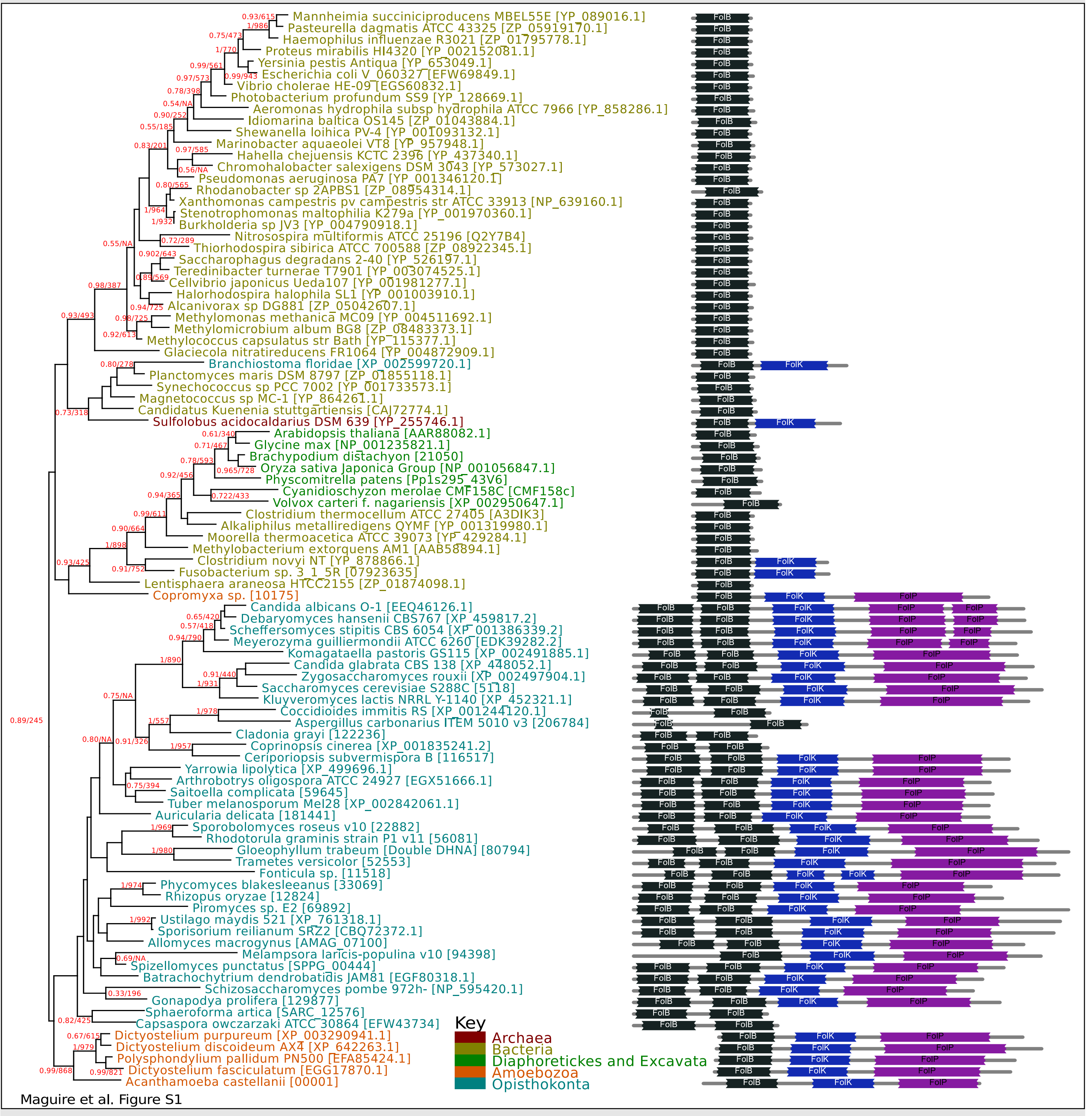
Basic Principles of Molecular Phylogenetics

Finlay Maguire
Finlay.Maguire.11@ucl.ac.uk
Natural History Museum, London; University College London; University of Exeter

Phylogenetics are a powerful tool in the study and elucidation of evolutionary processes

