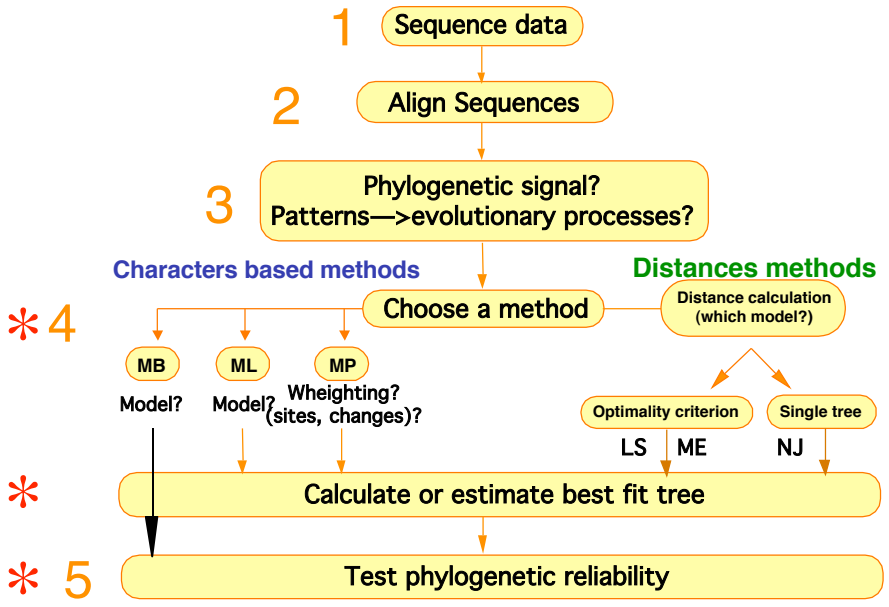
FROM PROTEIN SEQUENCES TO PHYLOGENETIC TREES

Robert Hirt
Department of Zoology, The Natural History Museum, London

Agenda

- **Remind you that molecular phylogenetics is complex**
  - **the more you know about the compared proterins and the method used, the better**
    - **Try to avoid the black box approach a much as possible!**
- **Give an overview of the phylogenetic methods and software used with protein alignments - some practical issues…**

From DNA/protein sequences to trees



Modified from Hillis et al., (1993). Methods in Enzymology 224, 456-487

Phylogenies from proteins

- Parsimony
- * Distance matrices
- * Maximum likelihood
- * Bayesian methods

Phylogenetic trees from protein alignments

- **Distance methods - model for distance estimation**
  - **Simple formula (e.g. Kimura,, use of Dij)**
  - **Complex models**
    - **Probability of amino acid changes - Mutational Data Matrices**
    - **Site rate heterogeneity**
- **Maximum likelihood and Bayesian methods- MDM based models are used for lnL calculations of sites -> lnL of trees**
    - **Site rate heterogeneity**
    - **Homogenous versus heterogeneous models**
    - **Estimations of data specific rate matrices (amino acid groupings - GTR like)**

Software: an overview

- **CLUSTALX - distance**
- **PHYLIP - distance, MP, and ML methods (and more)**
  - **Some complex protein models**
    - **PAM, JTT ± site rate heterogeneity**
  - **Bootstrapping - bootstrap support values**
- **PUZZLE - distance and a ML method**
  - **ML - quartet method**
  - **Complex protein models**
    - **JTT, WAG…matrices ± site rate heterogeneity**
  - **From quartets to n-taxa tree - PUZZLE support values**
  - **Some sequence statistics - aa frequency and heterogeneity between sequences**
  - **Tree comparisons - KH test**
- **MRBAYES - Bayesian**
  - **Complex protein models**
    - **JTT, WAG…matrices ± site rate heterogeneity**
    - **Data partitioning**
  - **Posteriors as support values**
- **P4**
  - **All the things you can dream off… almost… ask Peter Foster**
  - **Heterogeneous models among taxa or sites**
  - **Estimation of rate amino acid rate matrices for grouped categories (6x6 rate matrices can be calculated - much easier then 20x20)**

Software: alignment format

1) PHYLIP format (PHYLIP, PUZZLE, PAUP can read and export this format)

```
 4 500
Human   AAGGHTAG…TCTWC
Mouse   ATGGHTAA…TCTWC
Cat     ATGGKTAS…TCTWC
Fish    ASGGRTAA…SCTYC
```

