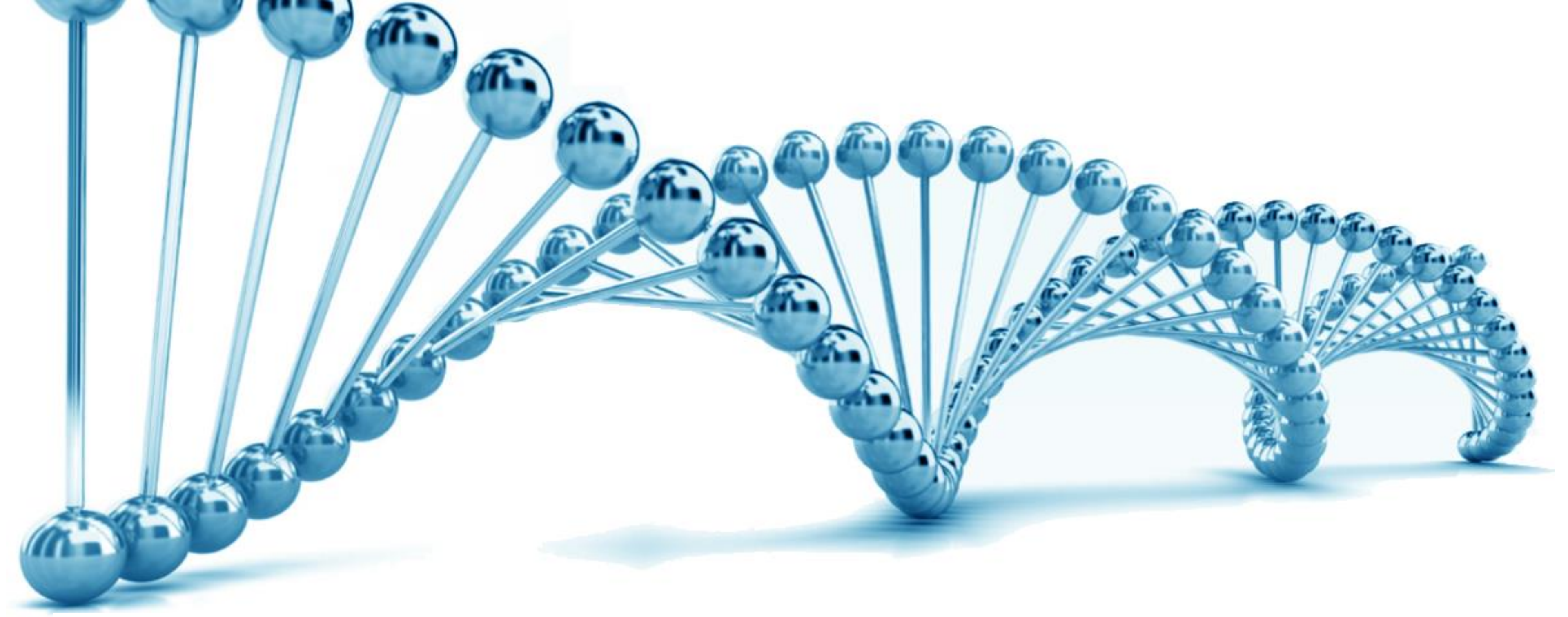

Comparing Selection Coefficients and Omega for Codon Substitution Rates

Albert Haque
Wilke Lab – Center for Computational Biology & Bioinformatics
University of Texas at Austin
October 11, 2013

Agenda

- Introduction
- Mutation Selection Model
    - Mutation Rate
    - Selection on Codon Usage
    - Selection on Protein
- Omega Model
    - Fixed Effects Likelihood
    - Random Effects Likelihood
- Omega-MutSel Comparison
- Research Goals

Introduction

- Codon Bias
  - Different frequencies for synonymous codons that code for the same amino acid
  - There is some external selective pressure
- How do we infer positive selection from DNA?

  Answer: Two models to examine selection at individual sites

  1. Selection Coefficient
  2. Omega

Motivation

- Everyone uses the Omega model
  - Easy to run
  - Been around for a long time
  - However, it leaves out information about underlying evolution

- Recently, mutation selection models have been developed to model selection pressure
- We don’t know which model is better

Software Packages

- Phylogenetic Analysis by Maximum Likelihood (PAML)
  - Estimates selective strengths on codon usage
- Hypothesis testing using Phylogenies (HyPhy)

- Software listed above can take a very long time (months)
- Need to get accustomed to software – what output, what parameters are required as input, etc.

Mutation Selection Model[1] (MutSel)

- Looks at the balance of mutation and selection
- Assume only one nucleotide change at a time
- Models the following:
  1. Nucleotide Mutation
  2. Selection on Codon Usage
  3. Selection on the Protein

[1] Yang, Z., and R. Nielsen. 2008. Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage.

Mutation Rate

First we need to define some variables:

μ_{ij} = mutation rate of nucleotide i to j in one generation

a_{ij} = nucleotide substitution rate from i to j from GTR[1] matrix

π_j^* = ***mutation bias***; we scale π_j^* such that $\sum \pi_j^* = 1$

Now we can calculate the mutation rate:

$\mu_{ij} = a_{ij}\pi_j^*$ where $a_{ij} = a_{ji}$ for all $i \neq j$

[1] Tavaré, S. 1986. "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences". Lectures on Mathematics in the Life Sciences (American Mathematical Society) 17: 57–86.