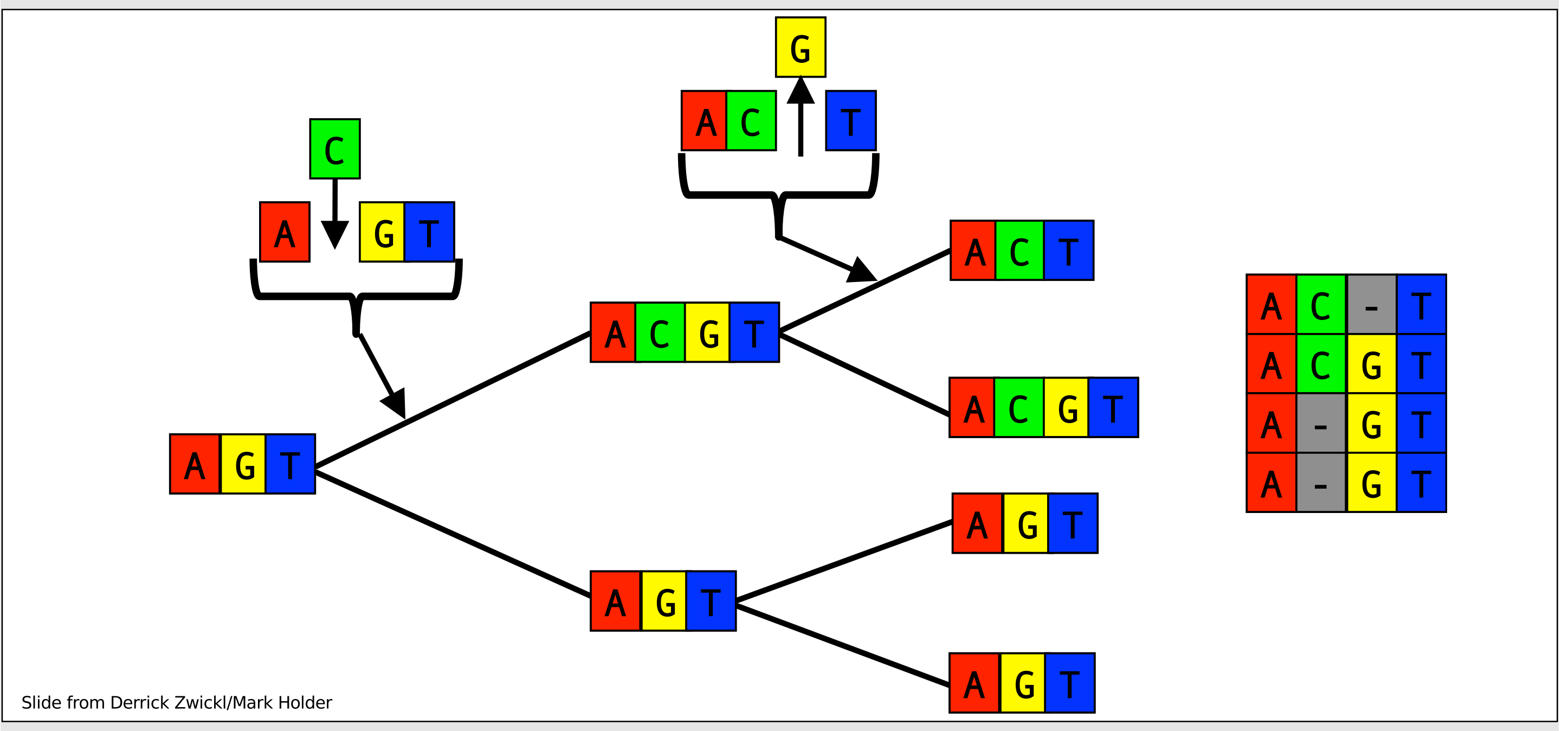
- Reconstruction of relationships between sequences and/or taxa
- Sequence identification
- Discovery of Horizontal Gene Transfer (HGT) events
- Exploration of equence and functional divergence
- Identification of evolutionary innovations
- Integral to many bioinformatic algorithms/applications



