

Comparative Genomics

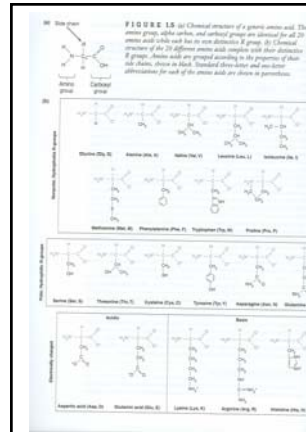
Lecture 8:
Phylogenetics III

Topics

- Stochastic models of protein evolution
- Rate variation
- Bayesian model selection
- Clock trees and non-clock trees
- Models of codon evolution and positive selection
- Mapping characters onto phylogenies
- Classification

Stochastic models of protein evolution

The 20 Amino Acids



- Four groups:
1. Hydrophobic neutral
 2. Hydrophilic neutral
 3. Acidic (- charge)
 4. Basic (+ charge)

Protein models 1

$$Q = \begin{pmatrix} & [A] & [C] & [D] & \dots & [Y] \\ [A] & - & \mu & \mu & \dots & \mu \\ [C] & \mu & - & \mu & \dots & \mu \\ [D] & \mu & \mu & - & \dots & \mu \\ \dots & \dots & \dots & \dots & - & \dots \\ [Y] & \mu & \mu & \mu & \dots & - \end{pmatrix}$$

The Poisson model
Essentially a Jukes-Cantor model

Protein models 2

$$Q = \begin{pmatrix} & [A] & [C] & [D] & \dots & [Y] \\ [A] & - & \pi_C \mu & \pi_D \mu & \dots & \pi_Y \mu \\ [C] & \pi_A \mu & - & \pi_D \mu & \dots & \pi_Y \mu \\ [D] & \pi_A \mu & \pi_C \mu & - & \dots & \pi_Y \mu \\ \dots & \dots & \dots & \dots & - & \dots \\ [Y] & \pi_A \mu & \pi_C \mu & \pi_D \mu & \dots & - \end{pmatrix}$$

The Equalin model
A generalized Felsenstein 1981 model

Protein models 3

$$Q = \begin{pmatrix} & [A] & [C] & [D] & \dots & [Y] \\ [A] & - & r_{AC}\pi_C\mu & r_{AD}\pi_D\mu & \dots & r_{AY}\pi_Y\mu \\ [C] & r_{AC}\pi_A\mu & - & r_{CD}\pi_D\mu & \dots & r_{CY}\pi_Y\mu \\ [D] & r_{AD}\pi_A\mu & r_{CD}\pi_C\mu & - & \dots & r_{DY}\pi_Y\mu \\ \dots & \dots & \dots & \dots & - & \dots \\ [Y] & r_{AY}\pi_A\mu & r_{AC}\pi_C\mu & r_{AD}\pi_D\mu & \dots & - \end{pmatrix}$$

The GTR model
208 free parameters
(189 rates, 19 state freqs)
Too many parameters for most datasets

Protein models 4

$$Q = \begin{pmatrix} & [A] & [C] & [D] & \dots & [Y] \\ [A] & - & r_{AC}\pi_C\mu & r_{AD}\pi_D\mu & \dots & r_{AY}\pi_Y\mu \\ [C] & r_{AC}\pi_A\mu & - & r_{CD}\pi_D\mu & \dots & r_{CY}\pi_Y\mu \\ [D] & r_{AD}\pi_A\mu & r_{CD}\pi_C\mu & - & \dots & r_{DY}\pi_Y\mu \\ \dots & \dots & \dots & \dots & - & \dots \\ [Y] & r_{AY}\pi_A\mu & r_{AC}\pi_C\mu & r_{AD}\pi_D\mu & \dots & - \end{pmatrix}$$

$$r_{AC} = 21, r_{AD} = 55, \dots, r_{YY} = 35$$

$$\pi_A = 0.032, \pi_C = 0.019, \dots, \pi_Y = 0.064$$

A fixed rate model
(Jones, Dayhoff, ...)

