Any questions?

- Blast
- Phylogenies
- Likelihood Ratio Tests

Based upon this BlastP result, is the query used homologous to lysin?

Related Structures

Sequences producing significant alignments:		Score (bits)	E Value
gi|604525|gb|AAC37229.1|	fertilization protein >gi|1097388|...	275	2e-73
gi|604527|gb|AAC37230.1|	fertilization protein >gi|1097389|...	248	2e-65
gi|604533|gb|AAC37231.1|	fertilization protein >gi|1097390|...	212	3e-54
gi|3513683|gb|AAC33930.1|	fertilization protein precursor [...	94	8e-19
gi|604529|gb|AAC37232.1|	fertilization protein >gi|1097391|...	82	5e-15
gi|604531|gb|AAC37233.1|	fertilization protein >gi|1097392|...	67	1e-10
gi|12084519|pdb|1GAK|A	Chain A, Crystal Structure Of Green ...	62	3e-09
gi|538388|gb|AAA21518.1|	lysin	40	0.022
gi|4704776|gb|AAD28265.1|	sperm lysin precusor [Tegula fune...	36	0.25
gi|602977|gb|AAB59216.1|	sperm lysin	36	0.25
gi|538406|gb|AAB59168.1|	lysin	36	0.32
gi|538400|gb|AAB59167.1|	lysin	36	0.32
gi|602973|gb|AAA57303.1|	sperm lysin	36	0.32

The current size of the NR protein database is 680,984,053 and the PDB is 3,816,875.

Which should you use in a Blast search to have the most sensitivity in detecting homologs with known 3D structures? Why?

BlastX translates a DNA sequence into the 6 reading frames and searches each one against the protein NR database. BlastP takes a known translation and searches against the protein NR database.

Do you expect both searches to have the identical E-value? Why or why not?

Analysis of variation in the dn/ds ratio between sites or lineages

- Reading for today on website
  - Review article of methods discussed today.
- Reading for Thursday on website
  - Example of adaptive evolution using variation between lineages.

d_N/d_S ratio (ω) estimated across all sites is inefficient at detecting positive selection

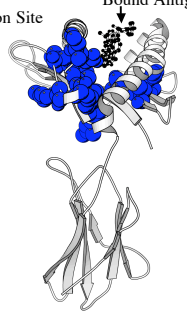
e.g. 3/44 (6.8%) sites subjected to positive selection

```
1  MSLAVLTFLVLCGFSFQHQAVGKWLTAAQKHPISGRMIRIRTKE
2  MSLAVLTFLVLCGWSFQHQAVGKWLTLAQKHPISGKMIRIRTKE
3  MSLAVLTFLVLCGYSFQHQAVGKWLTSAQKHPISGHMIRIRTKE
```

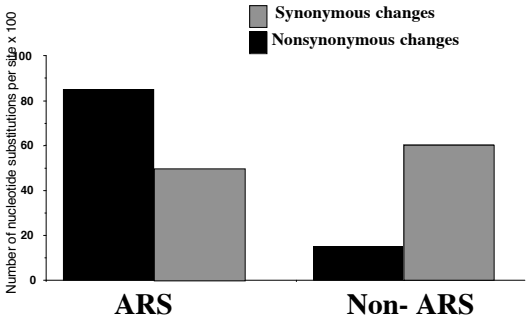
If functionally important regions are known, it is possible to analyze the regions separately

Antigen Recognition Site

Bound Antigen



Major histocompatability complex (MHC) has an antigen recognition site (ARS) and a structural region. The ARS binds foreign (i.e. viral) peptides and is known to be subjected to positive selection



This approach can work well if you know, a priori, where the functionally important parts of the gene are located (from functional or structural studies).

BUT, what if you don't have a prior functional or structural information?

