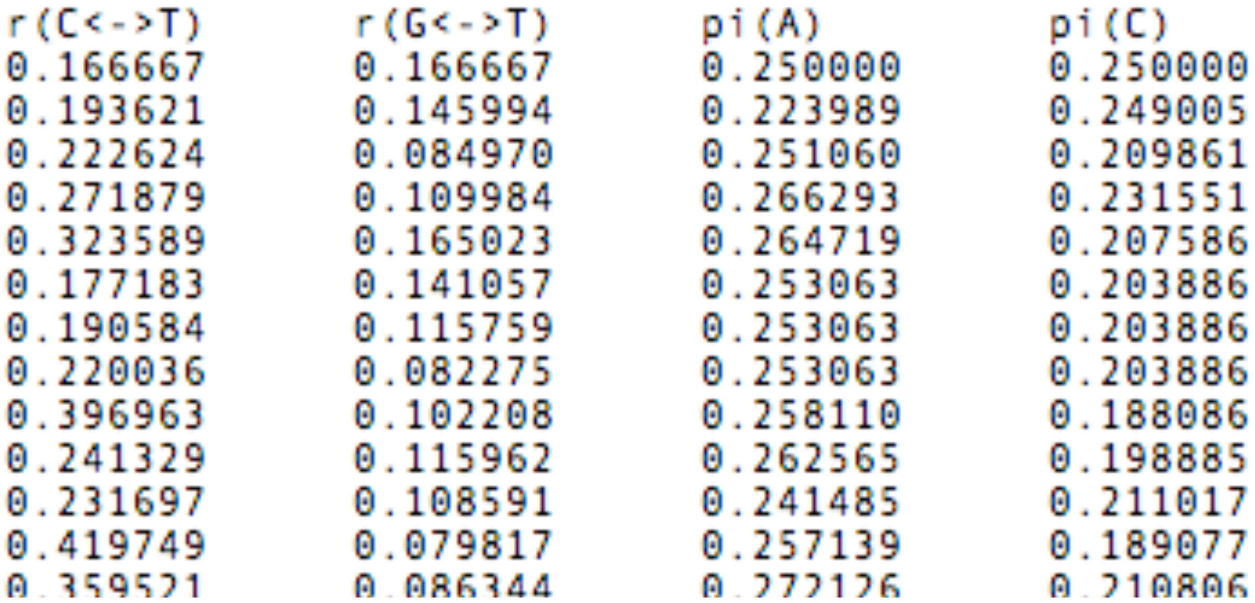
MCMCMC: Metropolis-coupled Markov Chain Monte Carlo

- **Problem**:
  - If there are multiple peaks in the probability landscape, then MCMC may get stuck on one of them
- **Solution**:
  - Metropolis-coupled Markov Chain Monte Carlo = MCMCMC = MC^3
- **MC^3 essential features:**
  - Run several Markov chains simultaneously
  - One chain “cold”: this chain performs MCMC sampling
  - Rest of chains are “heated”: move faster across valleys
  - Each turn the cold and warm chains may swap position (swap probability is proportional to ratio between heights)

➡ More peaks will be visited

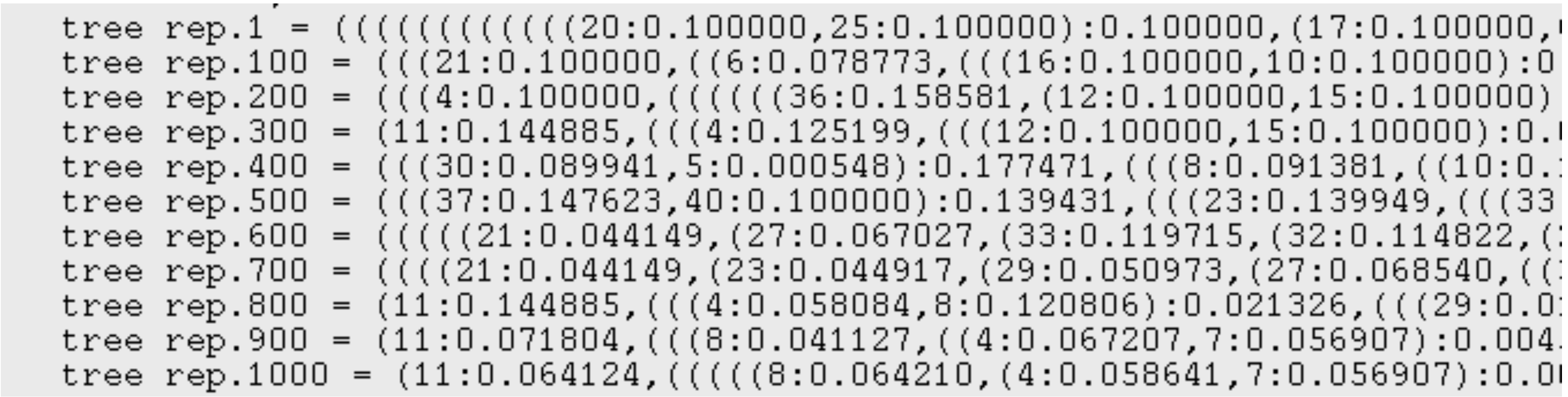
- More chains means better chance of visiting all important peaks, but each additional chain increases run-time (unless you use parallelization)

