Max. likelihood & Bayesian techniques are both likelihood-based.

Weaknesses of likelihood for phylogeny reconstruction:
1) Computational tractability
2) Based on overly simplistic evolutionary models.

But,

a) All phylogeny reconstruction methods are based on assumptions but some (e.g. parsimony) are not based on explicit ones. For methods based on unstated assumptions, we need to worry not just whether the assumptions are realistic but also we need to worry about what they are.

b) Likelihood methods allow assumptions to be rigorously tested. When an assumption is found to be particularly poor, it can be replaced with a better one (i.e., models will improve over time!)

Max. likelihood & Bayesian techniques are both likelihood-based.

Weaknesses of likelihood for phylogeny reconstruction:
1) Computational tractability
2) Based on overly simplistic evolutionary models.

But,

a) All phylogeny reconstruction methods are based on assumptions but some (e.g. parsimony) are not based on explicit ones. For methods based on unstated assumptions, we need to worry not just whether the assumptions are realistic but also we need to worry about what they are.

b) Likelihood methods allow assumptions to be rigorously tested. When an assumption is found to be particularly poor, it can be replaced with a better one (i.e., models will improve over time!)

Strengths of likelihood methods:

1. Explicit Assumptions – we know what we're assuming.
2. Use **all** information in a data set. Distance methods, for example, do not. This is part of the explanation for success of likelihood methods in simulations – they tend to yield estimates that are closer to the truth than other methods.
3. Likelihood approaches are consistent. Estimates get better as amount of data increases. (Caveat: violation of model assumptions may cause loss of consistency property)
4. Because likelihood applied to so many statistical situations in addition to phylogenetics, powerful theory & tools for performing likelihood analyses have developed. This theory and these tools (e.g., tools for hypothesis testing) can be applied to phylogenetics.
5. Likelihood lets you know how good estimate is, in addition to what estimate is.

Mechanistic versus Phenomenological Models of Sequence Evolution

see Ph.D. thesis by Nicolas Rodrigue ("Phylogenetic structural modeling of molecular evolution", 2008, University of Montreal)

(see also Rodrigue & Philippe. 2010. Trends in Genetics 26:248-252)

One good idea for more realistic models ...

TUFFLEY, C., and M. A. STEEL. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147:63–91.

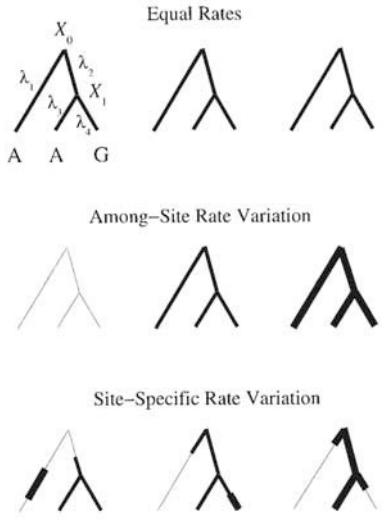


FIG. 1.—Distribution of rates across sites and lineages under three models of evolution. Each tree plot describes the distribution of rates across lineages for a particular site under the considered model. Three categories of rate are assumed, represented by different line thicknesses. Under the equal-rates (ER) model, all sites evolve at a constant, unique, moderate rate. Under the among-site rate variation (ASRV) model, each site has its own rate (low, moderate, or high), which is constant between lineages. Under the site-specific rate variation (SSRV) model, the rate of a site can switch between categories; a site has distinct rates in distinct lineages.

From Galtier. 2001. Mol. Biol. Evol. 18(5):866-873.

Tuffley/Steel -type model

		Slow				*Fast*			
		A	C	G	T	A	C	G	T
Slow	A	-	r	r	r	f	0	0	0
	C	r	-	r	r	0	f	0	0
	G	r	r	-	r	0	0	f	0
	T	r	r	r	-	0	0	0	f
Fast	A	s	0	0	0	-	q	q	q
	C	0	s	0	0	q	-	q	q
	G	0	0	s	0	q	q	-	q
	T	0	0	0	s	q	q	q	-

Substitution Rates: q>r
Switching rates: f (slow to fast), s (fast to slow)

Dayhoff model of protein evolution (see Dayhoff et al. 1972; Dayhoff et al. 1978) operates at the level of the 20 amino acid types.

π_i is the probability of amino acid type i

α_{ij} is the instantaneous rate of replacement from amino acid i to amino acid j

Dayhoff model is most general time-reversible 20-state model of amino acid replacement.

This means $\pi_i \alpha_{ij} = \pi_j \alpha_{ji}$ for all i and j.

It is important to separate the Dayhoff model of protein evolution from:

1. The procedure used by Dayhoff and collaborators to estimate the α_{ij} AND

2. The data set upon which the α_{ij} estimates were based.

Dayhoff and collaborators exploited the fact that the probability of replacements from amino acid type i to type j (i not equal to j) is approximately linear in time for small amounts of time.

In other words, the probability of a replacement from amino acid type i to a different type j is approximately $\alpha_{ij}t$ if t represents some small amount of time.

Subsequent studies (e.g., Jones et al. 1992) adopted the Dayhoff model but employed different data sets and parameter estimation procedures.

A	Ala																		
R	Arg	30																	
N	Asn	109	17																
D	Asp	154	0	532															
C	Cys	33	10	0	0														
Q	Gln	93	120	50	76	0													
E	Glu	266	0	94	831	0	422												
G	Gly	579	10	156	162	10	30	112											
H	His	21	103	226	43	10	243	23	10										
I	Ile	66	30	36	13	17	8	35	0	3									
L	Leu	95	17	37	0	0	75	15	17	40	253								
K	Lys	57	477	322	85	0	147	104	60	23	43	39							
M	Met	29	17	0	0	0	20	7	7	0	57	207	90						
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17					
P	Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7				
S	Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269			
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696		
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0	
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

Figure 80. Numbers of accepted point mutations (X 10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	2	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.

C	Cys	12																				
S	Ser	0	2																			
T	Thr	-2	1	3																		
P	Pro	-3	1	0	6																	
A	Ala	-2	1	1	1	2																
G	Gly	-3	1	0	-1	1	5															
N	Asn	-4	1	0	-1	0	0	2														
D	Asp	-5	0	0	-1	0	1	2	4													
E	Glu	-5	0	0	-1	0	0	1	3	4												
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4											
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6										
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp

Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of -10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair would be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data.

Table 23

Correspondence between Observed Differences and the Evolutionary Distance

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

Table 21

Relative Mutabilities of the Amino Acids[a]

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

[a]The value for Ala has been arbitrarily set at 100.

Table 22

Normalized Frequencies of the Amino Acids in the Accepted Point Mutation Data

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

Inspired by Lartillot and Philippe's CAT model of amino acid replacement that permits variation of preferred residues among sites, there is active development of sequence evolution models that allow variation of evolutionary processes among sites without prespecifying the number of categories, the nature of categories, or which sites are in which categories.

Key Ingredient: "Dirichlet Process" as a prior for the number of categories and for the probabilities of the categories.

Nicolas Lartillot and Hervé Philippe. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Mol. Biol. Evol. 21(6):1095-1109. 2004

Dirichlet Process Priors ("Chinese restaurant process", not same as Dirichlet distribution):

Useful to specify prior distribution for situations when number of categories is unknown and where prior probability of each possible category needs determination.

Additional applications in Evolution Include:

Characterization of population structure
Huelsenbeck and Andolfatto. 2007. Genetics. 175:1787-1802.

Variation in nonsyn. and synonymous rates among sites
Huelsenbeck et al. 2006. PNAS 103(16): 6263-6268.

Variation in evolutionary rate across a phylogeny
Heath et al. 2012. Mol. Biol. Evol. 29(3): 939-955.

Codon Models: Evolution occurs at the DNA level rather than at the amino acid level.

It makes sense to frame a model of protein evolution in terms of codons rather than amino acid types (Schoniger et al. 1990; Goldman and Yang 1994; Muse and Gaut 1994).

Codon-based models are typically framed in terms of 61 codon-states rather than 64 codon-states because the common genetic codes have three stop codons, and the possibility that a stop codon may appear or disappear from a sequence is not allowed.

One simplification that is often adopted holds that changes from one codon to another are only possible when the two codons differ at exactly one of the three codon positions.

The instantaneous rates of other changes between codons are set to 0.

Typical parameterization of a codon model when physicochemical differences between amino acids are ignored...