Maguire et al. Figure S1

Multiple Sequence Alignment

- Model of positional homology
- Data from which phylogeny is reconstructed
- Very important - all methods will mislead with a poor MSA
- Most tools implement progressive alignments using serial pairwise alignments followed by iterative improvement
- 'Guide tree' often generated to aid ancestral sequence reconstruction but can bias later reconstruction and inflate support (especially PRANK/ProtPal)
- Simultaneous inference of MSA and Phylogeny (e.g. BAliPhy) is potentially optimal solution but highly computationally expensive



Slide from Derrick Zwickl/Mark Holder

Masking

- Filtering of data from MSA to remove 'ambigous' sites:
  - Phylogenetically uninformative
  - Misleading
  - Poorly aligned i.e. alignment very unstable with different tools/settings
- Gapped sites often removed - many phylogenetic tools handle indel processes poorly (see work by Rivas *et al.*)
- Often conducted manually (Seaview) but low throughput and potential to bias reconstruction
- Automated tools (GBlocks, TrimAL) often highly heuristic

Distance and Parsimony methods

- Earliest methods
- Distance are based on clustering of sequences into tree using (weighted) distance scores
- MP is based on determining least evolutionary changes
- Fast computationally and can reduce compositional bias
- Prone to many other biases and inconsistency - especially LBA and variation in evolutionary rate within and among sites
- Typically used as a starting point for model-based methods (ML and BI) inplace of random topologies or as intermediate stages in other applications (e.g. MSA)

$$\delta = -2(lnL_1 - lnL_0), \delta \in \chi^2_{df=K}$$

$$BF = \frac{P(D|M_0)}{P(D|M_1)}$$

$$AIC_i = -2lnL_i + 2K$$

$$BIC_i = -2lnL_i + KlnN$$

Substitution models

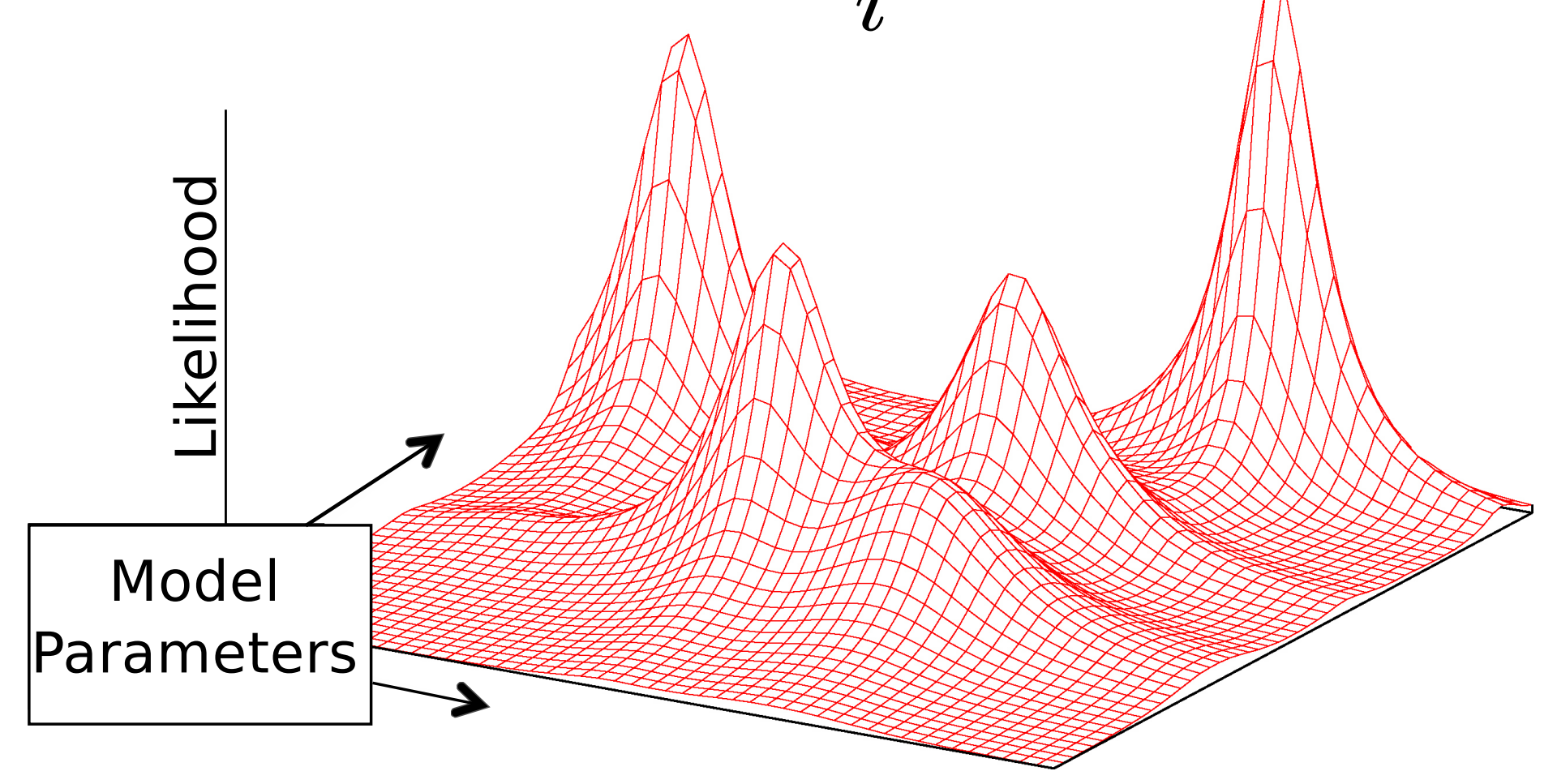
- Model based methods (ML and BI) require a statistical model of sequence evolution (i.e. P(G<-->T) etc.)
- Trade-off between variance and bias
- Inaccurate model selection can lead to inconsistency
- Multiple test criteria for model selection most of which implemented in tools to aid selection (jModelTest, ProtTest3)
- Nucleotide models are typically mechanistic (JC69/TN93) and nested within GTR (if all K are equal GTR = JC69)
- GTR parameters can fixed or estimated via DPP
- Codon models (GY94) also used but slow (3782 rates)
- Protein models (JTT/LG) typically empirical (observed rates in existing datasets) due to many state changes (380)
- ASRV handled by assigning sites different rates via:
  - Discrete Γ distribution (approximated from Γ using GLQ)
  - CAT model (typically 25-30 DPP assigned categories)
  - Invariant sites
  - Partitioning (if justifiable)
- Heterotachy (ATRV) can be incorporated by using covarion models and lineage specific models

