

Phylogenetic inference from Proteins

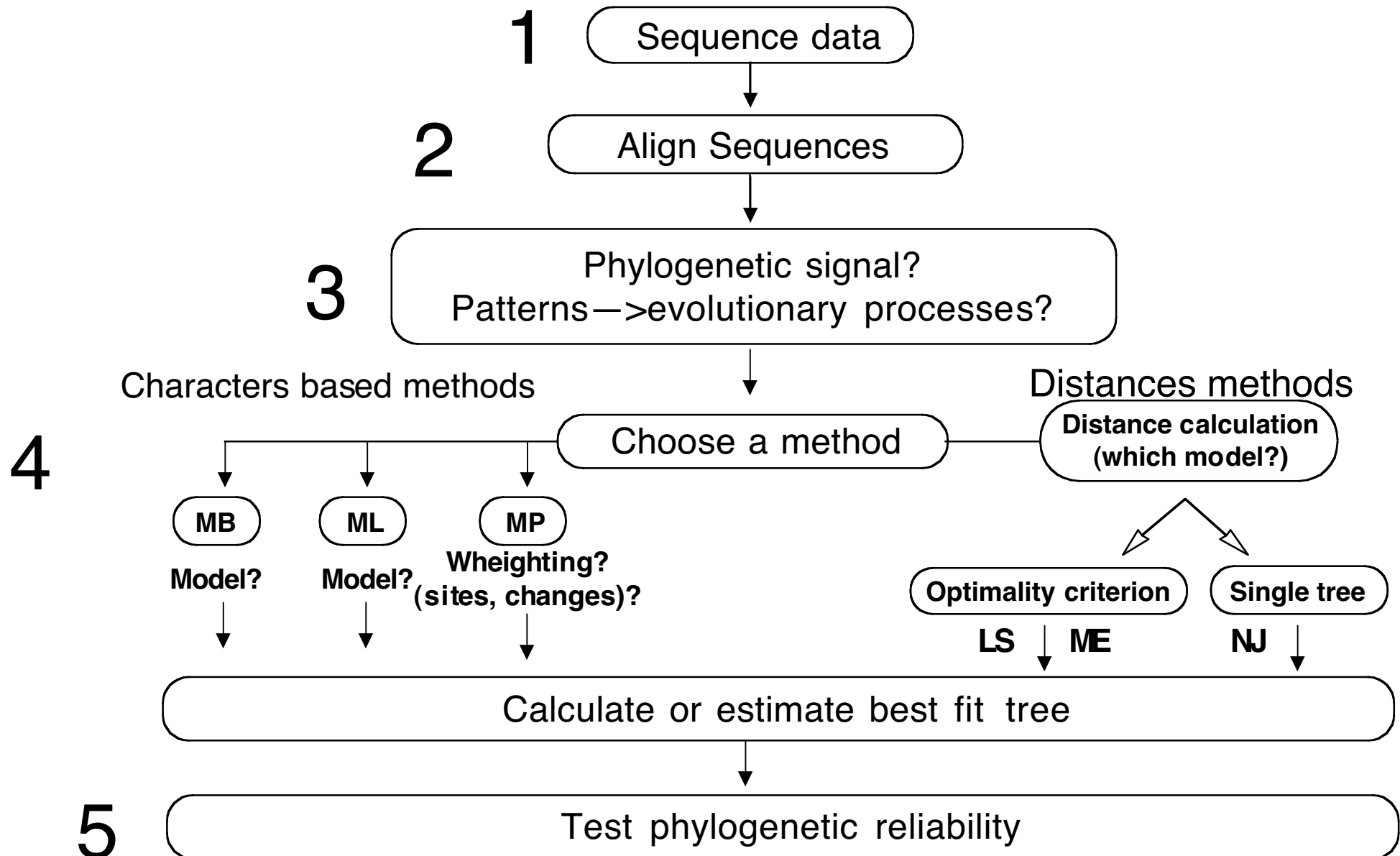
Aims

- **Remind you that phylogenetics is complex - the more you know about the compared sequences the better**
- **Introduce specific issues in protein sequence comparisons (models)**
- **Give an overview of the phylogenetic methods used with protein alignments**

Content

- **Comparing DNA and/or protein sequences?**
- **Empirical models (rate matrices) for protein sequence comparisons**
- **Inferring phylogenies from protein sequences - some programs**

The five steps in phylogenetics dancing



Why protein phylogenies?

