Estimating selection pressure on different genomic regions

Maryam Zaheri and Nicolas Salamin

Feb 2010

Data

- Detect selection pressure on a group of species
  - Protein coding genes (Codons)
  - For example: rbcL and pepC

Natural Selection

- Natural selection is the process by which heritable traits that make it more likely for an organism to survive and successfully reproduce become more common in a population over successive generations. It is a key mechanism of evolution.

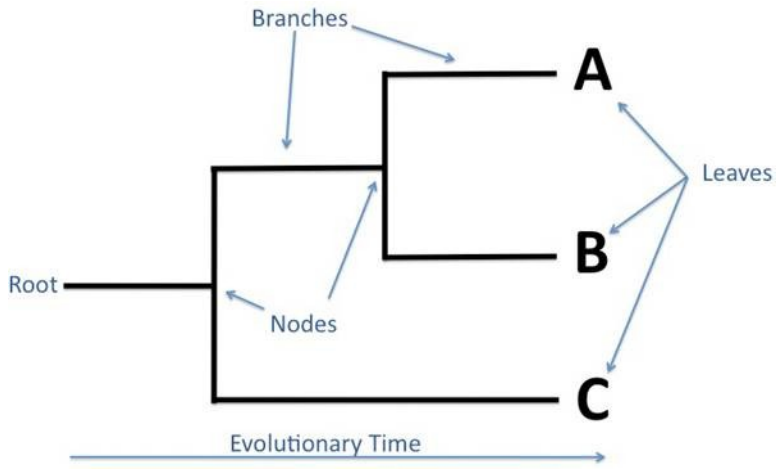
- “I have called this principle, by which each slight variation, if useful, is preserved, by the term of Natural Selection.” Darwin

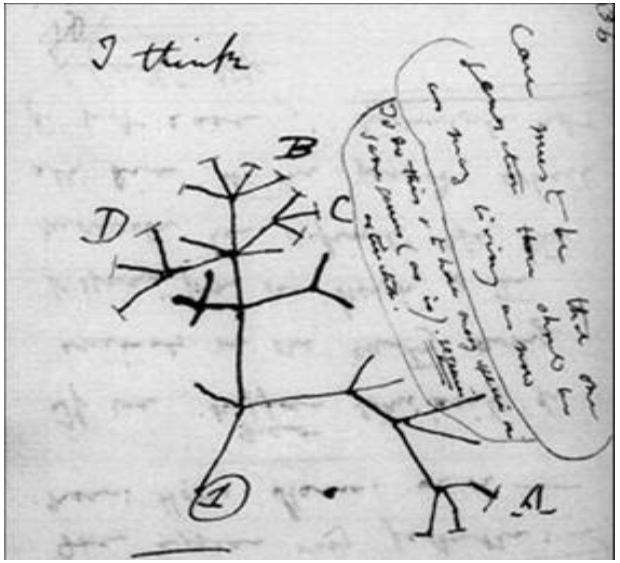
Selection pressure

- How to estimate selection pressure on a gene?
  - A codon substitution model, along phylogeny tree, for heterogeneous selection pressure among amino acid sites
  - Maximum likelihood estimation of parameters model

Phylogeny Tree

"The affinities of all the beings of the same class have sometimes been represented by a great tree... As buds give rise by growth to fresh buds, and these if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications." Darwin

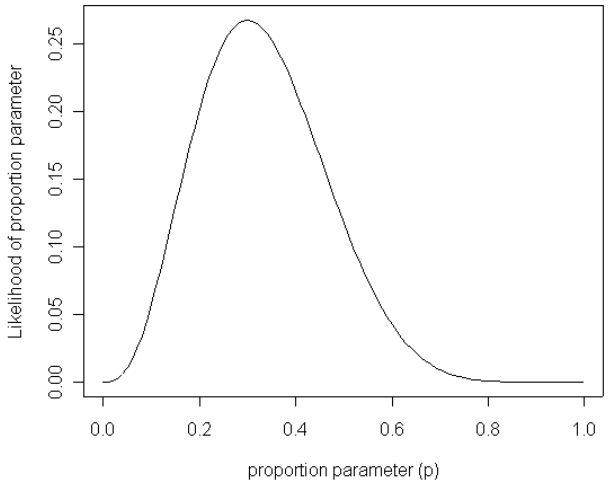




Likelihood

- Suppose you observe a random variable X in an experiment. Also assume that the distribution of the random variable X depends on a parameter θ. The function of probability density of X is given by fθ.

  – Example: take a coin and the binomial distribution



How to find the best value?