2) NEXUS format (PAUP, MRBAYES : only a subset of NEXUS' diversity)

```
#NEXUS
begin data;
  dimensions ntax=4 nChar=500;
  format datatype=protein gap=-
  missing=?;
  matrix
Human   AAGGHTAG…TCTWC;
Mouse   ATGGHTAA…TCTWC;
Cat     ATGGKTAS…TCTWC;
Fish    ASGGRTAA…SCTYC;
End;
```

3) GDE, PAUP, CLUSTALX, READSEQ…

- **Can read and export various format including PHYLIP and NEXUS…**

PHYLIP3.6

- **Protpars**: parsimony
- **Protdist**: models for distance calculations:
  - PAM1, JTT, Kimura formula (PAM like), others...
  - Correction for rate heterogeneity between sites! Removal of invariant sites? (not estimated, see PUZZLE!)
- **NJ and LS** distance trees (± molecular clock)
- **Proml**: protein ML analysis (no estimation of site rate heterogeneity - see PUZZLE)
  - Coefficient of variation (CV) versus alpha shape parameter $CV=1/alpha^{1/2}$
- **Bootstrapping**

Distance methods

A two step approach - two choices!

1) Estimate all pairwise distances

Choose a method (100s) - has an explicit model for sequence evolution

- **Simple formula**
- **Complex models - PAM, JTT, site rate variation**

2) Estimate a tree from the distance matrix

Choose a method: with (ME, LS) or without an optimality criterion (NJ)?

Simple and complex models

dij = -Ln (1 - Dij - (Dij^2/5)) **(Kimura)**

Simple and fast but can be unreliable - underestimates changes, hence distances, which can lead to misleading trees - PHYLIP, CLUSTALX

Dij is the fraction of residues that differs between sequence i and j (Dij = 1 - Sij)

dij = ML [P(□), (□, `pinv`), X_{ij}] **(bad annotation!)**

ML is used to estimate the dij based on the sequence alignment and a given model (MDM, gamma shape parameter and pinv - PHYLIP, PUZZLE. Each site is used for the calculation of dij, not just the Dij value.

More realistic complexity in relation to protein evolution and the subtle patterns of amino acid exchange rates…

Note: the values of the different parameters (alpha+pinv) have to be either estimated, or simply chosen (MDM), prior the dij calculations

1) Choosing/estimating the parameter of a model

1) Mutation Data Matrices: PAM, JTT, WAG…

- **What are the properties of the protein alignment (% identity, amino acid frequencies, globular, membrane)?**
- **Can be corrected for the specific dataset amino acid frequencies (-F)**
- **Compare ML of different models for a given data and tree**

2) Alpha and pinv values have to be estimated on a tree

- **PUZZLE can do that. Reasonable trees give similar values…**

2) Inferring the phylogenetic trees from the estimated dij

a) Without an optimality criterion

- **Neighbor-joining (NJ) (NEIGHBOR)**

  Different algorithms exist - improvement of the computing

  If the dij are additive, or close to it, NJ will find the ME tree…

b) With an optimality criterion

- **Least squares (FITCH)**
- **Minimum evolution (in PAUP)**

Fitch Margoliash Method 1968

- Seeks to minimise the weighted squared deviation of the tree path length distances from the distance estimates - uses an objective function

$$E = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} wij \; |dij - pij|^{\square}$$

E = the error of fitting dij to pij
T = number of taxa
if □ = 2 weighted least squares
wij = the weighting scheme

dij = F(Xij) pairwise distances estimate - from the data using a specific model (or simply Dij)
pij = length of path between i and j implied on a given tree
dij = pij for additive datasets (all methods will find the right tree)

Minimum Evolution Method

- For each possible alternative tree one can estimate the length of each branch from the estimated pairwise distances between taxa (using the LS method) and then compute the sum (S) of all branch length estimates. The minimum evolution criterion is to choose the tree with the smallest S value