- **For historical reasons - first sequences...**
- **Most genes encode proteins...**
- **To study protein structure,
function and evolution**
- **Comparing DNA and protein based
phylogenies can be useful**
 - **Different genes - e.g. 18S rRNA versus EF-2 protein**
 - **Protein encoding gene - codons versus amino acids**

**Protein were the first molecular sequences
to be used for phylogenetic inference**

- **Fitch and Margoliash (1967).**

Construction of phylogenetic trees.

Science 155, 279-284.

Character states in DNA and protein alignments

- DNA has four (five): A, C, G, T, and \pm indels**
 - Proteins have 20 (21): A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y and \pm indels**
- > more information in protein alignments?**

DNA->Protein: the code

- **3 nucleotides (a codon) code for one amino acid**
- **Degeneracy of the code: most amino acids are coded by several codons (61 codons -> codon based models)**
- > **more information in DNA?**

Codon degeneracy: protein alignments can be used to guide DNA alignments

Glu-Gly-Ser-Ser-Trp-Leu-Leu-Leu-Gly-Ser

Glu-Gly-Ser-Ser-Tyr-Leu-Leu-Ile-Gly-Ser

Asp-Gly-Ser-Ala-Trp-Leu-Leu-Leu-Gly-Ser

Asp-Gly-Ser-Ala-Tyr-Leu-Leu-Ala-Gly-Ser

GAA-GGA-AGC-TCC-TGG-TTA-CTC-CTG-GGA-TCC

GAG-GGT-TCC-AGC-TAT-CTA-TTA-ATT-GGT-AGC

GAC-GGC-AGT-GCA-TGG-TTG-CTT-TTG-GGC-AGT

GAT-GGG-TCA-GCT-TAC-CTC-CTG-GCC-GGG-TCA

DNA->Protein: code usage

- **Difference in codon usage can lead to large base composition bias - in which case one often needs to remove the 3rd codon... and possibly the 1st**
- **Comparing protein sequences can reduce the compositional bias problem**
 - > **more information in DNA or protein?**

Models for DNA and Protein evolution

- **DNA: 4 x 4 rate matrices**

Easy to estimate (can be combined with tree search)

- **Protein: 20 x 20 matrices**

More complex: time and estimation problems (rare changes?) -> empirical models from large datasets are typically used

- **Codon: 61 x 61 models**

Summary

- **Phylogenetics is complex: involves numerous steps, where choices have to be made**
- **It is often useful to use protein alignments to guide DNA alignments**
- **Non of the methods we use are perfect: comparing DNA and protein alignment based analyses can be useful - different methods are able to deal with different limitation(s) of a dataset**
- **Congruence between different genes is very important to give us some confidence in a tree**

Evolutionary models for amino acid changes

- All methods have explicit or implicit evolutionary models**
- Models for amino acid changes are typically 20 x 20 matrices**
- Rate heterogeneity between sites**

Models for protein evolution: amino acid exchange matrices

- Database searches**
 - Blast programs**
- Sequence alignments**
 - ClustalW**
- Phylogenetics**

Agenda

- **What are we comparing? Protein sequences - some basic feature**
- **Protein structure/function and its impact on patterns of mutations**
- **Amino acid exchange matrices (models): where do they come from?**

Proteins and amino acids

- **Proteins determine shape and structure of cells and carry most catalytic processes - 3D**
- **Proteins are polymers of 20 different amino acids**
- **Amino acids sequences determine the structure (2ndary, 3ary...) and function of the protein**
- **Amino acids can be categorized by their side chain physicochemical properties**
 - Polarity (hydrophobic versus hydrophilic, +/- charges)**
 - Size (small versus large)**

Amino acid physico-chemical properties

- Size**

- Polarity**

 - hydrophilic (polar, +/- charges)**

 - hydrophobic (non polar)**

Amino acids categories 1:

Doolittle (1985). Sci. Am. 253, 74-85.