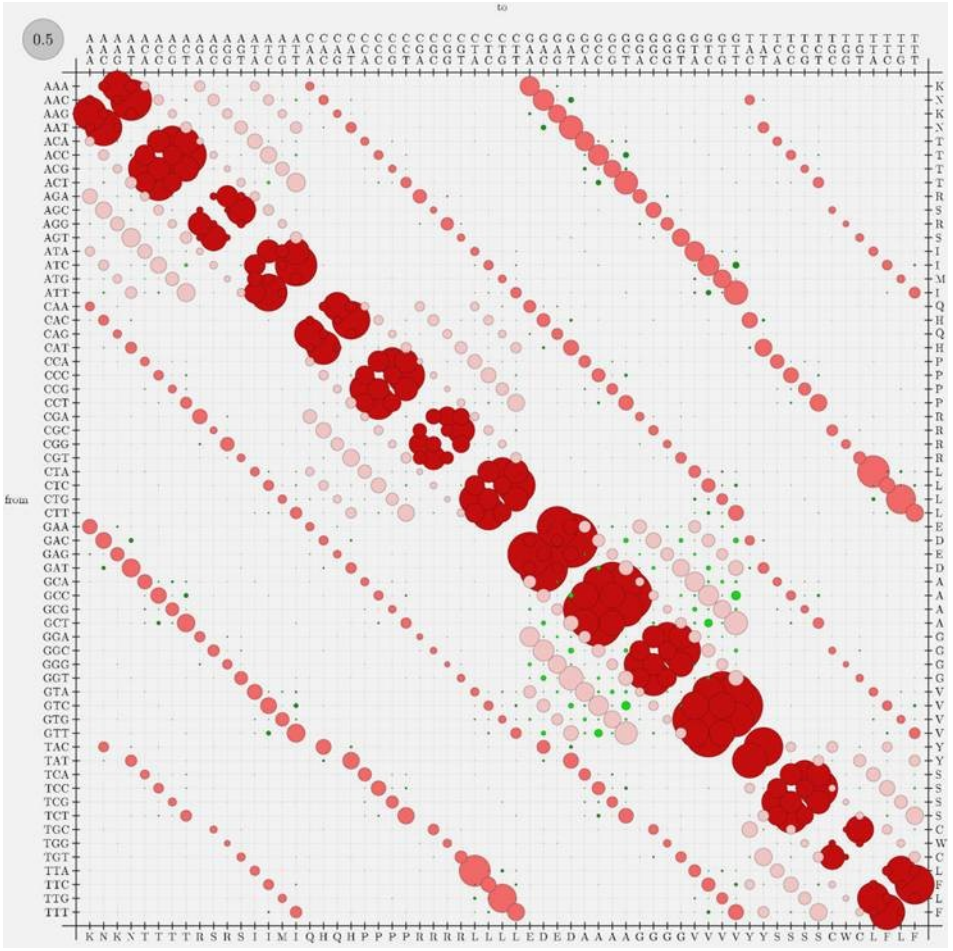
- The likelihood function L gives the probability of the parameter θ based on observed data X:
  $Lx(\theta) \approx f\theta(x)$

- In this method, we try to find the value of θ that maximizes likelihood function $Lx(\theta)$.

- If the likelihood function is differentiable, we find that estimator by calculating

$$\frac{\partial L_x \theta}{\partial \theta} = 0$$

Codon-based model

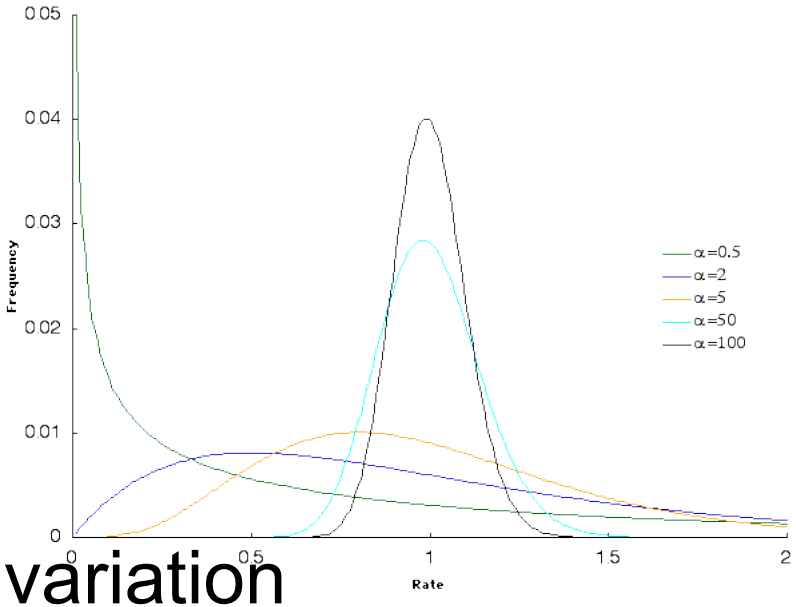


Heterogeneous selection pressure

- Assuming selection pressure among sites of each sequence follows a gamma distribution

$$f(\omega) = \beta^{\alpha} e^{-\beta\omega} \omega^{\alpha-1} / \Gamma(\alpha) \text{ for } \omega > 0.$$

$$F_G(\omega; \alpha, \beta) = \int_0^{\omega} \beta^{\alpha} e^{-\beta x} x^{\alpha-1} dx / \Gamma(\alpha).$$



- α < 1, strong among site variation

Tools

- Papers related to the topic
- Data sets of protein coding regions of several species
- Phylogenetic software for example Paml