MCMCMC for inference of phylogeny



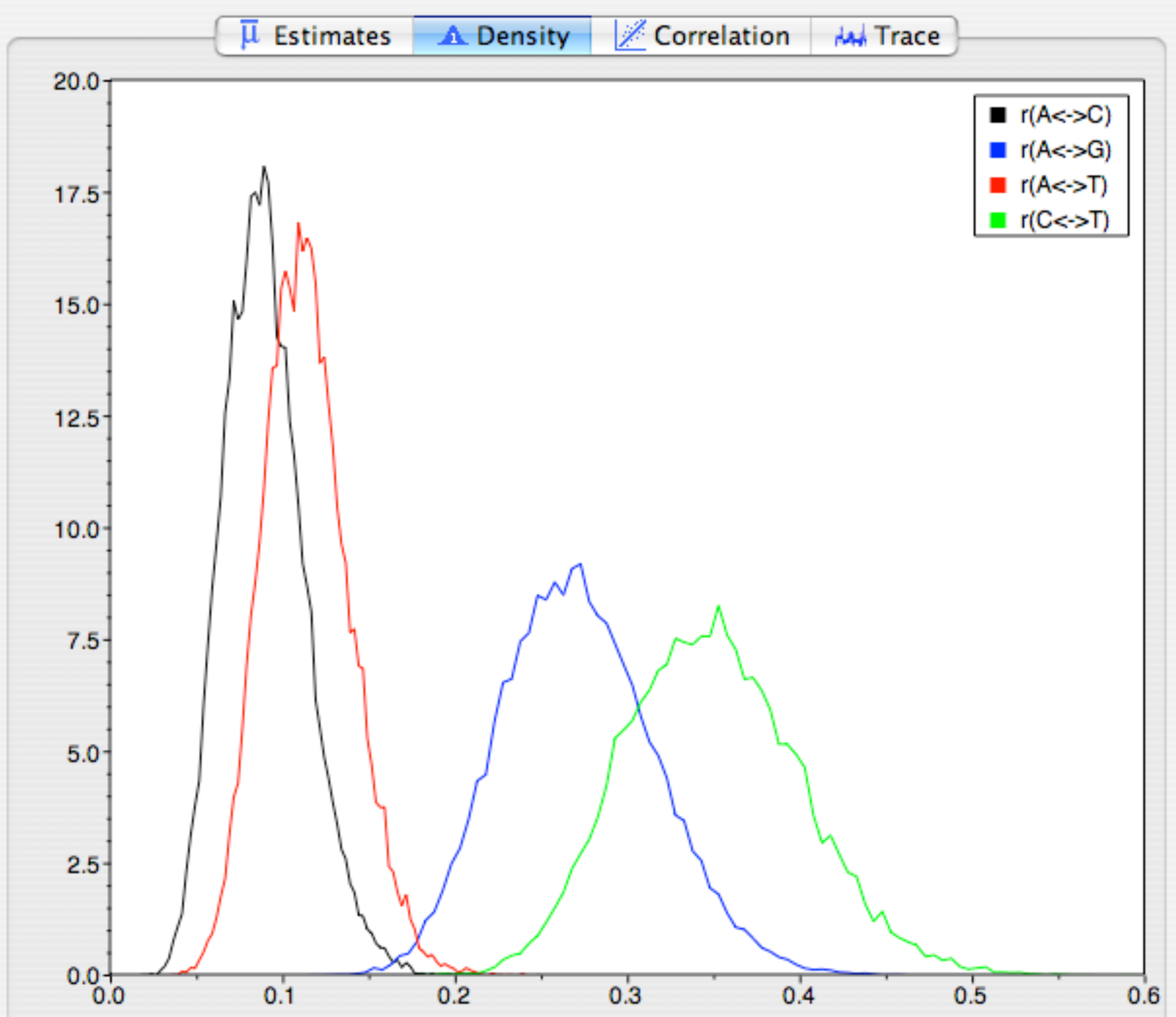
```
r(C<->T)      r(G<->T)      pi(A)         pi(C)
0.166667      0.166667      0.250000      0.250000
0.193621      0.145994      0.223989      0.249005
0.222624      0.084970      0.251060      0.209861
0.271879      0.109984      0.266293      0.231551
0.323589      0.165023      0.264719      0.207586
0.177183      0.141057      0.253063      0.203886
0.190584      0.115759      0.253063      0.203886
0.220036      0.082275      0.253063      0.203886
0.396963      0.102208      0.258110      0.188086
0.241329      0.115962      0.262565      0.198885
0.231697      0.108591      0.241485      0.211017
0.419749      0.079817      0.257139      0.189077
0.359521      0.086344      0.272126      0.210806
```

- Result of run:
  - Substitution parameters, nucleotide frequencies
  - Tree topologies