Small polar: S, G, D, N

Small non-polar: T, A, P, C

Large polar: E, Q, K, R

Large non-polar: V, I, L, M, F

Intermediate polarity: W, Y, H

Amino acids categories 2

Sulfhydryl: C

Small hydrophilic: S, T, A, P, G

Acid, amide: D, E, N, Q

Basic: R, K, H

Small hydrophobic : M, I, L, V

Aromatic: F, Y, W

Amino acid physico-chemical properties

- Major factor in protein folding
- Key to protein functions
 - > Major influence in pattern of amino acid mutations

As for Ts versus Tv in DNA sequences, some amino acid changes are more common than others: very important for sequence comparisons (alignment and phylogenetics!)

Small <—> small > small <—> big

Estimation of relative rates of residue replacement (models of evolution)

- **Differences/changes in protein alignments can be pooled and patterns of changes investigate.**
Selected sequence, alignment and counting method dependent!
- **Patterns of changes give insights into the evolutionary processes underlying protein diversification -> estimation of evolutionary models - How general is such a model?**
- **Choice of protein evolutionary models can be important for the sequence analysis we perform (database searching, sequence alignment, phylogenetics)**

Models of proteins evolution

20 x 20 matrices (relative rates of residue replacement $i \leftrightarrow j$)

- **Identity matrix (all rates are equal)**
- **Genetic code matrix**

- **Mutational data matrices (MDMs) - 100s...** (related to similarity scoring matrices)

MDMs are empirical models of amino acid replacement!

Models of proteins evolution: common assumptions

- All amino acid sites in an alignment evolve independently**
- The Markov process is assumed to be:**
 - stationary** (same composition for all taxa)
 - homogenous** (same model across sites & time)
 - reversible** (symmetrical matrices - no root!)

Amino acid substitution matrices based on observed substitutions: empirical models

- **Summarise the substitution pattern from large amount of existing data**
- **Selection of a protein dataset**
(globular/mitochondrial proteins?)
- **Selection of a counting method and the counted changes** (tree dependent/independent, restriction on the sequence divergence...)

Amino acid substitution matrices based on observed substitutions

- **Sequence based matrices**
PAM, JTT, BLOSUM, WAG...
- **Structure based matrices**
STR (for highly divergent sequences)

Estimation of relative rates of amino acid changes

- **Tree independent patterns - Pairwise comparisons (Blosum)**
- **Tree based patterns**
 - Maximum parsimony based methods (PAM, JTT)**
 - Maximum likelihood based methods (metREV, WAG)**

Amino Acid exchange matrices

$$\begin{pmatrix} - & s_{1,2} & s_{1,3} & \dots & s_{1,20} \\ s_{1,2} & - & s_{2,3} & \dots & s_{2,20} \\ s_{1,3} & s_{2,3} & - & \dots & s_{3,20} \\ \dots & \dots & \dots & \dots & \dots \\ s_{1,20} & s_{2,20} & s_{3,20} & \dots & - \end{pmatrix}$$