Instantaneous rate $\alpha_{i,j}$ from codon i to codon j is set to 0 if i and j differ at more than one nucleotide or if j encodes a premature stop codon. For cases where i and j differ by exactly one nucleotide, rate matrix entries are:

$$\alpha_{i,j} = \begin{cases} u\pi_j & \text{for a synonymous transversion} \\ u\pi_j\kappa & \text{for a synonymous transition} \\ u\pi_j\omega & \text{for a nonsynonymous transversion} \\ u\pi_j\kappa\omega & \text{for a nonsynonymous transition} \end{cases}$$

u, π_j, and κ reflect mutation rates

$\omega > 1$ means positive **diversifying** selection (i.e., nonsyn. rates higher than they would be if changes were synonymous)

Other kinds of positive selection exist (e.g., positive directional selection)

The previous rate matrix can be modified so that each codon k has its own parameter ω_k. The rates then become:

$$\alpha_{i,j} = \begin{cases} u\pi_h & \text{for a synonymous transversion} \\ u\pi_j\kappa & \text{for a synonymous transition} \\ u\pi_j\omega_k & \text{for a nonsynonymous transversion} \\ u\pi_j\kappa\omega_k & \text{for a nonsynonymous transition} \end{cases}$$

As with the rate heterogeneity among sites treatment, the distribution of ω_k values among codons can be modelled. Often, we want to know if certain codons have ω_k values that exceed 1.

Alternatively, we can assume all codons share the same value of ω but that ω values vary among branches on the tree. The rate matrix then becomes:

$$\alpha_{i,j} = \begin{cases} u\pi_j & \text{for a synonymous transversion} \\ u\pi_j\kappa & \text{for a synonymous transition} \\ u\pi_j\omega_B & \text{for a nonsynonymous transversion} \\ u\pi_j\kappa\omega_B & \text{for a nonsynonymous transition} \end{cases}$$

where ω_B is the parameter value for branch B. Many other possibilities for parameterizing codon models exist. and codon models can become very elaborate.

For example, Pedersen and colleagues (1998) carefully designed a codon model to reflect the fact that CpG dinucleotide levels are depressed in lentiviral genes.

Codon models have received attention for their potential ability to detect positive selection (Nielsen and Yang 1998).

Early methods for detecting positive selection from protein-coding DNA sequence data were designed to looked for an "excess" of nonsynonymous amino acid replacements throughout the sequence.

Codon methods offer the potential of detecting positive selection at individual sites and for detecting the existence of a small proportion of sites at which positive selection may operate.

Best statistical technique for detecting positive selection is a contentious issue at the moment...

Some future directions for codon-based models ...

Evolutionary changes that simultaneously affect two consecutive positions could be allowed (Averof et al. 2000 have claimed empirical evidence for these kinds of changes).

Reconciliation of codon-based models with classical population genetic models – some progress has been made (see Halpern and Bruno 1998).

Improved treatment of effects of chemical similarity of amino acids on protein evolution

For change from Sequence i to Sequence j where i & j differ only at one sequence position, evolutionary rate from i to j is R_{ij} where

R_{ij} = (Mutation Rate) x (Fixation Probability)

(see Halpern & Bruno. 1998. MBE 15:910-917)

For change from Sequence i to Sequence j where i & j differ only at one sequence position, evolutionary rate from i to j is R_{ij} where

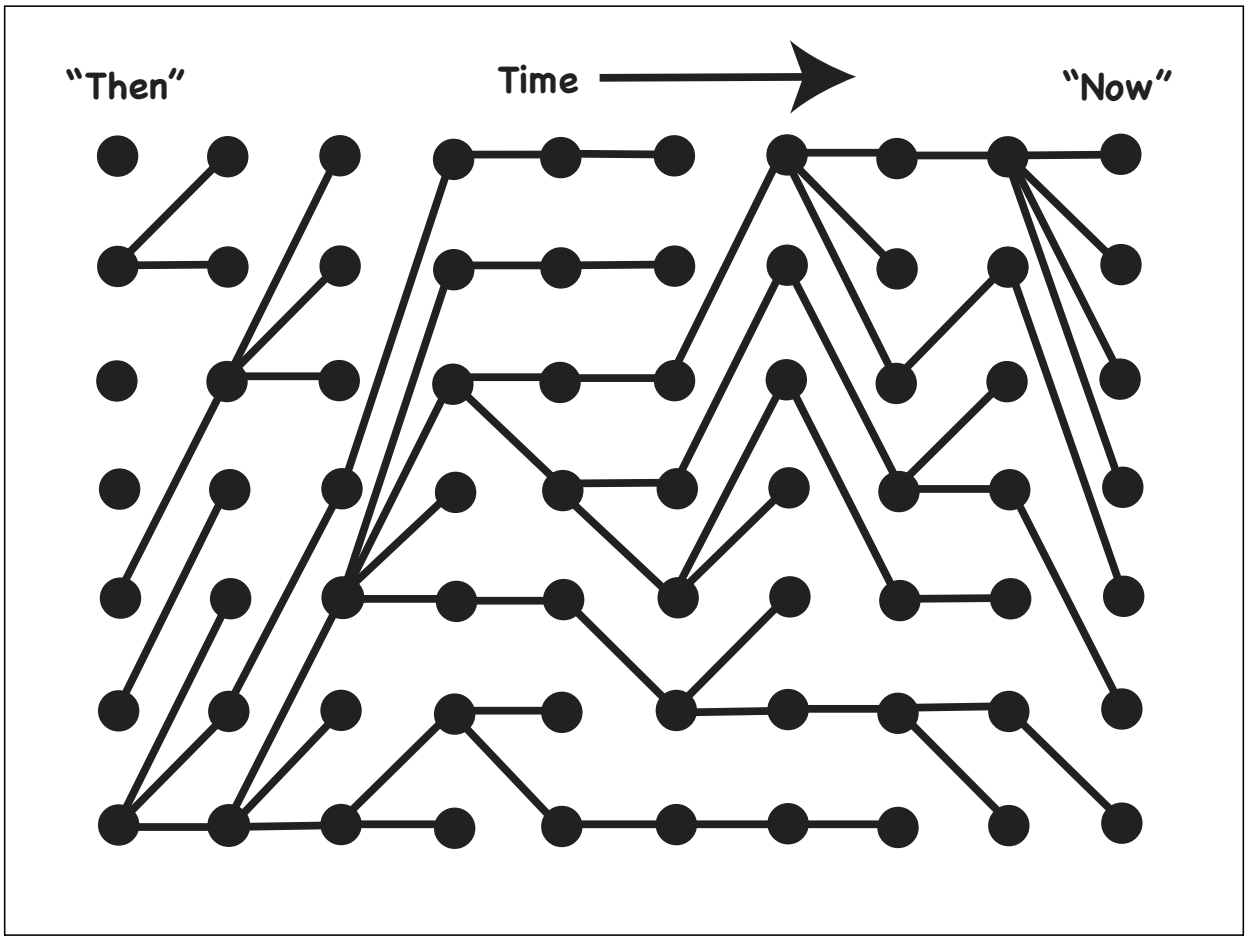
R_{ij} = (Mutation Rate) x (Fixation Probability)