$$S = \sum_{k=1}^{2T-3} \mathbf{V_k}$$

With $\mathbf{V_k}$ being the length of the branch k on a tree

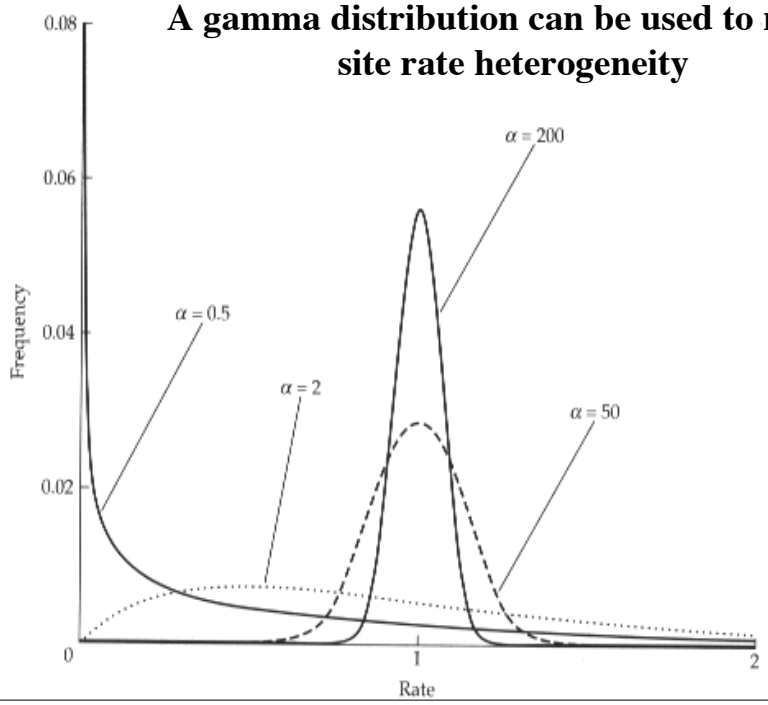
Distance methods

- **Advantages:**
  - Can be fast (NJ)
  - Some distance methods (LogDet) can be superior to more complex approached (ML) in some conditions
  - Distance trees can be used to estimate parameter values for more complex models and then used in a ML method
  - Provides trees with branch length
- **Disadvantages:**
  - Can loose information by reducing the sequence alignment into pairwise distances
  - Can produce misleading (like any method) trees in particular if distance estimates are not realistic (bad models), deviates from additivity

TREE-PUZZLE5.0

- Protein maximum likelihood method using "quartet puzzling"
  - With various protein rate matrices (JTT, WAG…)
  - Can include correction for rate heterogeneity between sites - pinv + gamma shape (can estimates the values)
  - Can estimate amino acid frequencies from the data
  - List site rates categories for each site (2-16)
  - Composition statistics
  - Molecular clock test
  - Can deal with large datasets
- Can be used for ML pairwise distance estimates with complex models - used with puzzleboot to perform bootstrapping with PHYLIP

A gamma distribution can be used to model site rate heterogeneity



Yang 1996
TREE, 11,
367-372

TREE-PUZZLE5.0

The quartet ML tree search method has four steps:

1) **Parameters (pinv-gamma) are estimated on a NJ n-taxa tree**
2) **Calculate the ML tree for all possible quartets (4-taxa)**
3) **Combine quartets in a n-taxa tree (puzzling step)**
4) **Repeat the puzzling step numerous times (with randomised order of quartet input)**
5) **Compute a majority rule consensus tree from all n-trees - has the puzzle support value**

Puzzle support values are not bootstrap values!

TREE-PUZZLE5.0

- **Models for amino acid changes:**
  - **PAM, JTT, BLOSUM64, mtREV24, WAG (with correction for amino acid frequencies)**
  - **Correction for specific dataset amino acid frequencies**
  - **Discrete gamma model for rate heterogeneity between sites 4-16 categories.**

    -> output gives the rate category for each site. Can be used to partition your data and analyse them separately…
- **Taxa composition heterogeneity test**
- **Molecular clock test**

TREE-PUZZLE5.0

- Can be used to calculate pairwise distances with a broad diversity of models - puzzleboot (Holder & Roger)
  - Can be used in combination with PHYLIP programs for bootstrapping:
  - SEQBOOT
  - NJ or LS…
  - CONSENSE

TREE-PUZZLE5.0

- Advantages:
  - Can handle larger numbers of taxa for maximum likelihood analyses
  - Implements various models (BLOSUM, JTT, WAG…) and can incorporate a correction for rate heterogeneity (pinv+gamma)
  - Can estimate for a given tree the gamma shape parameter and the fraction of constant sites and attribute to each site a rate category
- Disadvantages:
  - Quartet based tree search - amplification of the long branch attraction artefact within each quartet analysis?