$$X \text{diag}(\pi_1, \dots, \pi_{20}) = Q \text{ matrix}$$

Q Rate matrix

Q_{ij} Instantaneous rates of change of amino acids

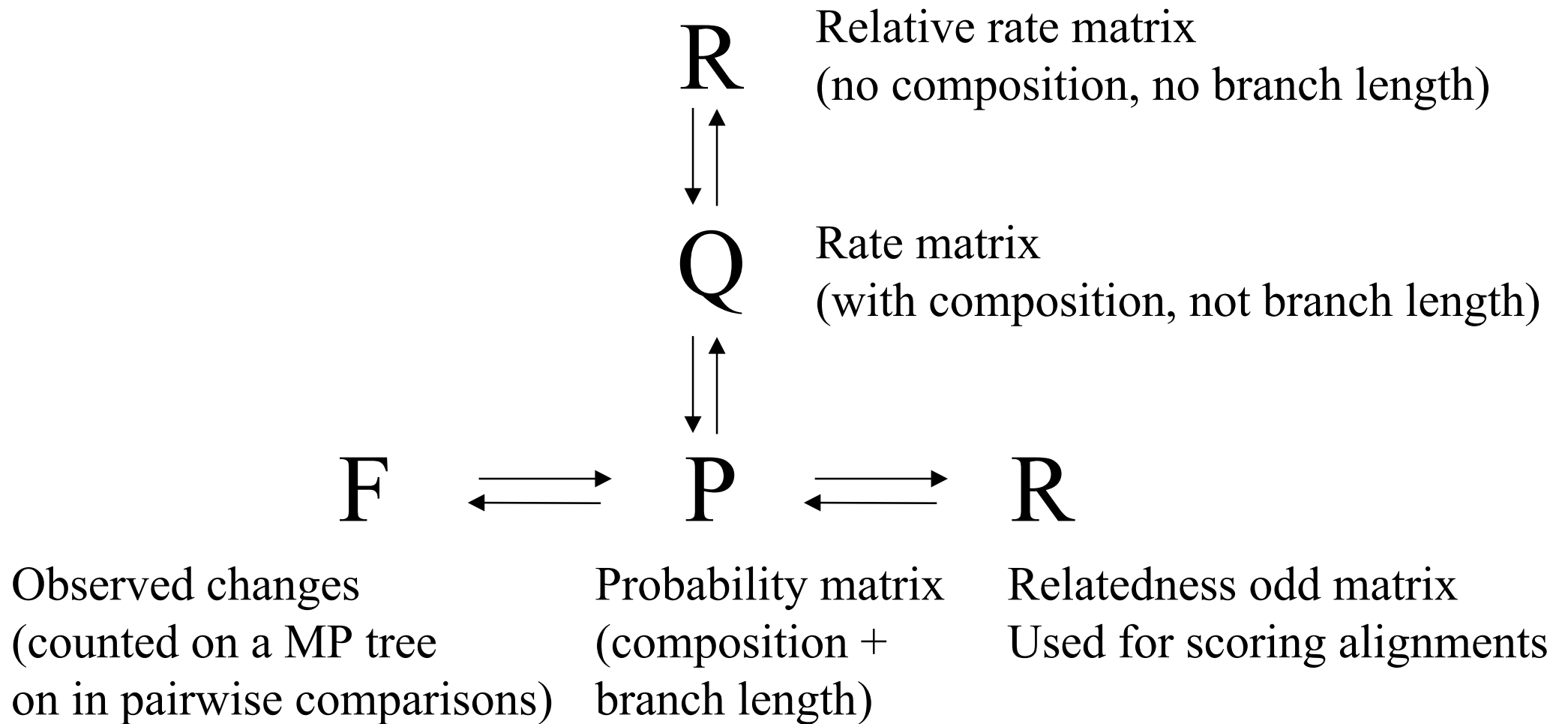
s_{ij} Exchangeabilities of amino acid pairs ij

$s_{ij} = s_{ji}$ Time reversibility

π_i Stationarity of amino acid frequencies

(typically the observed proportion of residues in the dataset)

Amino Acid exchange matrices



Modified from Peter Foster

The PAM and JTT matrices

- **PAM - Dayhoff et al. 1978**
 - **Nuclear encoded genes, 72 protein families consisting of 1,300 sequences**
- **JTT - Jones et al. 1992**
 - **59,190 accepted point mutations from 16,300 proteins**

The BLOSUM matrices

- **BLOcks SUBstitution Matrices**

The matrix values are based on 2000 conserved amino acid patterns from 500 families of related sequence (blocks)

Pairwise comparisons

- > more efficient for distantly related proteins
- > more agreement with 3D structure data