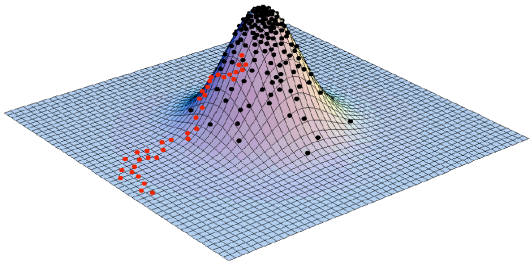
What is you have sequences from multiple species?

```
PINK      CGC CAC CGC TTC CGT TTT ATT CCA CAT
BLACK     --- CGC CGC TAT CAA TTT GTT CAA CAT
RED       --- CGC AGC TGG CAT TAT GTT GAA CCC
WHITE     --- CGC CGC TGG CAT TAT GTT CCA CCC
PINTO     --- CGC AGC TGG ACT TAT GTT CAA CCC
FLAT      --- CGC CGC TGG AAT TTT GTT ACA CCC
GREEN     --- CGC CGA TGG ACT TTT GTT CGA TAT
```

Markov models of codon evolution

Goldman & Yang 1994 *MBE* **11**:725-736

Muse & Gaut 1994 *MBE* **11**:715-724



Why use a likelihood model of codon evolution?

1. We can take advantage of the phylogeny
2. Computation of transition probabilities accomplishes the following in 1 step:
   i. estimation of parameters (t, κ, ω)
   ii. correction for multiple hits
   iii. weight evolutionary pathways between codons

Codon models

Important parameters:

- Transition/transversion rate ratio: κ
- Biased codon usage: π_j for codon j
- Nonsynonymous/synonymous rate ratio: $\omega = d_N/d_S$

Likelihood calculations require…

- An explicit model of substitution that specifies change probabilities for a given branch length:

$$\mathbf{Q} = \begin{pmatrix} \pi_A r_{AA} & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{CA} & \pi_C r_{CC} & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{GA} & \pi_C r_{GC} & \pi_G r_{GG} & \pi_T r_{GT} \\ \pi_A r_{TA} & \pi_C r_{TC} & \pi_G r_{TG} & \pi_T r_{TT} \end{pmatrix}$$

Jukes-Cantor
Kimura 2-parameter
Hasegawa-Kishino-Yano (HKY)
Felsenstein 1981, 1984
General time-reversible

- An estimate of optimal branch lengths in units of expected amount of change (v = rate x time)

$$\mathbf{P}(v) = e^{\mathbf{Q}v}$$

Rate matrix $Q = \{q_{ij}\}$

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at 2 or 3 positions} \\ \pi_j, & \text{for syn. transversion} \\ \kappa\pi_j, & \text{for syn. transition} \\ \omega\pi_j, & \text{for nonsyn. transversion} \\ \omega\kappa\pi_j, & \text{for nonsyn. transition} \end{cases}$$

$P(t) = \{p_{ij}(t)\} = e^{Qt}$

Rates to CTG

Synonymous

CT**C** (Leu) → CT**G** (Leu): π_{CTG}

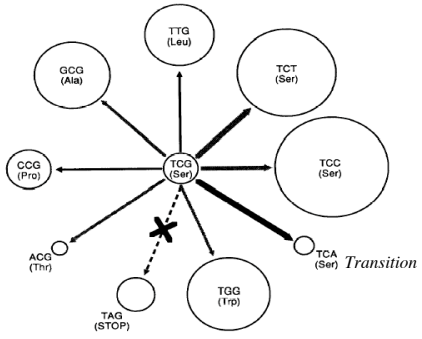
TTC (Leu) → **C**T**G** (Leu): $\kappa\pi_{CTG}$

Nonsynonymous

GTG (Val) → **C**TG (Leu): $\omega\pi_{CTG}$

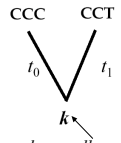
CCG (Pro) → CTG (Leu): $\kappa\omega\pi_{CTG}$

Maximum likelihood methods incorporate models of codon evolution / bias.



Simple case (only two codons):

The likelihood of observing the data of a pair of species.



Summed over all possible 61 codon ancestors for codon k

$$L_h(CCC, CCT) = \sum_k \pi_k p_{kCCC}(t_0) p_{kCCT}(t_1)$$

The likelihood of observing the entire sequence alignment is the product of the probabilities at each site.

$$L = L_1 \times L_2 \times L_3 \times \ldots \times L_N = \prod_{h=1}^{N} L_h$$

The log likelihood is a sum over all sites.

$$\ell(t, \kappa, \omega) = \ln\{L\} = \ln\{L_1\} + \ln\{L_2\} + \ln\{L_3\} + \ldots + \ln\{L_N\} = \sum_{h=1}^{N} \ln\{L_h\}$$

***t*, κ, ω estimated by MCMC. π_i estimated from codon frequencies.**

Remember: we are interested in adaptive evolution

ω = 1: neutral evolution

ω < 1: purifying (negative) selection

ω > 1: diversifying (positive) selection

Pairwise comparisons

- Calculate dn/ds from data
- Determine if value is significantly > 1.