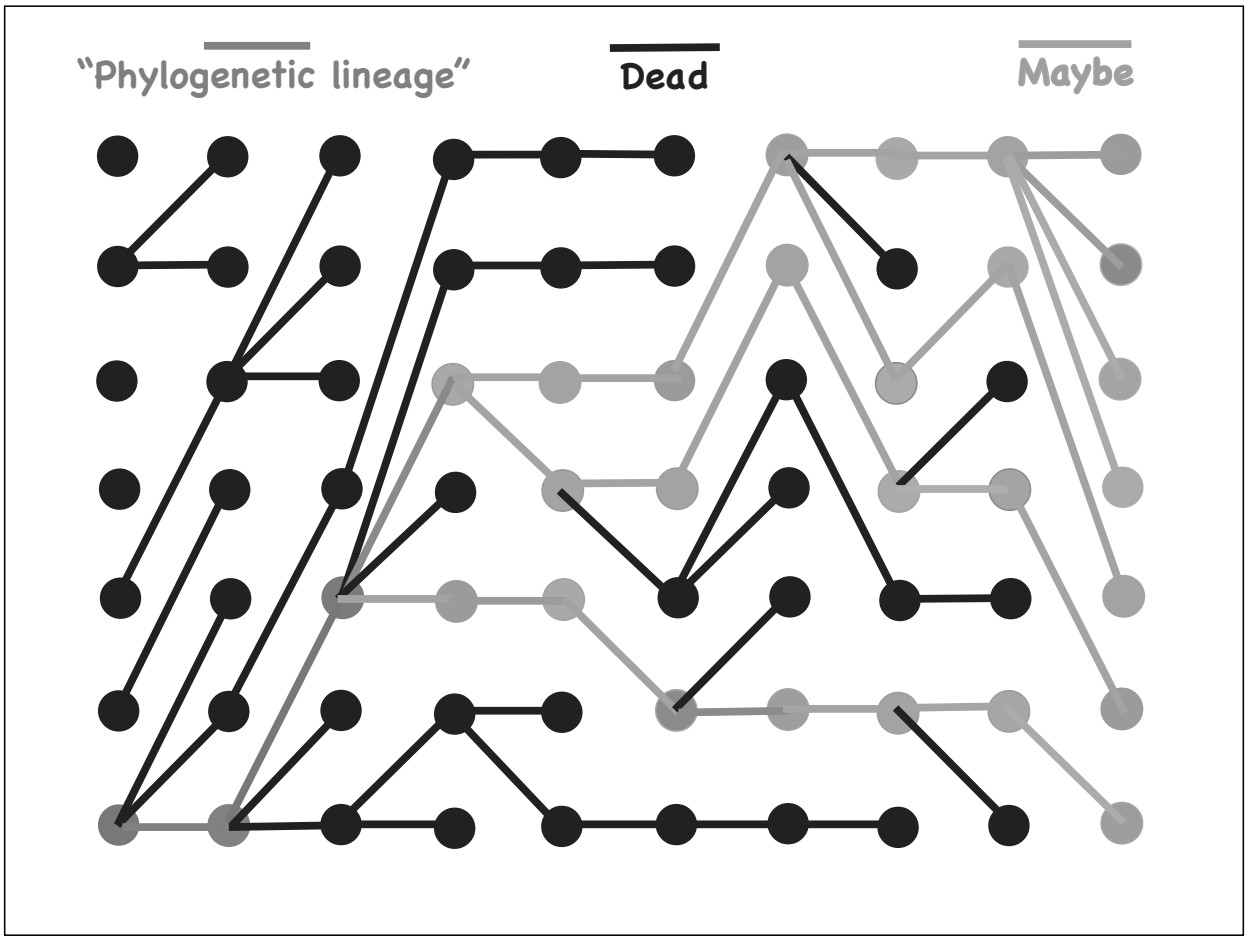
(see Halpern & Bruno. 1998. MBE 15:910-917)

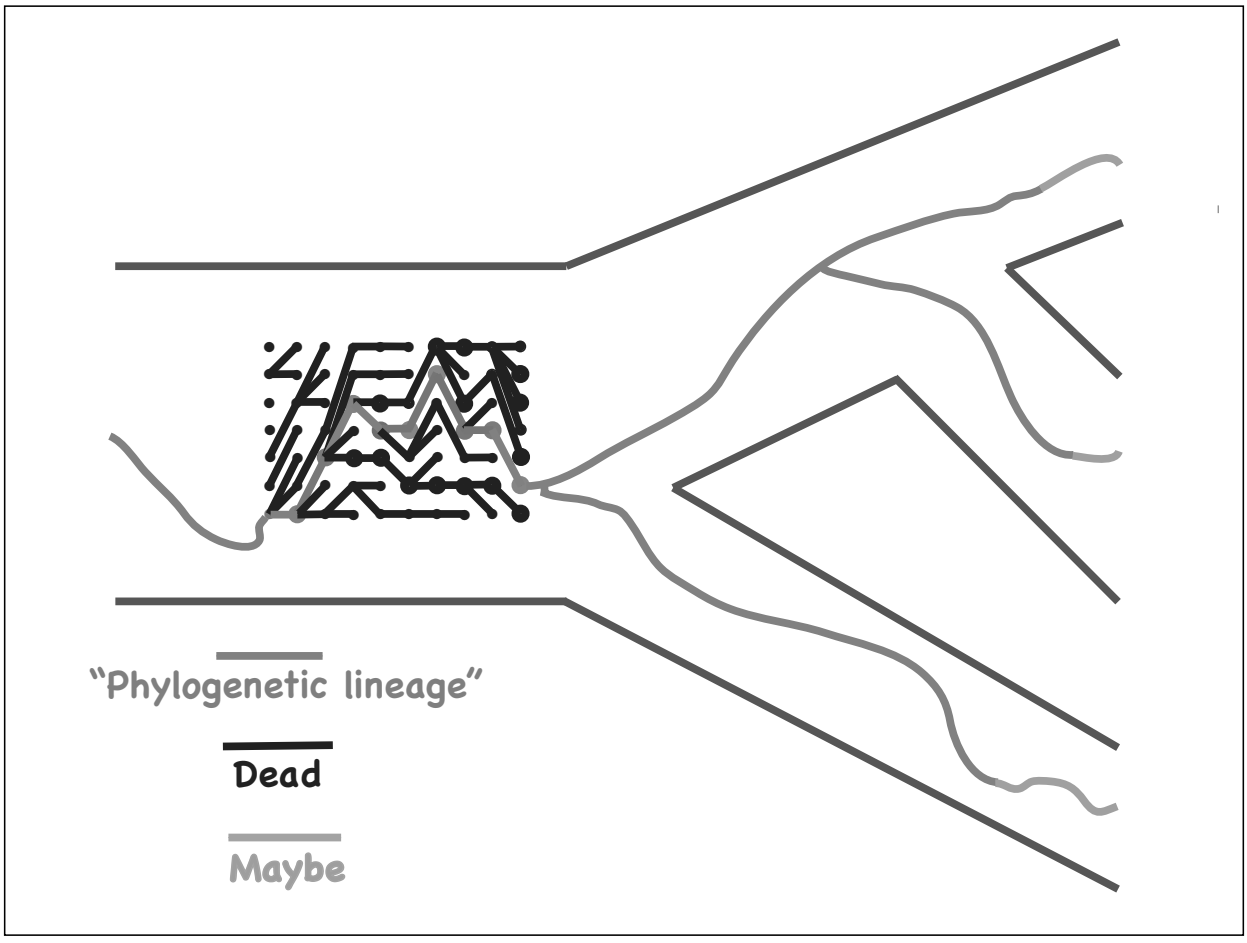
With low mutation rates, this depends on effective pop'n size "N" and relative fitness of j minus i (call this difference "s")

Population Genetic formulae for fixation probability allows estimation of Ns

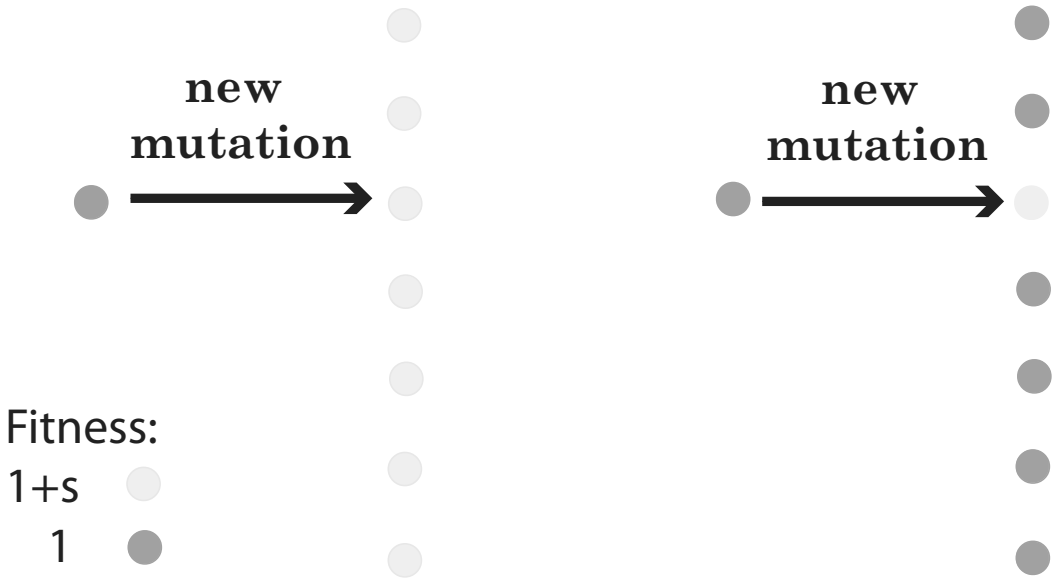
What justifies the assumption of phylogenetic models that sequences change over time according to a Markov process?