BLOSUM62 Amino Acid Log-odd Substitution Matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W			
C	9																				C	sulfhydryl	
S	-1	4																				S	
T	-1	1	5																			T	
P	-3	-1	-1	7																		P	small hydrophilic
A	0	1	0	-1	4																	A	
G	-3	0	-2	-2	0	6																G	
N	-3	1	0	-2	-2	0	6															N	
D	-3	0	-1	-1	-2	-1	1	6														D	acid, acid-amide
E	-4	0	-1	-1	-1	-2	0	2	5													E	and hydrophilic
Q	-3	0	-1	-1	-1	-2	0	0	2	5												Q	
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8											H	
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5										R	basic
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5									K	
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5								M	
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4							I	small hydrophobic
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4						L	
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4					V	
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6				F	
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7			Y	aromatic
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11		W	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W			

$x_{ij} < 0$ freq. less than chance
 $x_{ij} = 0$ freq. expected by chance
 $x_{ij} > 0$ freq. great than chance

The WAG matrix

Whelan and Goldman (2001) Mol. Biol. Evol. 18, 691

- **Globular protein sequences**
 - **3,905 sequences from 182 protein families**
- **Produced a phylogenetic trees for every family and used maximum likelihood to estimate the relative rate values in the rate matrix (overall lnL over 182 different trees)**
 - **Better fit of the model with most data** (significant improvement of the lnL of a tree when compared to PAM or JTT matrices)

Summary

- **Many amino acid rate matrices exist and one needs to choose one, among many, for protein comparisons (alignment, phylogenetics...) do not hesitate to experiment!**
- **One should make a rational choice:**
 - **How was the rate matrix produced?**
 - **What are the structural features of the sequences you are comparing? Globular/membrane protein?**
 - **Level of sequence identity**

Always try to correct for rate heterogeneity between sites!

Phylogenies from proteins alignments:

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

Phylogenetic trees from protein alignments

- **Parsimony based methods - unweighted/weighted**
- **Distance methods - model for distance estimation**
 - **probability of amino acid changes, site rate heterogeneity**
- **Maximum likelihood - model for ML calculations**
 - **probability of amino acid changes, site rate heterogeneity**

Trees from protein alignment: Parsimony methods - cost matrices

- **All changes weighted equally**
- **Weights based on the minimal number of amino acid substitutions (PHYLIP -PROTPARS)**
- **Weights based on observed frequency of amino acid substitutions in alignments**
- **Weights based on physico-chemical properties of amino acids**

Parsimony: unweighted matrix for amino acid changes

Ile -> Leu cost = 1

Trp -> Asp cost = 1

Ser -> Arg cost = 1

Lys -> Asp cost = 1

Parsimony: weighted matrix for amino acid changes 1: minimal changes of amino acids

Ile -> Leu cost = 1

Trp -> Asn cost = 3

Ser -> Arg cost = 2

Lys -> Asp cost = 2

Phylogenetic trees from protein alignments

- **Parsimony based methods - unweighted/weighted**
- **Distance methods - model for distance estimation**
 - **probability of amino acid changes, site rate heterogeneity**
- **Maximum likelihood - model for ML calculations**
 - **probability of amino acid changes, site rate heterogeneity**

Distance methods

A two step approach - two choices!

1) Estimate all pairwise distances

Choose a method (100s) - has an explicit model for sequence evolution (PAM, JTT, WAG...)

2) Estimate a tree from the distance matrix

Choose a method: with or without an optimality criterion?

Distance methods

Choice one: how to estimate pairwise distances?

1) Estimate all pairwise distances

Choose a method (100s) - these are explicit model for sequence evolution and can include correction for:

- Different substitution rates (rate matrices)**
- Variation in amino acid composition (-F option)**
- Rate heterogeneity between sites (pinv, gamma)**

Calculation of pairwise distances

```

123456789.....33
Taxa a: AgggCTggTTCGgAGTCgTTAAg-ggAT--AAA
Taxa b: AAgg-TggCTCTgAATTgTTCgg-gCTT-CgAA
Taxa b: AAggCTgACTTTgAATTgTTCAgCgCTTACgAg
Taxa b: AAgg-TTgCTCTgAACTgTTCggCgCTTACgAA

      *  *   *  *  *  *   **   **   **
Taxa i: AGGGCTGGTTCGGAGTCGTTAAG-GGAT--AAA
Taxa j: AAGG-TGGCTCTGAATTGTTCGG-GCTT-CGAA
    
```