For pairwise comparisons, we must determine if d_N/d_S is ***significantly*** greater than 1

- Estimate d_N/d_S ratio for pairwise comparison.
- Estimate d_N/d_S ratio with d_N/d_S fixed at one.
- Compare likelihoods using likelihood ratio test.

Example pairwise estimates

Estimating the dn/ds ratio:

H._sorenseni ... 1 H._rufescens
lnL = -567.039906
t= 0.1406 S= 108.2 N= 257.8 dN/dS= 6.0570 dN= 0.0622 dS= 0.0103

Fixing the dn/ds ratio equal to one:

H._sorenseni) ... 1 H._rufescens
lnL = -569.395789
t= 0.1390 S= 105.2 N= 260.8 dN/dS= 1.0000 dN= 0.0463 dS= 0.0463

Likelihood ratio test:

Example pairwise estimates

Estimating the dn/ds ratio:
H._sorenseni ... 1 H._rufescens
lnL = -567.039906
t= 0.1406 S= 108.2 N= 257.8 dN/dS= 6.0570 dN= 0.0622 dS= 0.0103

Fixing the dn/ds ratio equal to one:
H._sorenseni) ... 1 H._rufescens
lnL = -569.395789
t= 0.1390 S= 105.2 N= 260.8 dN/dS= 1.0000 dN= 0.0463 dS= 0.0463

Likelihood ratio test: -2[-569.4 - (-567.0)] = 4.8

+ selection

Example pairwise estimates

Estimating the dn/ds ratio:
H._walallensis ... 2 H._sorenseni
lnL = -568.604732

t= 0.1292 S= 95.2 N= 270.8 dN/dS= 1.2521 dN= 0.0454 dS= 0.0363

Fixing the dn/ds ratio equal to one:
H._walallensis ... 2 H._sorenseni
lnL = -568.661120

t= 0.1292 S= 94.5 N= 271.5 dN/dS= 1.0000 dN= 0.0431 dS= 0.0431

Likelihood ratio test:

Example pairwise estimates

Estimating the dn/ds ratio:
H._walallensis ... 2 H._sorenseni
lnL = -568.604732

t= 0.1292 S= 95.2 N= 270.8 dN/dS= 1.2521 dN= 0.0454 dS= 0.0363

Fixing the dn/ds ratio equal to one:
H._walallensis ... 2 H._sorenseni
lnL = -568.661120

t= 0.1292 S= 94.5 N= 271.5 dN/dS= 1.0000 dN= 0.0431 dS= 0.0431

Likelihood ratio test: -2[-568.6 - (-568.7)] = 0.2

Can not reject null model = neutral evolution.

Problem: averaging over a pair

In a pairwise analysis we must average the ω ratio over:

1. all sites
2. the entire evolutionary history

In a large-scale pairwise database search, only 17 out of 3,595 genes were found to be under positive selection, at <0.5% (Endo *et al*. 1996 *MBE* **13**: 685-690)

Problem: averaging over a pair has very low power if the questions are about "when" or "where"!

Solution: Phylogenetic estimation of selection pressure

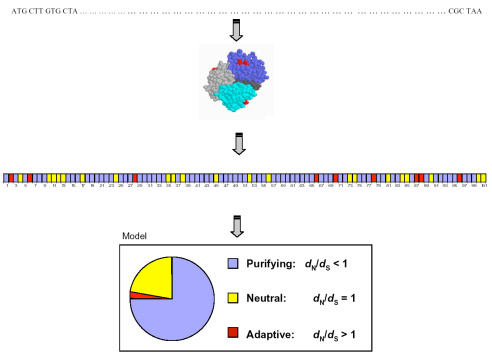
- variable ω over branches (when?)

- variable ω over sites (what fraction?)