"Phylogenetic lineage"
Dead
Maybe


"Phylogenetic lineage"
Dead
Maybe

Fixation probabilities depend on the other alleles in the population



Towards more general dependence among sequence positions in molecular evolution...

Hwang, D.G., and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. U.S.A. 101(39):13994-14001

Jensen, J. L., and A. K. Pedersen. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. 32:499-517

Pedersen A. -M. K. and J. L. Jensen. 2001. A Dependent-Rates Model and an MCMC-Based Methodology for the Maximum-Likelihood Analysis of Sequences with Overlapping Reading Frames. Mol. Biol. Evol. 18(5):763-776.

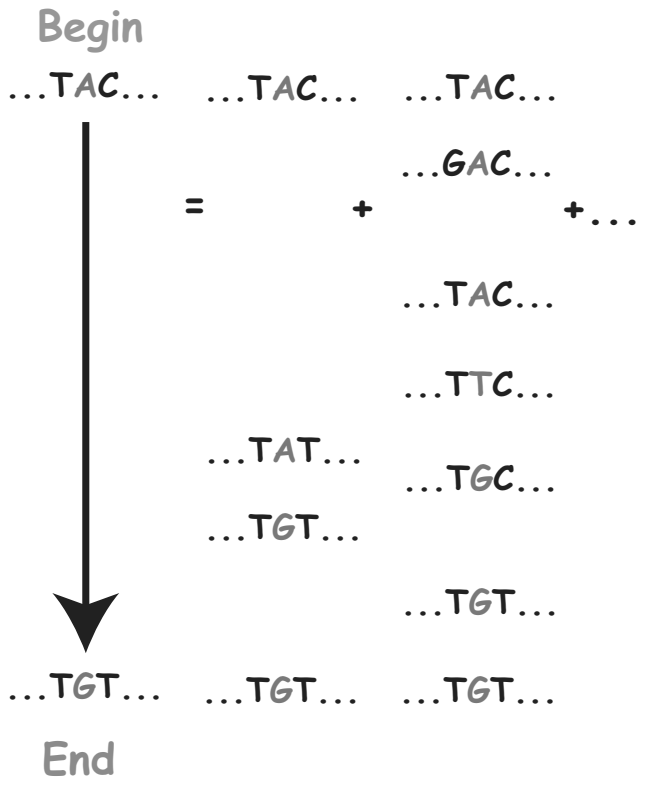
Robinson, D.M., D.T. Jones, H. Kishino, N. Goldman, and J.L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20(10): 1692-1704.

Siepel, A., and D. Haussler. 2004a. Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood. Mol. Biol. Evol. 21:468-488.

Siepel, A., and D. Haussler. 2004b. Combining phylogenetic and hidden Markov models in biosequence analysis. J Comput Biol. 11:413-428.

4-state substitution model

From \ To	A	C	G	T
A	–	+	+	+
C	+	–	+	+
G	+	+	–	+
T	+	+	+	–



$R_{i,j}$

From \ To	*AA...AA*	*AA...AC*	*AA...AG*	*AA...AT*	*AA...CA*	...	*TT...GT*	*TT...TA*	*TT...TC*	*TT...TG*	*TT...TT*
AA...AA	-	+	+	+	+		0	0	0	0	0
AA...AC	+	-	+	+	0		0	0	0	0	0
AA...AG	+	+	-	+	0		0	0	0	0	0
AA...AT	+	+	+	-	0		0	0	0	0	0
AA...CA	+	0	0	0	-		0	0	0	0	0
...											
TT...GT	0	0	0	0	0		-	0	0	0	+
TT...TA	0	0	0	0	0		0	-	+	+	+
TT...TC	0	0	0	0	0		0	+	-	+	+
TT...TG	0	0	0	0	0		0	+	+	-	+
TT...TT	0	0	0	0	0		+	+	+	+	-

4^N by 4^N rate matrix

Rate away from sequence i is

$$R_{i\bullet} = \sum_{j, j \neq i} R_{ij}$$

where R_{ij} is rate from sequence i to sequence j.

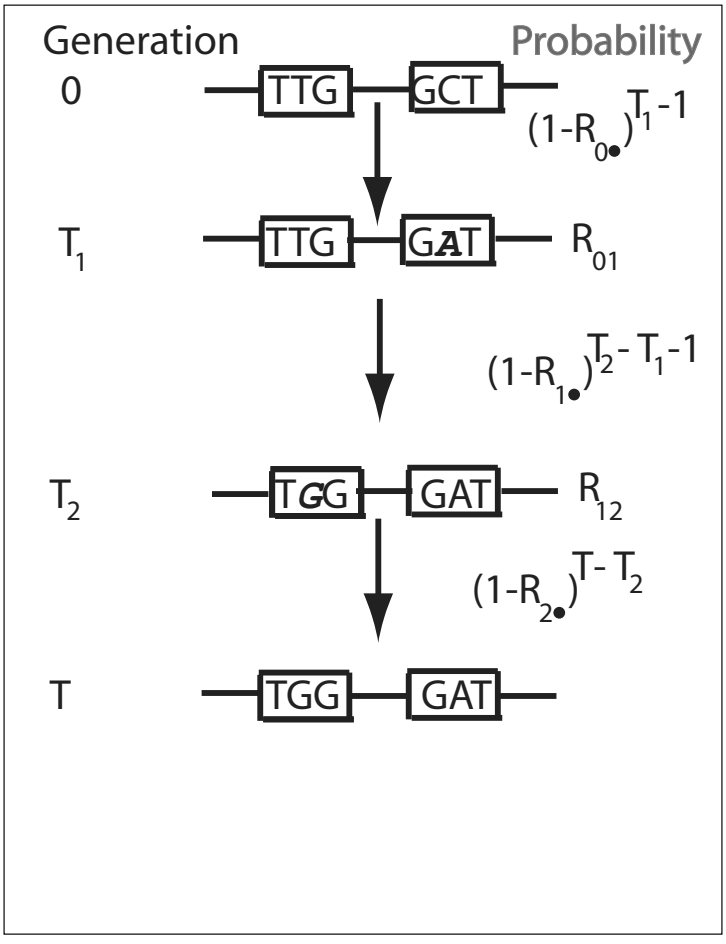
Consider T generations of evolution where

... Sequence 0 changes to Sequence 1 in generation T_1

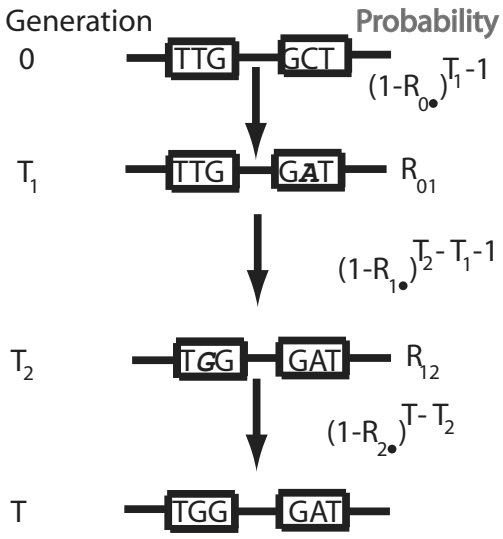
... Sequence 1 changes to Sequence 2 in generation T_2

... No other changes occur

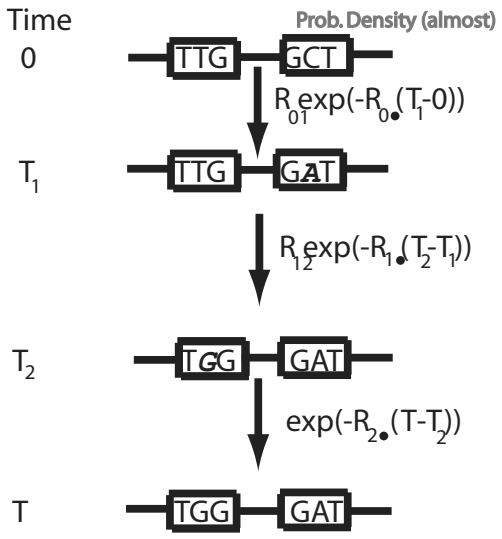
What is probability of this possible history?