Selection on Codon Usage

Definitions:

f_I = fitness parameter for codon I

$s_{IJ} = f_J - f_I$ = ***selection coefficient*** for the mutation that changes codon I into J

To calculate the fixation probabilities:

$S_{IJ} = 2Ns_{IJ} = 2N(f_J - f_I)$ = ***scaled selection coefficient***

N = population size

$h(S_{IJ}) = S_{IJ}/(1 - e^{-S_{IJ}})$ = **ratio** of fixation probability of the $I \rightarrow J$ mutation to the fixation probability of a neutral mutation

Selection on Codon Usage

- Let Q denote the codon substitution matrix:
  - $q_{IJ} = \begin{cases} 0 & \text{if more than one change} \\ \mu_{i_k j_k} h(S_{IJ}) & \text{if synonymous substitution} \\ \omega \mu_{i_k j_k} h(S_{IJ}) & \text{if non-synonymous substitution} \end{cases}$

  Where k is the codon position in the sequence

- Why use ω?
  - Because it is simple and it produces similar estimates of mutation parameters as models that incorporate chemical properties[1]

[1] Yang, Z., and R. Nielsen. 2008. Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage.

Selection on the Protein

- Averaged over time, the proportion of $I \rightarrow J$ mutations among all mutations is:

$$m_{IJ} = \frac{\pi_I \mu_{i_k j_k}}{\sum_{I \neq J} \pi_I \mu_{i_k j_k}} \quad \text{and} \quad m_{IJ}^{+} = \frac{\pi_I \mu_{i_k j_k} \mathbb{1}}{\sum_{I \neq J} (\pi_I \mu_{i_k j_k} \mathbb{1})}$$

Note: $\mu_{ij} = a_{ij} \pi_j^{*}$ = mutation rate; S_{IJ} = scaled selection coefficient

- Where $\mathbb{1}$ is the indicator function:
  - $\mathbb{1} = 1$ if $S_{IJ} > 0$, and 0 otherwise
  - Only include advantageous mutations

- Thus, the strength of ***positive selection*** on the protein is:

$$\bar{S}_{+} = \sum_{I \neq J} (m_{IJ}^{+} S_{IJ} \mathbb{1})$$

Omega Models

- Compare synonymous and non-synonymous mutations
- $\omega = \frac{dN}{dS}$

  $if\ \omega < 1$ implies purifying selection

  $\omega = 1$ implies neutral mutations

  $\omega > 1$ implies diversifying positive selection

- Typically calculated by taking average over all codons
- Problem: It becomes difficult for $\omega > 1$
- Possible Solution: Create statistical models for ω

[1] Z. Yang, et al. 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites.

Random Effects Likelihood (REL)

- If we use one ω for each site, we get too many parameters
- Probability of observing data x_h given site h:

$$f(x_h) = \sum_{k=1}^{2} p_k f(x_h|\omega_k) = p_1 f(x_h|\omega_1) + p_2 f(x_h|\omega_2)$$

$h = \{1,2,\dots,\mathrm{n}\}$ and p = proportion of codon sites in categories

Two categories:

1. Non-synonymous mutations are neutral
2. Non-synonymous sites are eliminated by selection

R. Nielsen and Z. Yang. 1998. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene.

Fixed Effects Likelihood (FEL)

- Keep the model parameters fixed:
  - Branch lengths
  - Nucleotide rate biases
  - Tree topology
- Using a FEL rate matrix[1], we can compute each site
- Apply a likelihood test to determine significance

- Can process gene-size alignments of several hundred sequences in a few hours on a small cluster

[1] S. Pond, and S. Frost. 2005. Not So Different After All: A comparison of Methods for Detecting Amino Acid Sites Under Selection

Omega-MutSel Comparison

$\lambda_a =$ parameter determining frequency of amino acid A

“scaled selection coefficient”

$$F(a) \sim e^{-\lambda int(aa)} \text{ = fitness}$$

$$\pi_{a \rightarrow b} = \frac{1 - [F(a)/F(b)]^{\frac{1}{N}}}{1 - F(a)/F(b)}$$

$$K = \mu N \sum_{a} [F(a) \sum_{b} \pi_{a \rightarrow b}]$$

- If we perform some algebra on $\pi_{a \rightarrow b}$, we can eliminate N from the K equation.

[1] D. Ramsey, M. Scherrer, T. Zhou, and C. Wilke. The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. Genetics 2011.

Omega-MutSel Comparison

$$dN = \frac{K_N}{N_N} = \frac{\mu N \sum_i \sum_{j \in \mathcal{M}} F(i) \pi_{i \to j}}{\sum_i \sum_{j \in \mathcal{M}} F(i)}$$

$$dS = \frac{K_s}{N_s} = \mu$$

- No omega was used to calculate dN or dS
- Remember:
  - $q_{IJ} = \begin{cases} 0 & \text{if more than one change} \\ \mu_{i_k j_k} h(S_{IJ}) & \text{if synonymous substitution} \\ \omega \mu_{i_k j_k} h(S_{IJ}) & \text{if non-synonymous substitution} \end{cases}$

Conclusion

Current and Next Steps

- Software currently exists, but it requires long computation
- We are running a MutSel model on PAML to understand various (input and intermediate) parameters and output

Research Goals

- Compare Omega Models with Selection Coefficient
- Is one model better than the other?
- When is one model more appropriate? Under what conditions?