Statistical methods (maximum likelihood) have been developed to test models of positive selection using genes. If a selection model fits the data better than a neutral model, one can identify sites subjected to positive selection and infer functional importance.

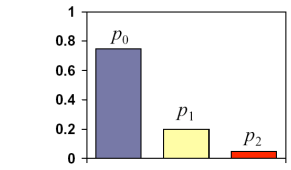
Variation between sites

- Model different classes of codon sites
  - Purifying, neutral, positive selection.
- Compare selection model with one class that has dn/ds > 1, to neutral model where all classes have dn/ds < 1 using likelihood ratio test.



What fraction ? — site models

Discrete model (M3) with $K = 3$ site classes



Site class 1: $\omega_0 = 0$, 60% of codon sites

Site class 2: $\omega_1 = 1$, 38% of codon sites

Site class 3: $\omega_2 = 5$, 02% of codon sites

What fraction? —Models of variable ω ratios among codons

Model	Code	NP	Parameters
One-ratio	M0	1	ω
Neutral	M1	1	p_0
Selection	M2	3	p_0, p_1, ω_2
Discrete	M3	2K-1	$p_0, p_1, \ldots, p_{K-2}$ $\omega_0, \omega_1, \ldots, \omega_{K-2}$
Frequency	M4	5	$p_0, p_1, \ldots, p_4$
Gamma	M5	2	α, β
2Gamma	M6	4	$p_0, \alpha_0, \beta_0, \alpha_1$
Beta	M7	2	p, q
Beta&ω	M8	4	p_0, p, q, ω
Beta&gamma	M9	5	p_0, p, q, α, β
Beta&normal+1	M10	5	p_0, p, q, α, β
Beta&normal>1	M11	5	p_0, p, q, μ, σ
0&2normal>1	M12	5	$p_0, p_1, \mu_2, \sigma_1, \sigma_2$
3normal>0	M13	6	$p_0, p_1, \mu_2, \sigma_0, \sigma_1, \sigma_2$

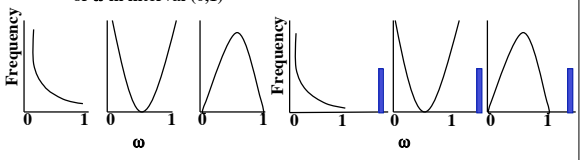
Variation between sites analyzed by likelihood ratio tests. Compare likelihood of neutral vs. selection models[1]

Neutral models

M1: $\omega_1 = 0$, $\omega_2 = 1$
or
M7: Assume distribution of ω in interval (0,1)

Selection models

M2: $\omega_1 = 0$, $\omega_2 = 1$, ω_S estimated
or
M8: distribution, ω_S estimated



[1]Method of Nielsen and Yang (1998), Yang et al. (2000)

1st: Testing for positive selection using likelihood ratio test statistic:

$$\Delta l = \log\left(\frac{\max\{L(\text{neutral model})\}}{\max\{L(\text{selection model})\}}\right)$$

$$= \log(\max\{L(\text{neutral model})\}) - \log(\max\{L(\text{selection model})\})$$

$-2\Delta l$ approximates χ^2 with n degrees freedom, where n is the difference in number of parameters between the nested models.

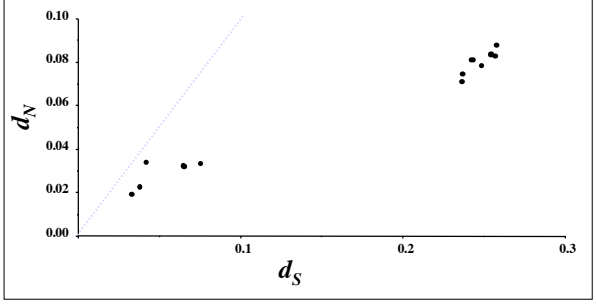
2nd: Identifying sites subjected to selection

If likelihood ratio test shows a significant difference, then positive selection is indicated. Given the distribution of ω, we then use an empirical Bayes approach to predict sites subjected to positive selection.