Number of aligned positions : **n = 33**

Total number of differences: **n_d = 12**

Observed

$$D_{ij} = \frac{n_d}{n} = 0.3636\dots$$

Corrections - model dependent

$$d_{ij} = F(D_{ij})$$

Distance methods

Choice two: how to estimate a “best” tree

2) Estimate a tree from the distance matrix

- Single tree algorithms, produce a unique tree e.g. the Neighbor joining method, a heuristic method to produce a ME tree. No effort is made to compare alternative trees. Efficient if distances are additive**
- A method with an optimality criterion, e.g. Minimum evolution and Least-Squares (Fitch-Margoliash) methods**

Phylogenetic trees from protein alignments

- **Parsimony based methods - unweighted/weighted**
- **Distance methods - model for distance estimation**
 - **probability of amino acid changes, site rate heterogeneity**
- **Maximum likelihood - model for ML calculations**
 - **probability of amino acid changes, site rate heterogeneity**

Protein phylogenies: some software

- **PAUP4: parsimony only!**
- **PHYLIP3.6: distance and parsimony
PROTPARS, PROTDIST & PROML**
- **MOLPHY2.3: maximum likelihood
PROTML**
- **TREE-PUZZLE5.0: maximum
likelihood (quartet) and distance**
- **MrBayes1.11: Bayesian analyses (ML)**