Discrete Time



Continuous Time



(T represents many generations, rates per generation are small)

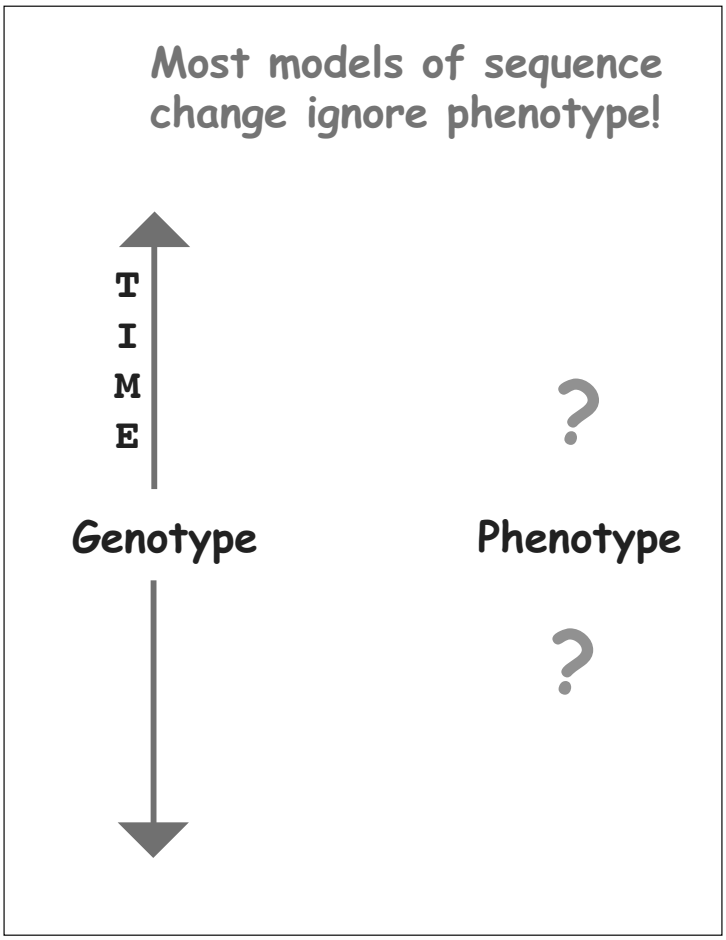
Data Augmentation:

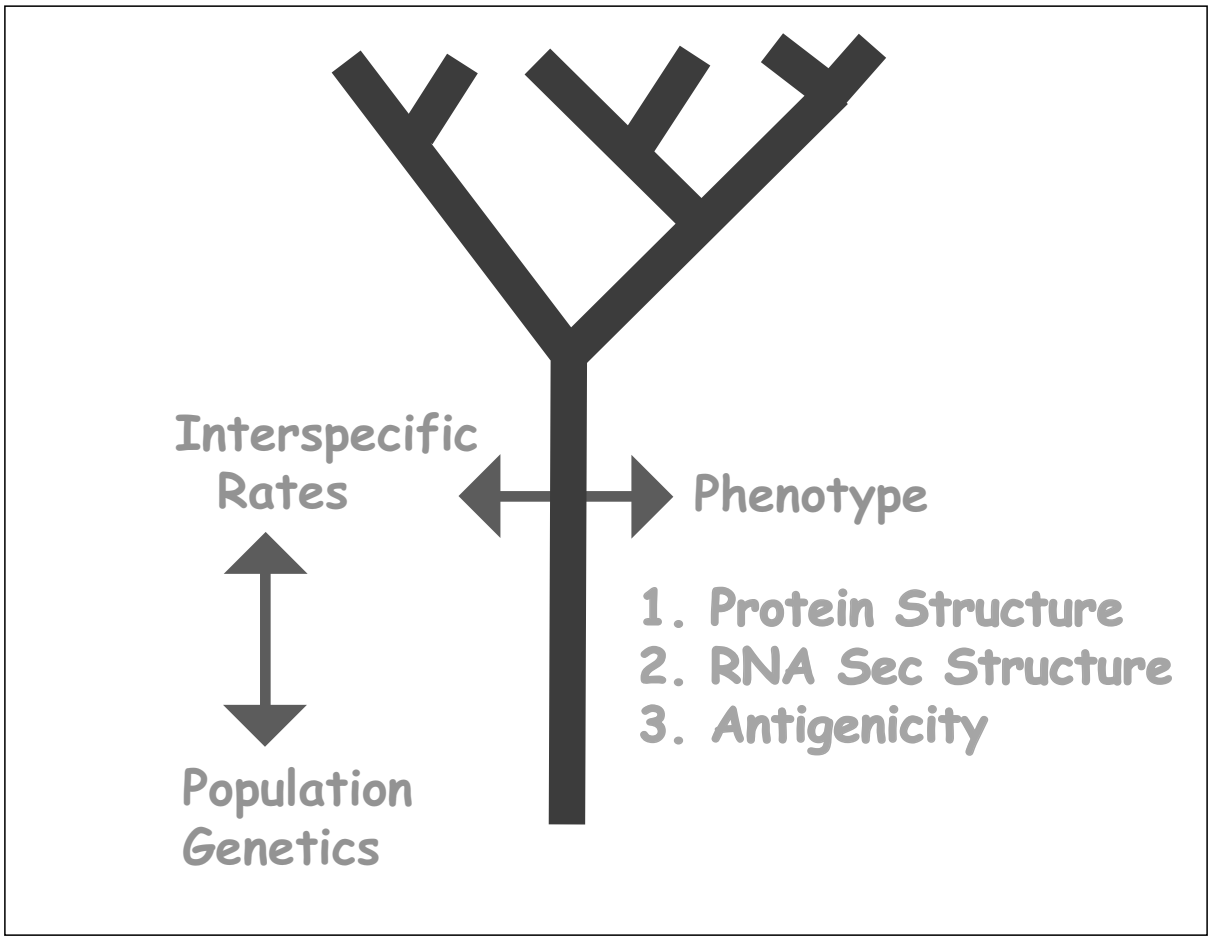
... an inference strategy for case where it is hard to calculate likelihood of observed data

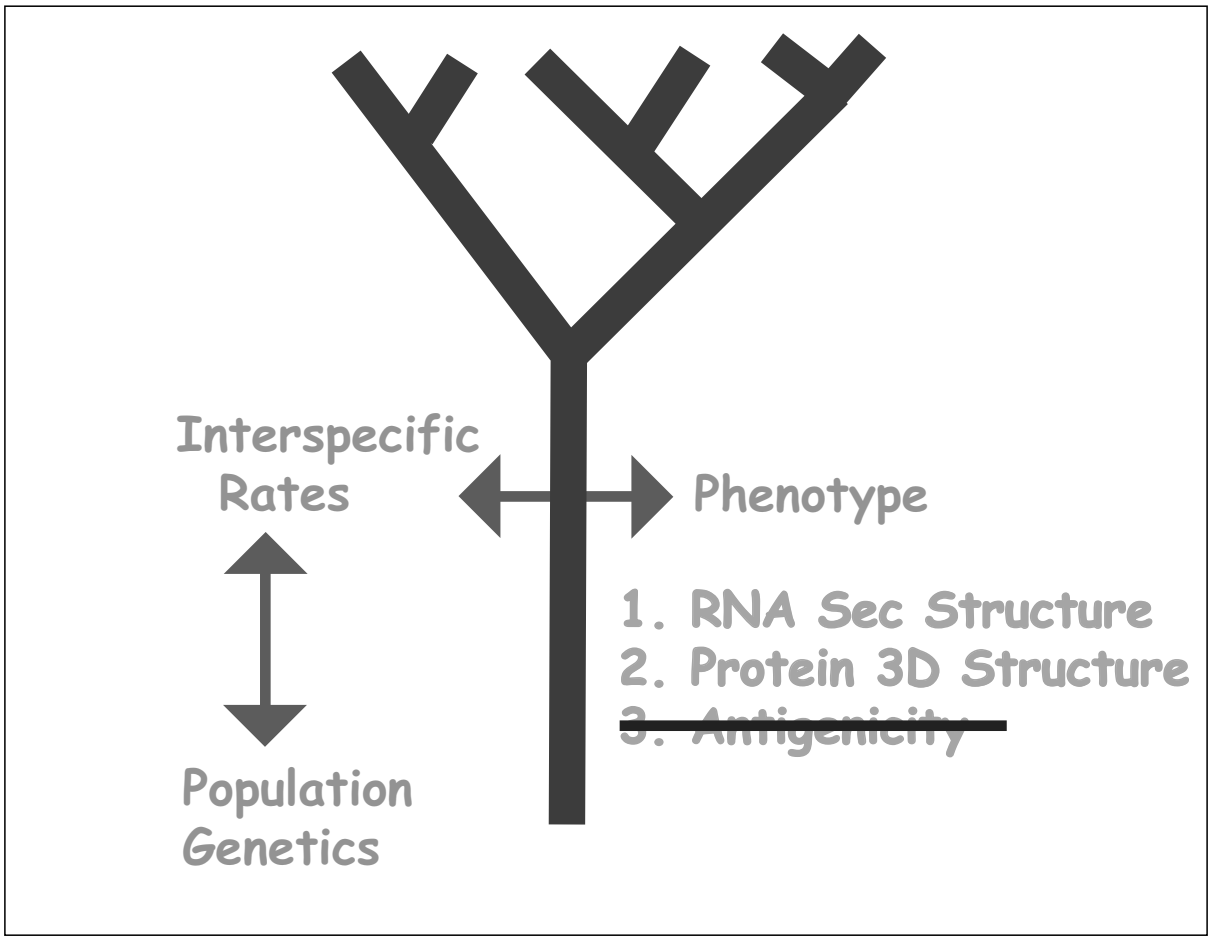
... Strategy is to facilitate likelihood computation by pretending that more is observed than actually is observed.