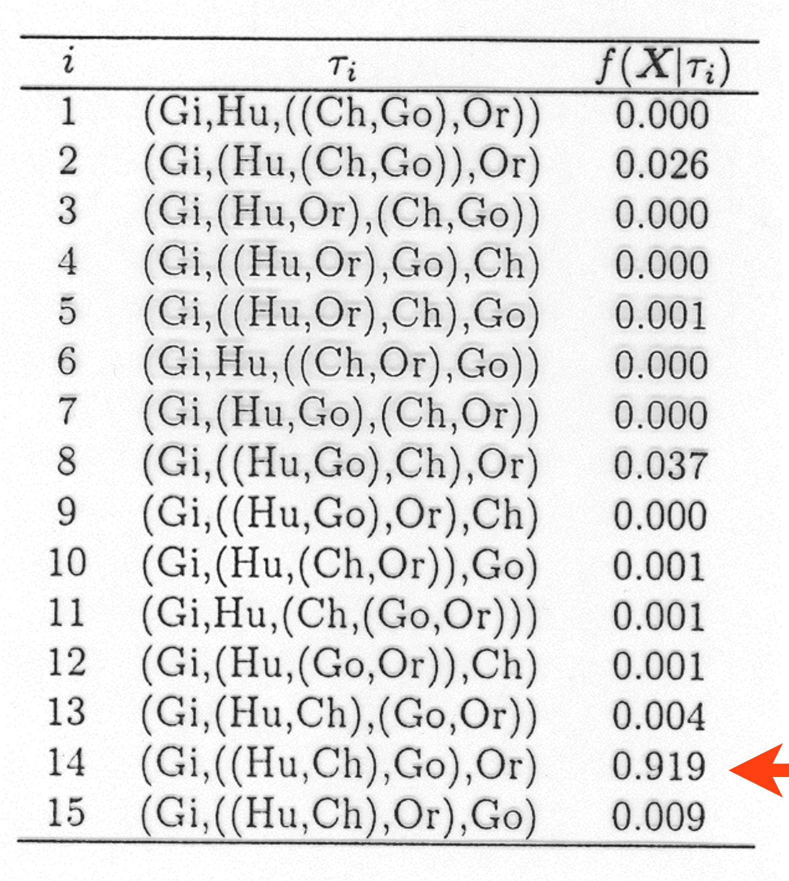


```
   tree rep.1 = (((((((((((20:0.100000,25:0.100000):0.100000,(17:0.100000,
   tree rep.100 = (((21:0.100000,((6:0.078773,(((16:0.100000,10:0.100000):0
   tree rep.200 = (((4:0.100000,((((((36:0.158581,(12:0.100000,15:0.100000)
   tree rep.300 = (11:0.144885,(((4:0.125199,(((12:0.100000,15:0.100000):0.
   tree rep.400 = (((30:0.089941,5:0.000548):0.177471,(((8:0.091381,((10:0.
   tree rep.500 = (((37:0.147623,40:0.100000):0.139431,(((23:0.139949,(((33
   tree rep.600 = (((((21:0.044149,(27:0.067027,(33:0.119715,(32:0.114822,(
   tree rep.700 = ((((21:0.044149,(23:0.044917,(29:0.050973,(27:0.068540,((
   tree rep.800 = (11:0.144885,(((4:0.058084,8:0.120806):0.021326,(((29:0.0
   tree rep.900 = (11:0.071804,(((8:0.041127,((4:0.067207,7:0.056907):0.004
   tree rep.1000 = (11:0.064124,(((((8:0.064210,(4:0.058641,7:0.056907):0.0
```

Posterior probability distributions of substitution parameters



Posterior Probability Distribution over Trees



i	τ_i	$f(X\|\tau_i)$
1	(Gi,Hu,((Ch,Go),Or))	0.000
2	(Gi,(Hu,(Ch,Go)),Or)	0.026
3	(Gi,(Hu,Or),(Ch,Go))	0.000
4	(Gi,((Hu,Or),Go),Ch)	0.000
5	(Gi,((Hu,Or),Ch),Go)	0.001
6	(Gi,Hu,((Ch,Or),Go))	0.000
7	(Gi,(Hu,Go),(Ch,Or))	0.000
8	(Gi,((Hu,Go),Ch),Or)	0.037
9	(Gi,((Hu,Go),Or),Ch)	0.000
10	(Gi,(Hu,(Ch,Or)),Go)	0.001
11	(Gi,Hu,(Ch,(Go,Or)))	0.001
12	(Gi,(Hu,(Go,Or)),Ch)	0.001
13	(Gi,(Hu,Ch),(Go,Or))	0.004
14	(Gi,((Hu,Ch),Go),Or)	0.919
15	(Gi,((Hu,Ch),Or),Go)	0.009

- MAP (maximum a posteriori) estimate of phylogeny: tree topology occurring most often in MCMCMC output
- Clade support: posterior probability of group = frequency of clade in sampled trees.
- 95% credible set of trees: order trees from highest to lowest posterior probability, then add trees with highest probability until the cumulative posterior probability is 0.95