MrBayes 3.0

- Bayesian approach
  - Iterative process leading to improvement of trees and model parameters and that will provide the most probable trees (and parameter values)
- Complex models for amino acid changes:
  - PAM and JTT, WAG (with correction for amino acid frequencies, but you have to type it!?!?!)
  - Correction for rate heterogeneity between sites (pinv, discrete gamma, site specific rates)
- Powerful parameter space search
  - Tree space (tree topologies)
  - Shape parameter (alpha shape parameter, pinv)
  - Can work with large dataset
  - Provides probabilities of support for clades

MrBayes 3.0

- MrBayes will produce a population of trees and parameter values - obtained by a Markov chain (mcmcmc). If the chain is working well these will have converged to "probable" values
- In practice we plot the results of an mcmcmc to determine the region of the chain that converged to probable values. The "burn in" is the region of the mcmcmc that is ignored for calculation of the consensus tree
  - Trees and parameter values from the region of equilibrium are used to estimate a consensus tree
  - The number of trees recovering a given clade corresponds to the posterior for that clade, the probability that this clade exists
  - The mcmcmc uses the lnL function to compare trees

  Most methods provide a single tree and parameters value
  - Bootstrapping provide a distribution of tree topologies
  - Puzzling steps also provides a distribution tree topologies
    - Bootstrap values - Puzzle support values - Posteriors values ???
    - But not to sure how to interpret these different support values. Posteriors are typically higher then bootstrap and puzzle support values?!?

MrBayes 3.0: an example

```
#NEXUS
begin data;
 dimensions ntax=8 nChar=500;
 format datatype=protein gap=- missing=?;
 matrix

Etc…

Begin mrbayes;
        log start filename=d.res.nex.log replace;
        prset aamodelpr=fixed(wag);
        lset rates=invgamma Ngammacat=4;
        set autoclose=yes;
        mcmc ngen=5000 printfreq=500 samplefreq=10 nchains=4 savebrlens=yes
        startingtree=random filename=d.res.nex.out;
  quit;
end;
[
Begin mrbayes;
log start filename=d.res.nex.con.log replace;
sumt filename=d.res.nex.out.t burnin=150 contype=allcompat;
end;
]
```

Block I

Block II

A Bayesian analysis

-Propose a starting tree topology and parameters values (branch length, alpha, pinv), calculate lnL

-Change one of these and compare the lnL with previous proposal

-If the lnL is improved accept it

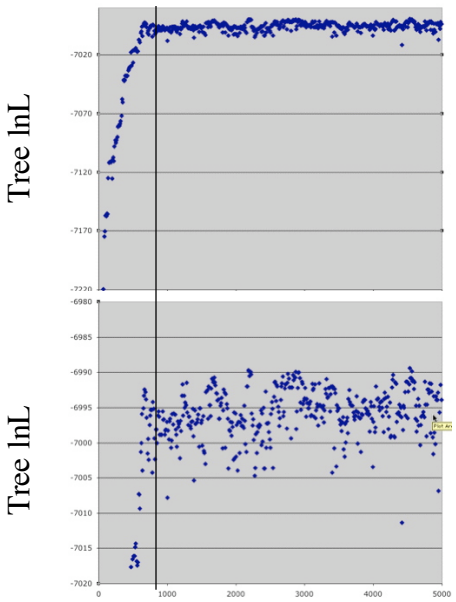
-If not, accept it only sometimes

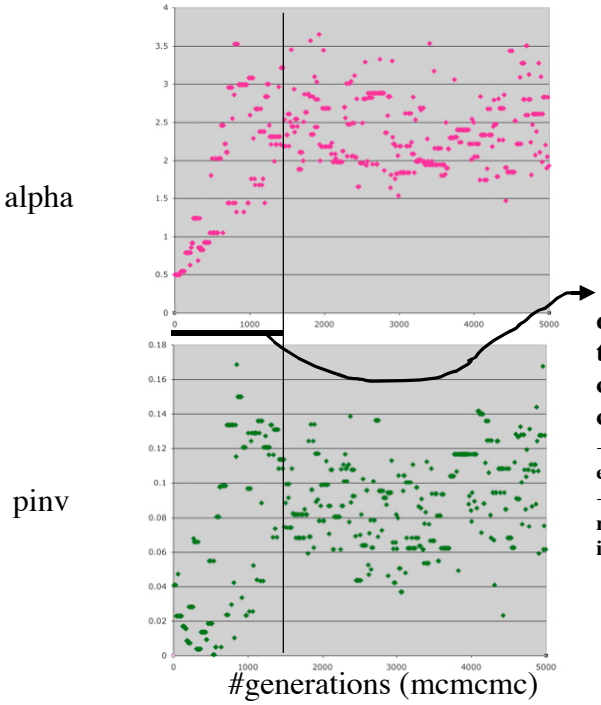
-Do many of these…

-Plot the change of lnL in relationship to the number of generations run

-Determine the region where the chain converged and calculate the consensus tree for that region