$$P(Aligment|Model) =: L_{Alignment}(Model)$$

$$P(D|M) =: L_D(M)$$

$$\hat{M} =: \underset{M}{argmax}\, L_D(M)$$

$$P(D|M) = \prod_i P(d_i|M)$$



Maximum-Likelihood Inference

- Find the most likely model (tree topology, branch lengths and substitution model) for the data (MSA) (see above)
- Optimisation problem but with highly correlated parameters, discrete topologies and branch length optimisation is topology dependent
- Topology optimisation via NNI and SPR
- ML is consistent when model assumptions are fulfilled
- Uses all data, tests topologies and better br optimisation
- Most appropriate when inferential signal is strong and datasets are large (RAxML current SoA)
- Inference robustness test via pseudoreplicate resampling

Bayesian Inference

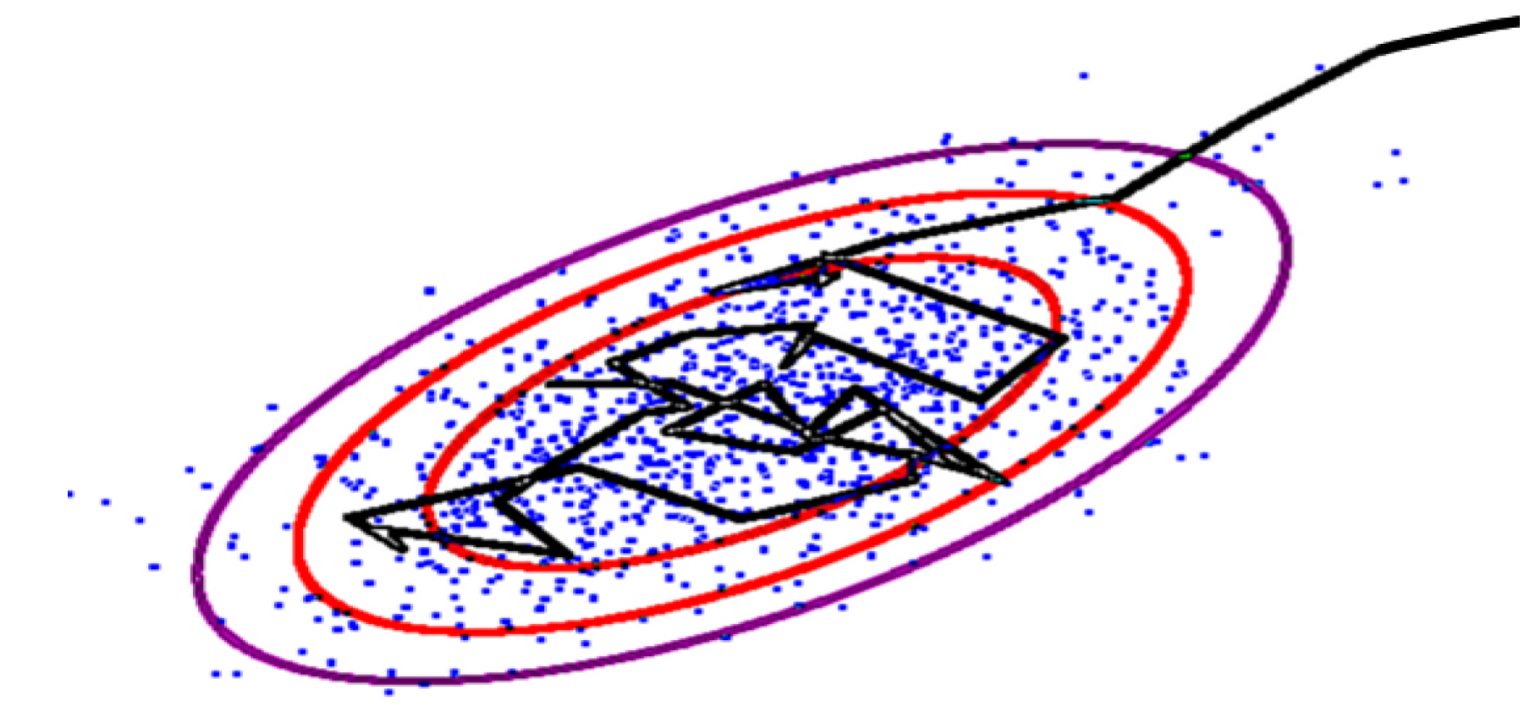
- Summarise posterior probability density of model (topology, branch lengths and substition model)
- Marginal probability term computationally intractable therefore compute ratio between models
- Sampling of posterior densities using MCMC with acceptance of new states dependent on ratios
- Run multiple chains saving model state every n generations and assess convergence
- Once convergence summarise model
- Metropolis-Coupling of MCMC with heated chains allows faster traversal of parameter space
- Posterior probabilities assigned to parameter values means support values are 'built-in'
- Best with low signal to parameter ratio i.e. complex models or little signal

