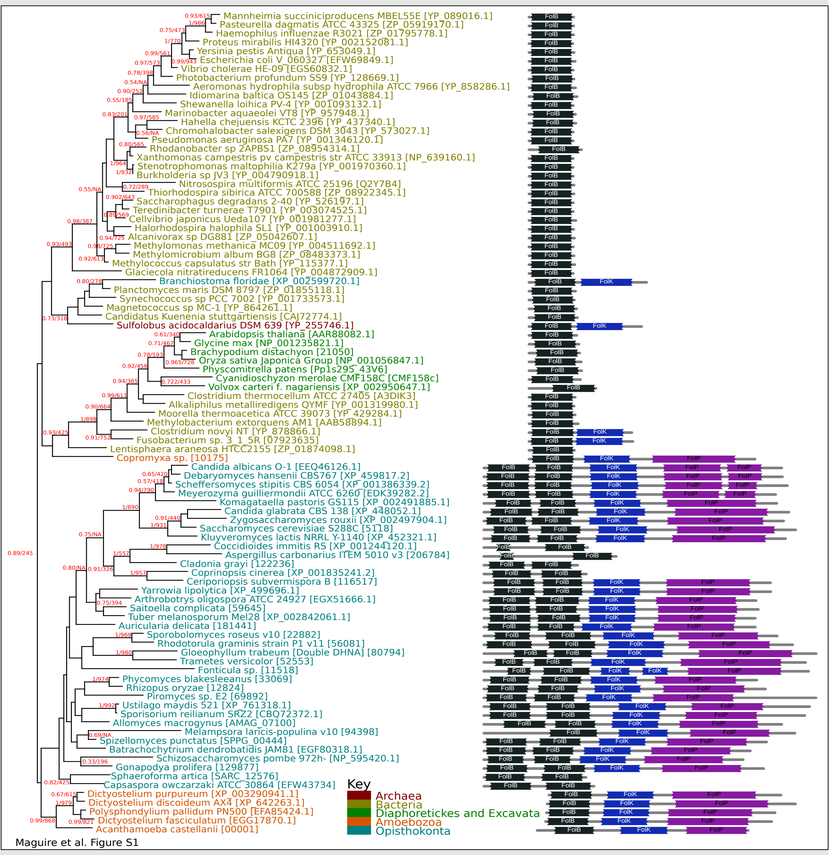
Basic Principles of Molecular Phylogenetics

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:

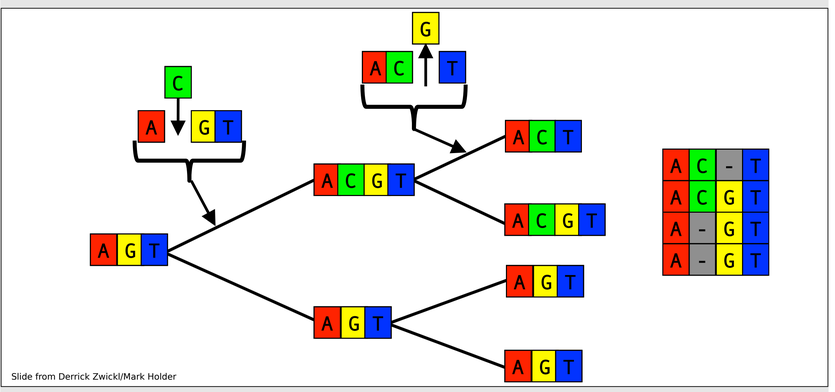
- Reconstruction of relationships between sequences and/or taxa
- Sequence identification
- Discovery of Horizontal Gene Transfer (HGT) events
- Exploration of equence and functional divergence
- Identification of evolutionary innovations
- Integral to many bioinformatic algorithms/ applications



Maguire et al. Figure S1

Multiple Sequence Alignment

- Model of positional homology from which tree is constructed
- Most important part of analysis
- Dynamic programming algorithms using serial pairwise alignments followed by iterative improvement
- Simultaneous inference of MSA and Phylogeny (e.g. BAliPhy) is potentially optimal solution but highly computationally expensive



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.*)

Distance and Parsimony methods

- Earliest methods
- Fast computationally but prone to bias and inconsistency
- Main utility as a diagnostic tool and starting point

Substitution model selection criteria

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

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

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

$$BIC_i = -2lnL_i + KlnN$$

Substitution models

- Model based methods (ML and BI) require a statistical model of sequence evolution (i.e. P(G<-->T) etc.)
- Multiple test criteria (above) for model selection most of which are implemented in tools to aid selection
- Nucleotide models are typically mechanistic (JC69/ TN93) and nested within GTR (if all K are equal GTR = JC69)
- Protein models (JTT/LG) typically empirical (observed rates in existing datasets) due to many state changes (380)
- Models also incorporate changes in the rate of evolution in different sites (ASRV) or lineages (ATRV) by a variety of methods

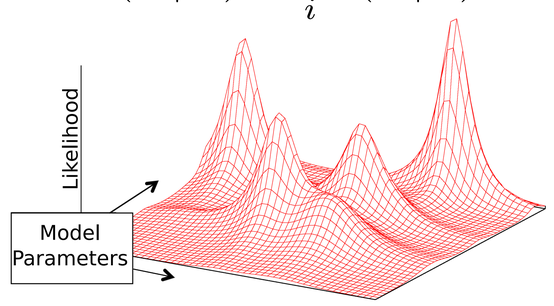
Maximum Likelihood Inference

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

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

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

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



Maximum-Likelihood Inference (ML)

- Find the most likely phylogenetic model (tree topology, branch lengths and substitution model) for the data (MSA) (see above)
- Optimisation problem but with highly correlated parameters, discrete topologies and topology dependent branch length optimisation
- Topology and branch length are optimised via Nearest Neighbour Interchange (NNI) and Subtree Pruning and Regrafting (SPR) perturbations
- ML is consistent when model assumptions are fulfilled
- Most appropriate when inferential signal is strong and datasets are large (RAxML current SoA)

Bayesian Inference (BI)

- Generates a posterior probability density of a phylogenetic model based on priors and data likelihood
- Posterior probability density is sampled using Monte-Carlo Markov-Chains
- MCMC randomly perturb model parameters and accept or reject new parameter state by comparison of likeihood with old state
- MCMC eventually discover likelihood peak and generate a pool of 'plausible' phylogenetic models
- This distribution can then be summarised to recover the most probable model
- Best with low signal to parameter ratio i.e. complex models or little signal
- Allows incorporation of extra knowledge (via informative priors)

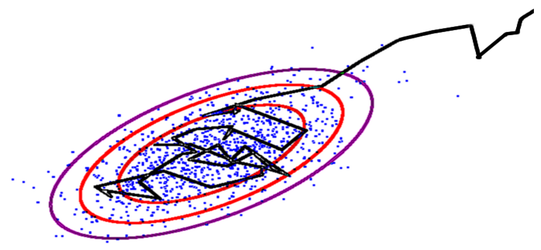
Bayesian Inference

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

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

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

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



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