For example, might not be able to calculate likelihoods with models of sequence evolution but might be able to calculate likelihoods if entire evolutionary history was known.

Landis et al. ("Bayesian analysis of biogeography when the number of areas is large". Systematic Biology. 2013, Advance Access) employ this statistical strategy to infer history of ancestral ranges of species.


Most models of sequence change ignore phenotype!
TIME
?
Genotype
Phenotype
?


Interspecific Rates
Phenotype
1. Protein Structure
2. RNA Sec Structure
3. Antigenicity
Population Genetics



Biological Inspiration:

Parisi & Echave. 2001.
Mol. Biol. Evol. 18:750-756.

Statistical Inspiration:

Jensen & Pedersen. 2000.
Adv. Appl. Prob. 32:499-517

Pedersen & Jensen. 2001.
Mol. Biol. Evol. 18:763-776

Rate notation and assumptions

Rate R_{ij} from Sequence i to j is 0 if j has stop codon or if i and j differ at more than 1 position

Otherwise, assume i and j differ at 1 position where j has nucleotide type h

Model with independence among codons

$R_{ij} = \ldots$

$u\pi_h$ if synonymous transversion

$u\pi_h\kappa$ if synonymous transition

$u\pi_h\omega$ if nonsyn. transversion

$u\pi_h\kappa\omega$ if nonsyn. transition

$\omega > 1$ is positive selection

Protein structure changes far more slowly than protein sequence. There seem to be constraints on protein sequence evolution that maintain protein structure.

We assume tertiary structure known and unchanging

Fold recognition and sequence-structure compatibility

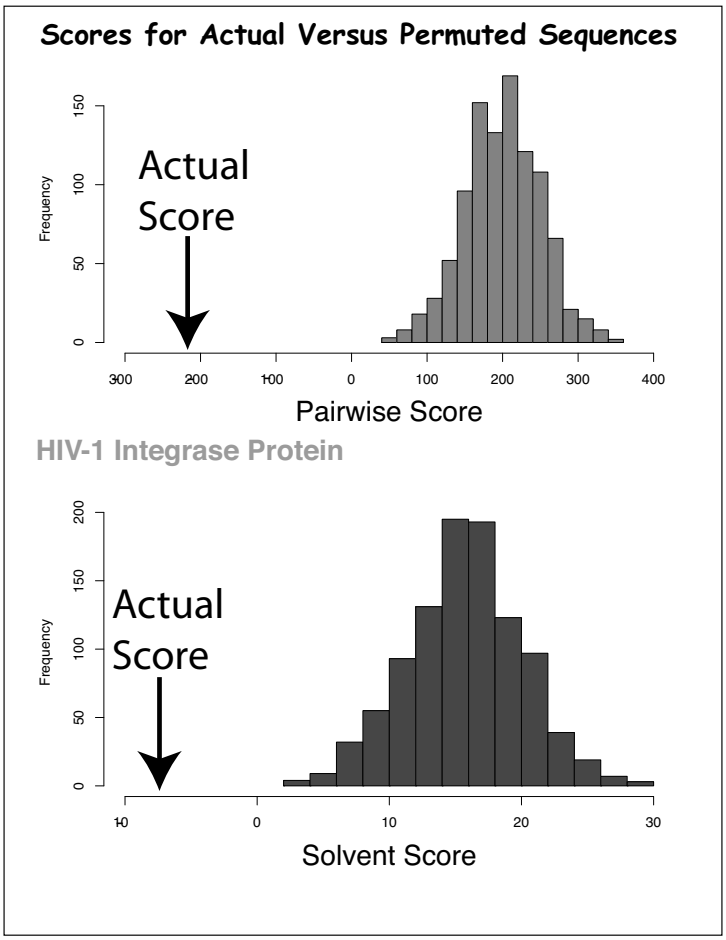
Idea underlying our model: Rate from sequence i to j should be low if j does not fold as well into known structure as i and high if j folds into known structure better than i

Sequence-structure compatibility assessed by GenThreader software of David Jones

$E_f(i)$ is solvent accessibility score of sequence i folded into known structure

$E_p(i)$ is pairwise interaction score of sequence i folded into known structure

(low scores fit better than high scores)