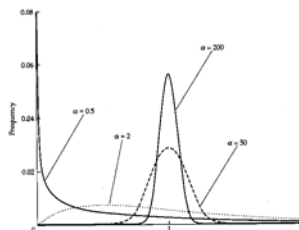
Fixed rate models

- Jones: General
- Dayhoff: General
- Mtrev: Mitochondrial proteins
- Mtmam: Mammal mitochondrial proteins
- Wag: General
- Rtrev: Reverse transcriptase, retroviruses
- Cprev: Chloroplast proteins
- Blosum: General
- Vt: Vertebrate proteins

(Do "Citations" in MrBayes to find out more)

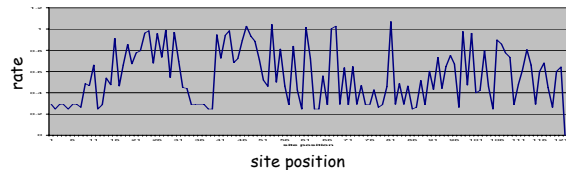
Rate variation across sites

Rate Variation Across Sites



Gamma distribution
The shape of the distribution is determined by a single parameter, the shape parameter α

Rate variation across sites in a protein-coding gene (first positions in replicase; based on nine bacteriophages)



Spatial autocorrelation is effected by codon position in protein-coding genes

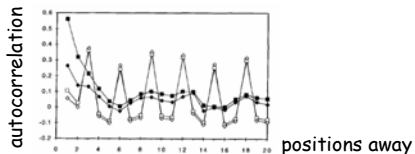


FIGURE 5.—Serial correlation of substitution rates along the mtDNA sequence, which are predicted by assuming the HKY+C+AdG (■), HKY+C+dG (●), HKY+AdG (□) and HKY+dG (○) models; $K = 8$ categories are used in these discrete-gamma models. The tree topology of Figure 4 is assumed. The graph shows the correlation coefficient between predicted rates (r) at sites separated by 1, 2, ..., 20 nucleotides. For the HKY+C+AdG and HKY+C+dG models, which assume different rate parameters for codon positions, only the random variable (r) from the (discrete) gamma distribution is used in the calculations.

(from Yang 1995)

Bayesian model testing

Bayesian Model Testing

- The normalizing constant in Bayes' theorem, the marginal probability of the model or $f(X)$, can be used for model testing
- $f(x)$ can be estimated by taking the harmonic mean of the likelihood values from the MCMC run (MrBayes will do this automatically with 'sump')
- Critical values in Kass and Raftery (1997)
- Any models can be compared: nested, non-nested, data-derived
- With Bayes factor comparisons, you do not need to decide first on the prior probability of the models (implicitly equal probability)

Bayes' theorem

$$f(\theta | X) = \frac{f(\theta)f(X | \theta)}{\int f(\theta)f(X | \theta)d\theta} = \frac{f(\theta)f(X | \theta)}{f(X)}$$

Model likelihood

Bayesian Model Testing

Posterior model odds:

$$\frac{f(M_1)f(X/M_1)}{f(M_0)f(X/M_0)}$$

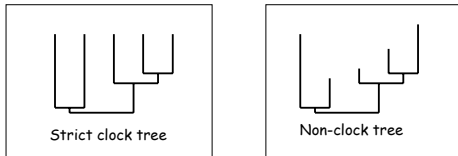
Bayes Factor:

$$B_{10} = \frac{f(X/M_1)}{f(X/M_0)}$$

Bayes Factor Comparisons

$2\ln B_{10}$	B_{10}	Interpretation
0 to 2	1 to 3	Barely worth a mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

Standard models of branch lengths



- Clock trees allow dating; non-clock trees do not
- But clock trees poor fit to most data