-> consensus tree with posteriors for clade support



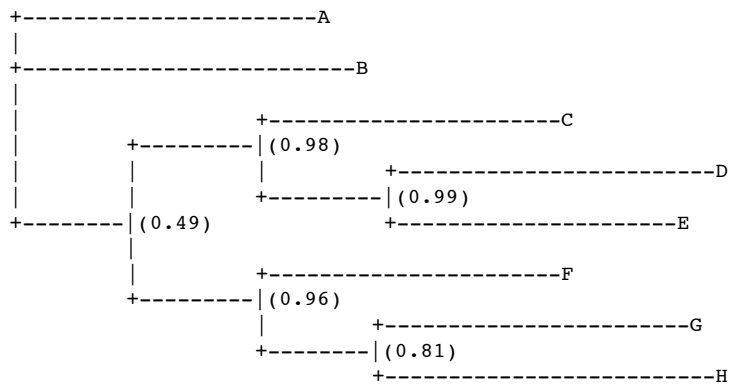


"Burn in" determines the trees to be ignored for consensus tree calculation

-Was the chain run long enough?

-Do we get the same result from an independent chain?

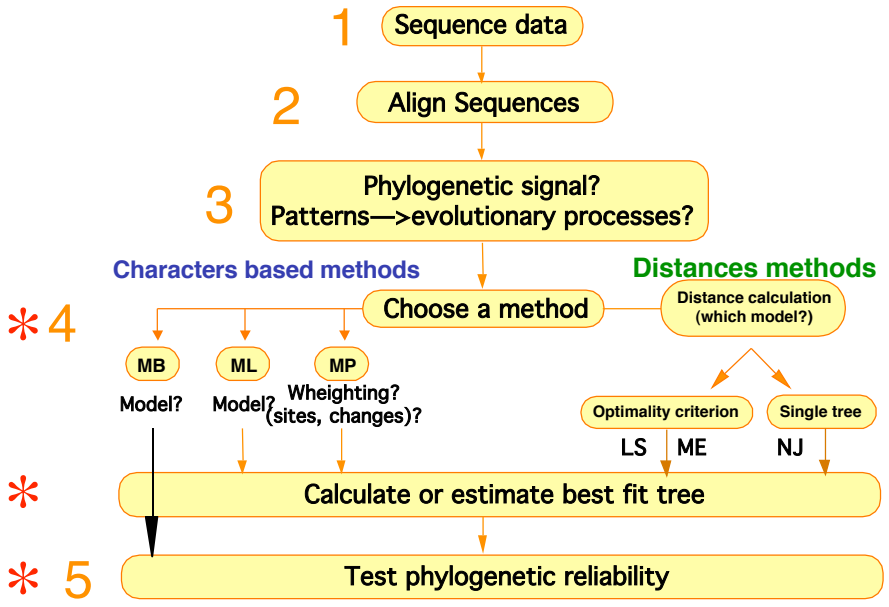
Consensus tree with a burn in of 1500 (150)
Showing posterior values for the different clades - probability for a given clade to be correct (for the given data and method used!!!)



Summary

- **No single program allows thorough phylogenetic analyses of protein alignments**
- **Combination of PHYLIPv3.6, TREE-PUZZLEv5.1, MRBAYESv3 and P4 allow detailed protein phylogenetics**
- **Remember that experimenting with your data and available methods/models can lead to interesting and biologically relevant results (data <-> method)**
  - **Incorporate site rate heterogeneity correction in the model or reduce heterogeneity by data editing (with and without invariant sites?)**
  - **Partitioning of the alignment (variant - various rates, invariant sites, secondary structure, protein domains…)**
  - **Amino acid groupings (6 categories - GTR like)**
  - **LogDet for proteins?**
- **Do not take support values as absolute. Any support values is for a given method and data, only!**

From DNA/protein sequences to trees



Modified from Hillis et al., (1993). Methods in Enzymology 224, 456-487