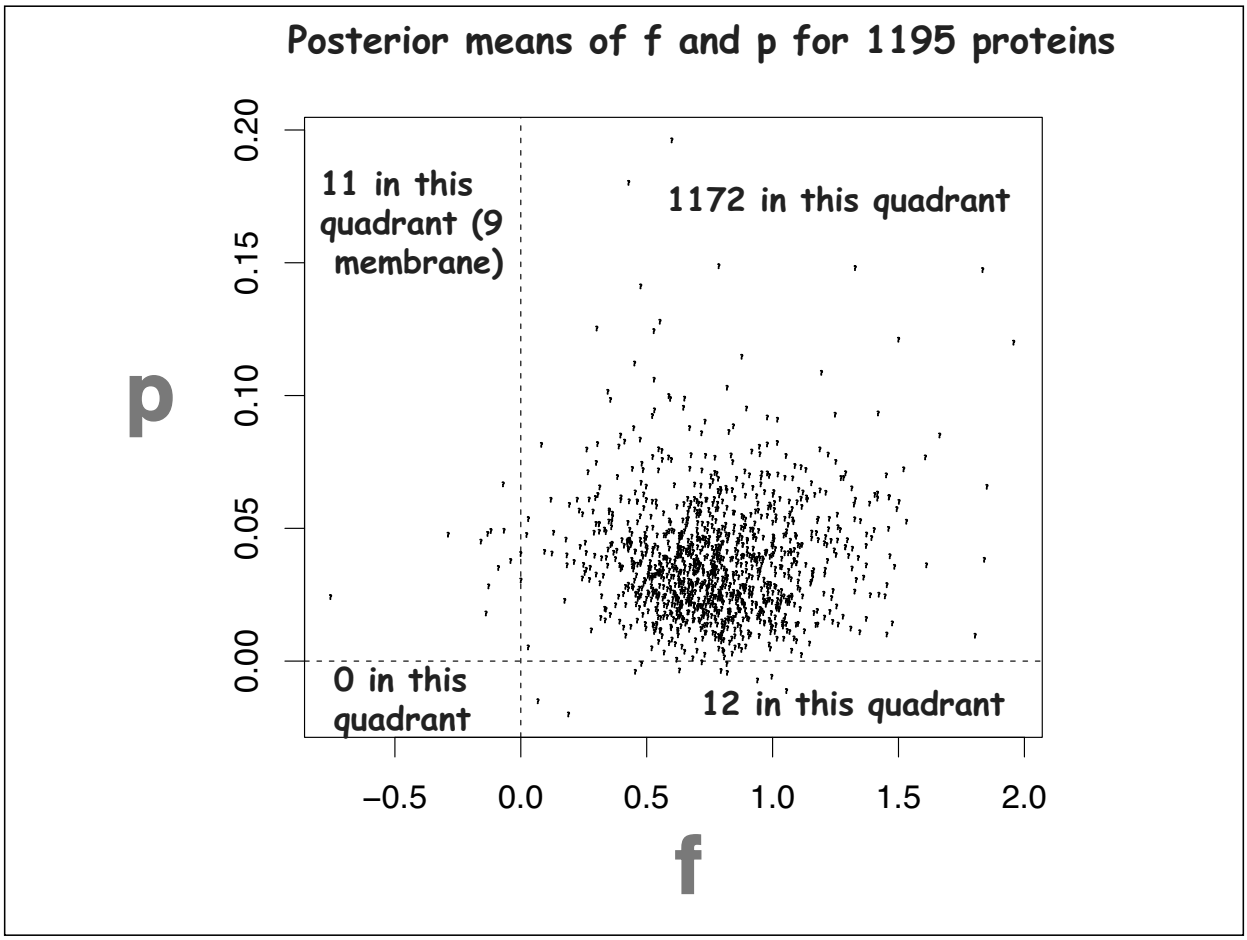


Protein tertiary structure as phenotype

$E_f(i)$ ($E_p(i)$) is solvent accessibility (pairwise) score of i

f & p relate scores to evolutionary rates

$$R_{i,j} = \begin{cases} u\pi_h & \text{syn. transversion} \\ u\pi_h\kappa & \text{syn. transition} \\ u\pi_h\omega e^{(E_f(i)-E_f(j))f+(E_p(i)-E_p(j))p} & \\ & \text{nonsyn. transv.} \\ u\pi_h\kappa\omega e^{(E_f(i)-E_f(j))f+(E_p(i)-E_p(j))p} & \\ & \text{nonsyn. transi.} \end{cases}$$



Why model dependence among codons due to protein structure?

1. Quantify impact of protein structure on protein evolution

2. Ancestral Sequence Reconstruction

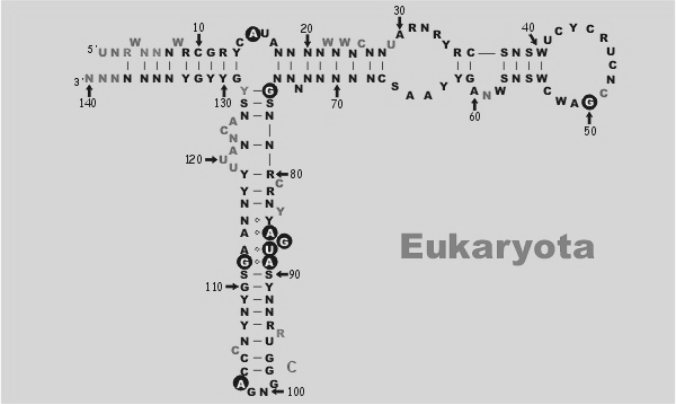
3. Detect positive selection

4. Infer order of selectively beneficial nucleotide substitutions

5. Predict evolution? (probably not)

5S rRNA secondary structure

(from http://rose.man.poznan.pl/5SData/)



red and green positions are insertions/deletions relative to most sequences

black circles with yellow letters are highly conserved throughout eukaryotes

(following results from Jiaye Yu)

RNA secondary structure as phenotype

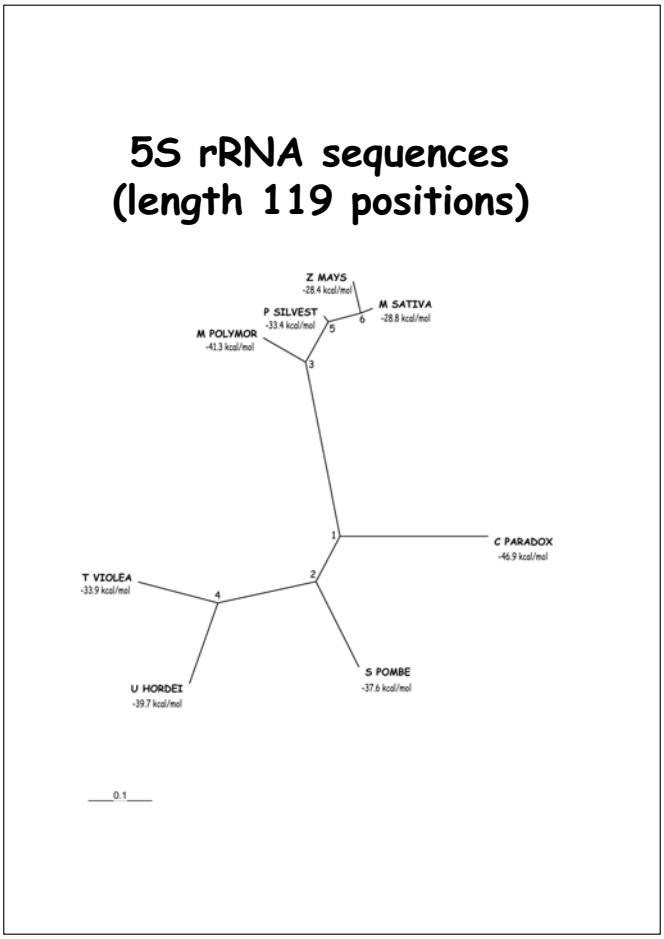
$E(i)$ is approximate energy of Sequence i using known secondary structure