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

$$P(M|D) = \frac{P(D|M) \cdot P(M)}{\int_{M'} P(D|M') \cdot P(M')dM'}$$

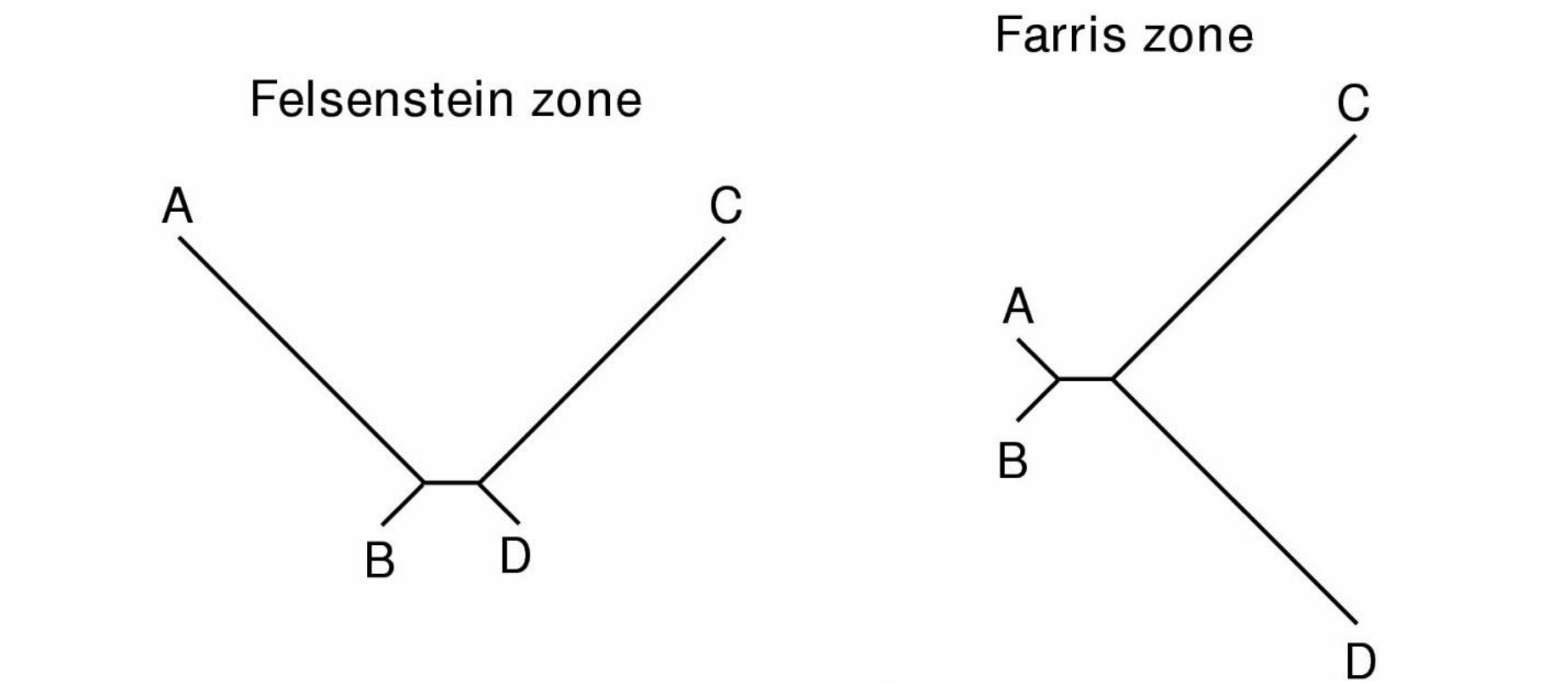
$$\frac{P(M_A|D)}{P(M_B|D)} = \frac{\frac{P(D|M_A)\cdot P(M_A)}{\int_{M'} P(D|M')\cdot P(M')dM'}}{\frac{P(D|M_B)\cdot P(M_B)}{\int_{M'} P(D|M')\cdot P(M')dM'}}$$

$$\frac{P(M_A|D)}{P(M_B|D)} = \frac{P(D|M_A) \cdot P(M_A)}{P(D|M_B) \cdot P(M_B)}$$



Common Problems and Pitfalls

- Hidden paralogy (misidentification of paralogs as orthologs often due to loss of one copy) - improve taxon sampling
- Long Branch Attraction (LBA) - use ML/BI and attempt to break up long branches by adding intermediate taxa. In extreme cases remove long branch from MSA and test topology change
- Poor taxon sampling - amplifies other artefacts (e.g. LBA) and can produce misleading relationships
- Overreliance on a single methodology - most journals now expect trees to be built via ML and BI methodologies with summary of support values
- Differences between different models and inferences can be informative - try many variants of reconstruction
- Incorrect usage of programs - bioinformatics documentation is generally poor unfortunately however mailing lists can be useful



References
Rivas, E., Eddy, S. R., and Haussler, D. (2008). Probabilistic phylogenetic inference with insertions and delet-ions. PLoS Computational Biology, 4(9):e1000172.
Paul O. Lewis Woods Hole Molecular Evolution Workshop 2012 Lectures
Alexander Stamatakis RAxML 7.3 Manual
Derrick Zwickl Woods Hole Molecular Evolution Workshop 2012 Lectures
John Hulsenbeck MrBayes 3.2 Manual