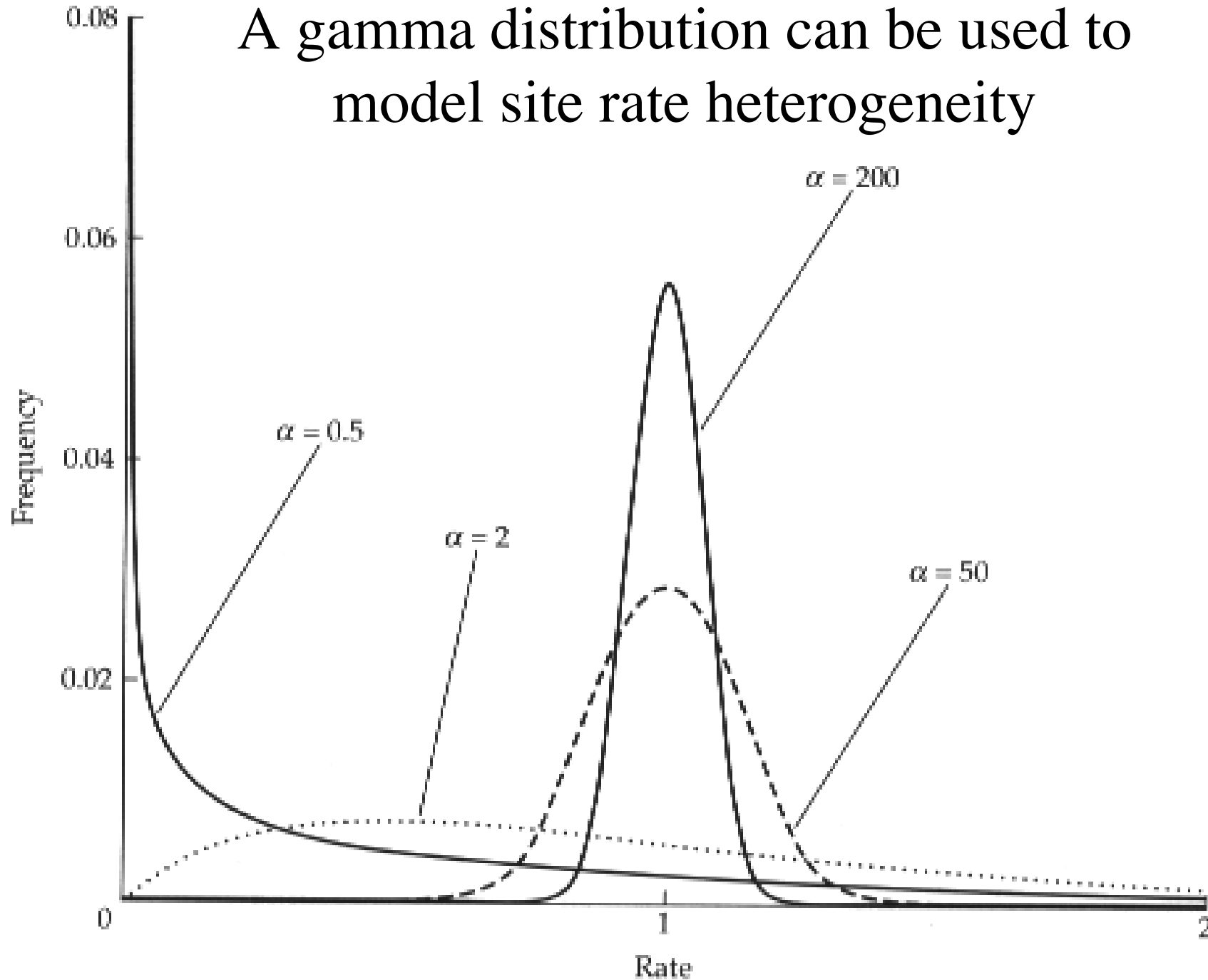
PHYLIP3.6

- **Protpars: parsimony**
- **Protdist: models for distance calculations:**
 - **PAM1, Kimura formula (PAM like), others...**
 - **Correction for rate heterogeneity between sites! Removal of invariant sites?**
- **NJ and LS distance trees**
- **Proml: protein ML analysis**
- **Bootstrapping**

TREE-PUZZLE5.0

- **Maximum likelihood method - “quartet puzzling” - with various protein rate matrices and can include correction for rate heterogeneity between sites - pinv + gamma shape**
- **ML pairwise distance estimates with complex models (as above) - puzzleboot**

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 tree**
- 2) Calculate the ML tree for all possible quartets**
- 3) Combine quartets in a n-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 - puzzle support values**

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)
 - Discrete gamma model for rate heterogeneity between sites -> output gives the rate category for each site
- **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 - long branch attraction?

PROTML2.3

- **Maximum likelihood analyses of protein alignments with several models (e.g. JTT, PAM, mtREV24)**
- **Has several tree search methods**
 - **Exhaustive search, small datasets only (-e)**
 - **Heuristic searches (-q, -s, -R). NNI search - needs a starting tree - gives also local bootstrap proportions (caution: LBP are typically overestimates!!!)**
 - **Semi-constrained analyses, performed with -e searches**

PROTML2.3

- **Advantages:**
 - **Perform tree search with all taxa (not quartets)**
 - **Several tree search strategies (semi-constrained, heuristic, NNI, exhaustive search)**
 - **Several protein rate matrices (PAM, JTT, mtREV) with or without corrections dataset specific amino acid frequencies**

PROTML2.3

- **Disadvantages:**
 - **No models for correction for rate heterogeneity between sites - correction is possible by manual removal of invariant sites (estimated with TREE-PUZZLE5.0)**
 - **No implemented “real” bootstrapping (but ask James for PHYCON)**
 - **LBP are typically overestimates of BP**
 - **No “sophisticated” heuristic search approach (such as TBR in PAUP, but it has NNI)**

MrBayes 1.11

- **Complex models for amino acid changes:**
 - Poisson, ..., PAM and JTT, (with correction for amino acid frequencies)
 - Correction for rate heterogeneity between sites (pinv, discrete gamma, site specific rates)
- **Powerful parameter space search (shape parameter and tree space combined!) using a Bayesian analysis based approach**
- **Posterior probabilities for a given clade are often close to one even when bootstrap values are low - overestimates?**

Summary

- **No single program allows thorough phylogenetic analyses of protein alignments**
- **Combination of PHYLIP3.6, TREE-PUZZLE5.0, PROTML2.3 and MrBayes1.11 allows detailed protein phylogenetics**
- **Remember that experimenting with your data and available methods/models is very important!!!**

The five steps in phylogenetics dancing