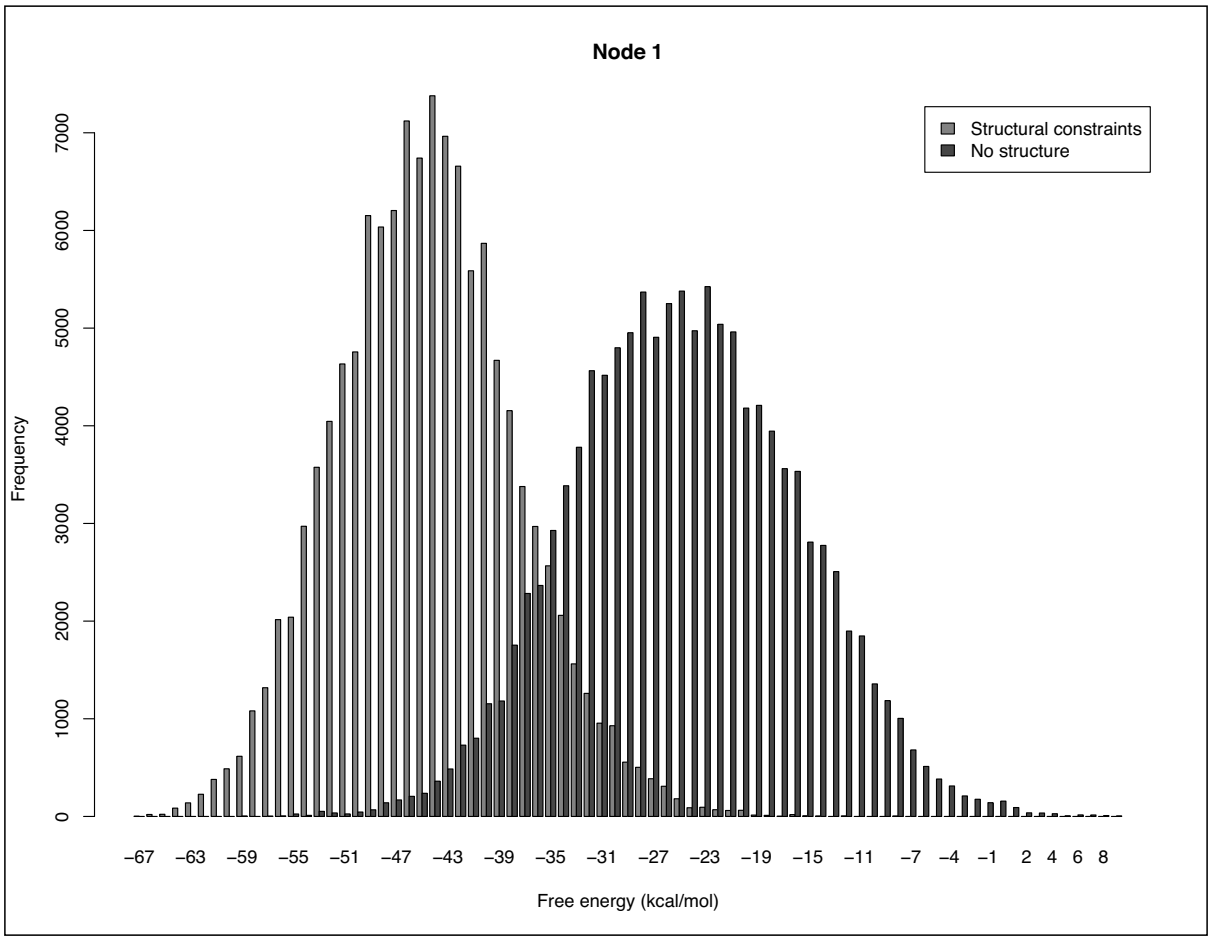
f relates energy to evolutionary rates

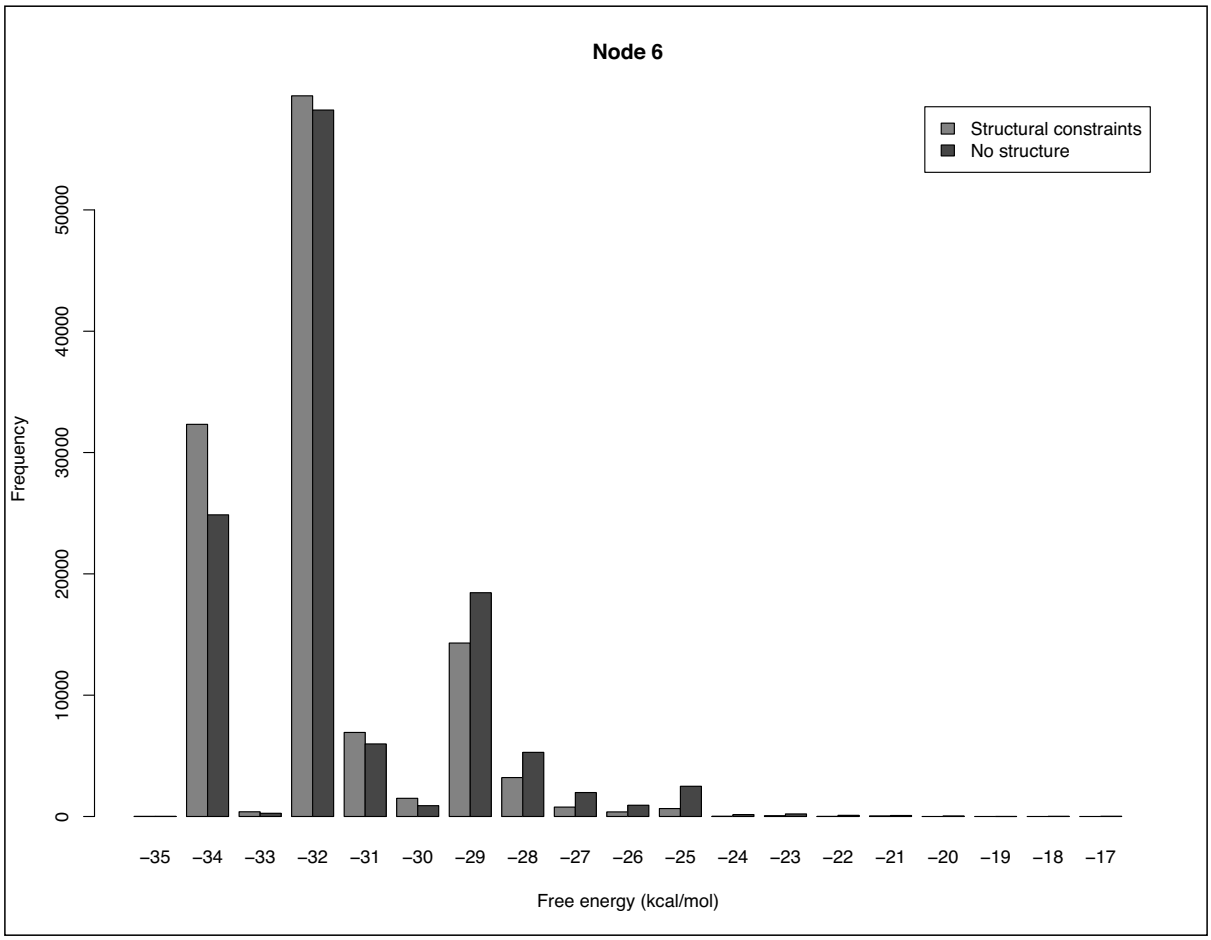
h is nucleotide type in Sequence j at sole position where i and j differ

$$R_{i,j} = \begin{cases} u\pi_h e^{(E(i)-E(j))f} & \text{for a transversion} \\ u\pi_h \kappa e^{(E(i)-E(j))f} & \text{for a transition.} \end{cases}$$

$e^{(E(i)-E(j))f} > 1$ is positive selection


5S rRNA sequences
(length 119 positions)
Z MAYS
-28.4 kcal/mol
M SATIVA
-28.8 kcal/mol
P SILVEST
-33.4 kcal/mol
M POLYMOR
-41.3 kcal/mol
C PARADOX
-46.9 kcal/mol
T VIOLEA
-33.9 kcal/mol
S POMBE
-37.6 kcal/mol
U HORDEI
-39.7 kcal/mol
0.1


Node 1
Structural constraints
No structure
Frequency
0
1000
2000
3000
4000
5000
6000
7000
−67 −63 −59 −55 −51 −47 −43 −39 −35 −31 −27 −23 −19 −15 −11 −7 −4 −1 2 4 6 8
Free energy (kcal/mol)



Protein Evolution References

Averof, M., A. Rokas, K.H. Wolfe, and P.M. Sharp. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science. **287**:1283–1286.

Cao Y, Adachi J, Janke A, Paabo S, Hasegawa M (1994) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. J Mol Evol 39: 519–527

Dayhoff, M.O., R.V. Eck, and C.M. Park. 1972. A model of evolutionary change in proteins. Pp. 89–99 in M.O. Dayhoff, ed. Atlas of protein sequence and structure, vol. 5, National Biomedical Research Foundation, Washington D.C.

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. Pp. 345–352 in M.O. Dayhoff, ed. Atlas of protein sequence structure, vol. 5, suppl. 3. National Biomedical Research Foundation, Washington D.C.

Goldman N, Yang Z. 1994. A codon–based model of nucleotide substitution for protein–coding DNA sequences. Mol. Biol. Evol. 11:725–736.

Gonnet, G.H., M.A. Cohen, and S.A. Benner. 1992. Exhaustive matching of the entire protein sequence database. Science 256:1443–1445.

Halpern, A., and W.J. Bruno. 1998. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. Mol. Biol. Evol. **15**:910–917.

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160

Muse, S.V. 1996. Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. **13**:105–114.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. Mol. Biol. Evol. 11:715–724.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

Parisi G. and J. Echave. 2001. Structural Constraints and Emergence of Sequence Patterns in Protein Evolution. Mol. Biol. Evol. 18(5):750-756.

Pedersen, A-M. K., C. Wiuf, and F.B Christiansen. 1998. A codon-based model designed to describe lentiviral evolution. Mol. Biol. Evol. 15:1069-1081

Pollock, D.D., W.R. Taylor, and N. Goldman. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287**:187–198.

Robinson, D.M., D.T. Jones, H. Kishino, N. Goldman, and J.L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20(10):1692-1704.

Schöniger, M., G.L Hofacker, and B. Borstnik. Stochastic traits of molecular evolution – acceptance of point mutations in native actin genes. J. Theor. Biol. **143**:287–306.

Models of Sequence Evolution: Nucleotide Substitution

Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. Bull Math Biol 51:79–94

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376. **(the paper that made maximum likelihood practical for phylogenies**

Felsenstein J., and G.A. Churchill. (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104

Jensen, J.L., and A.-M. K. Pedersen. 2000. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. Adv. Appl. Prob. 32:499-517.

Lockhart PJ, MA Steel, MD Hendy, D Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605-612 **(the LogDet)**

Pedersen, A.-M.K., and J.L. Jensen. 2001. A dependent-rates model and an MCMC-based methodology for the maximum likelihood analysis of sequences with overlapping reading frames. Mol. Biol. Evol. 18:763-776.

Yang Z (1993) Maximum–likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 39:306–314

Yang Z (1995) A space–time process model for the evolution of DNA sequences. Genetics 139:993–1005.