Codon models

Codon models

- If the change involves more than one nucleotide substitution, the rate is 0
- If the change involves one nucleotide substitution, the rate is equivalent to that nucleotide substitution rate
- If the change is non-synonymous, the rate is a factor ω of the base rate
- $\omega > 1$ \rightarrow positive selection
- $\omega < 1$ \rightarrow negative selection
- We can let ω vary over sites and infer the evolutionary pressure at each site

The Universal Code

		Second position							
		U		C		A		G	
First position	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
		UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
		UUA	Leu	UCA	Ser	UAA	End	UGA	End
		UUG	Leu	UCG	Ser	UAG	End	UGG	Trp
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
		CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
		CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
		CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
		AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
		AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
		AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
		GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
		GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
		GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Codon models

$$Q = \begin{pmatrix} & \text{[AAA]} & \text{[AAC]} & \text{[AAG]} & \dots & \text{[TTT]} \\ \text{[AAA]} & - & \omega\pi_{AAC}\mu & \pi_{AAG}\mu & \dots & 0 \\ \text{[AAC]} & \omega\pi_{AAA}\mu & - & \omega\pi_{AAG}\mu & \dots & 0 \\ \text{[AAG]} & \pi_{AAA}\mu & \omega\pi_{AAC}\mu & - & \dots & 0 \\ \dots & \dots & \dots & \dots & - & \dots \\ \text{[TTT]} & 0 & 0 & 0 & \dots & - \end{pmatrix}$$

Goldman-Yang/Muse-Gaut model
60+1 parameters

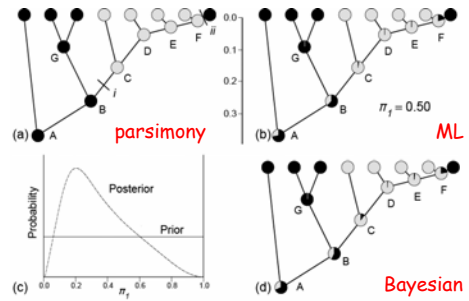


Protein structure of the influenza hemagglutinin protein, chains A and B. The seven positively selected residues shown in red. They were identified by simulation from the posterior distribution of a Bayesian MCMC analysis.

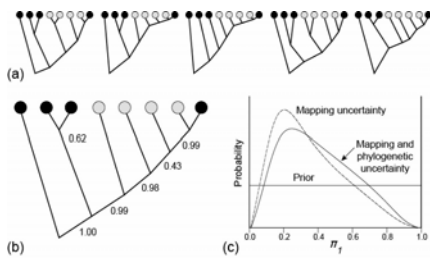
Huelsenbeck et al. Science 294:2310, 2001

Mapping characters onto phylogenies

Mapping Uncertainty

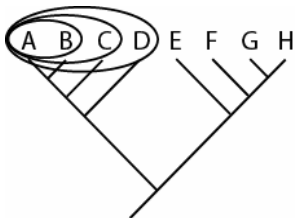


Phylogenetic and Mapping Uncertainty



Phylogenetic classification

Phylogenetic Classification



Only monophyletic groups (clades, natural groups) should be recognized in a biological classification

Monophyletic groups include AB, ABC, ABCD, GH, FGH, and EFGH

Examples of non-monophyletic groups include: AC, EF, ECD, and AG. These should not be recognized as groups in a biological classification.

Can we detect molecular adaptations in diving mammals?

Can we find an appropriate model organism for studying disease X?

Where and how did disease X originate?

How can we find suitable targets for design of antiviral drugs?

How often does mutation X, which causes disease Y, originate?

Why is species X similar to the unrelated species Y?

Is the current classification of organism group X correct?

Are there genes that evolve in a clock-like fashion such that we can use a molecular clock to date past events?

Which gene is best for identification of strains of virus X?

Can we use phylogenies to identify a virulence factor in a group of disease-causing bacteria?

Is there a correlation between the structure of gene X, implied in aging processes, and longevity?

Can we use the comparative approach to find a heat stable version of an enzyme involved in the production of vitamin C?