Example using MHC

- 6 alleles of human class I MHC aligned
  - Known to be subjected to positive selection
- Calculated dn/ds across all sites
- Test for variation in the dn/ds ratio between sites
  - Identification of sites under selection

MHC class I d_N/d_S ratio averaged across all sites



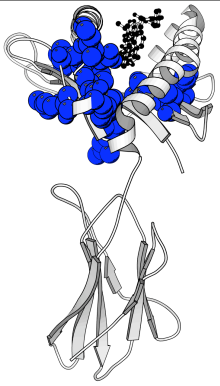
Tests of Positive Selection

Major Histo-Compatibility gene is known to be under positive selection

"Housekeeping" gene (Carbonic Anhydrase) is not under positive selection

Gene	*n*	*codons*	d_N/d_S	-2Δ l *M1 vs. M2*	-2Δ l *M7 vs. M8*	*Parameter estimates*	*Positively selected sites*
Positive control							
MHC	6	362	0.55	13.2**	14.7**	$p_1 = 0.073$, $\omega_s = 4.0$ $p_0 = 0.927$, B(p =0.10 ,q =0.27)	44, 48, 52, 55, 56, 82, 99, 101, 121, 126
Negative control							
CA	6	261	0.34	4.6	0.6	$p_0 = 1.0$, B(p =0.14 ,q =0.28)	None

** p<0.01

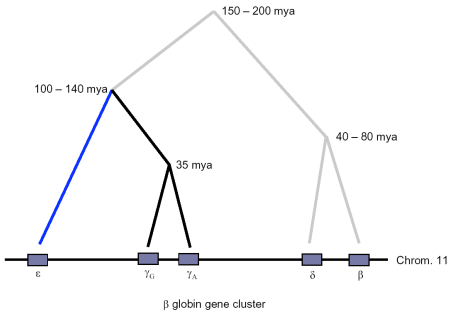


***MHC* residues predicted to be under positive selection are located in the antigen recognition site**

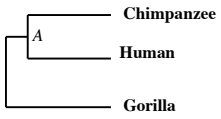
Variation between lineages

- What types of changes are lineage specific.
- Neutral theory predicts no variation in the d_N/d_S ratio between lineages.
- Likelihood ratio test between model of no variation and "free" estimates for each lineage.

Selection pressure varies in time



One method to detect lineage variation is to estimate ancestral sequence and perform pairwise comparisons:



A - Chimpanzee dn/ds = ?

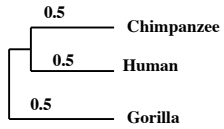
A - Human dn/ds = ?

Gorilla - A dn/ds = ?

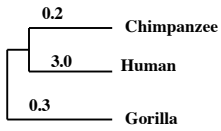
This method does not take into account uncertainty in estimating the ancestral sequence

Another method is to use a LRT to compare models of codon evolution:

One ratio (Neutral)

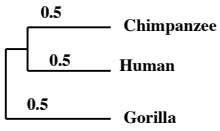


Free ratio (Selection)

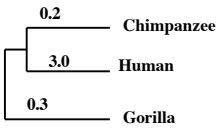


How many degrees of freedom?

One ratio (Neutral)



Free ratio (Selection)



The free ratio has 3 estimates of ω compared to 1 estimate for the neutral model. Degrees of freedom = 2.

Problems with lineage analysis

- Estimating dn/ds across all sites
  - There are "branch - site" models
- Many branches increase the degrees of freedom
  - Can limit estimates to particular branches, such as human lineage or following duplication

Summary

1. Pairwise methods have very low power to detect adaptive evolution.
2. Branch models allow variation among branches but assume one ω for all sites, and have low power to detect positive selection.
3. Site models assume allow variation among sites but assume selection pressure does not change among branches, and will have higher power if positive selection is long term

1. Why is averaging the dn/ds ratio across all sites in a gene inefficient at detecting adaptive evolution
2. What is one way to detect variation in rates of evolution along lineages.
3. Model 7 is a neutral model of DNA evolution estimates for a beta distribution with the parameters p and q. Model 8 add an additional proportion of sites with a dn/ds ratio estimated from the data. What are the degrees of freedom between these two models?
4. Describe a selective and neutral model that would be a useful for testing adaptive evoloution by a likelihood ratio tests. Include what parameters are different between the models.
5. How do you determine the degrees of freedom for a likelihood ratio test?
6. What are some important factors to consider when calculating dn/ds ratios?

PAML demo