What Is Conservation?

Lee A. Newberg

February 22, 2005

A Central Dogma

Junk DNA mutates at a background rate, but functional DNA exhibits *conservation*.

Today's Question

What is this conservation?

Definition Possibilities

Sequence is said to be conserved across species . . .

Parsimony

if there are few base mismatches.

Statistical #1

if the best model shows reduced phylogenetic distances between the species.

Statistical #2

if the best model requires the incorporation of selection pressures.

Statistical #1 *vs.* Statistical #2

They aren't necessarily the same!

Example: Selection Is Ten Times as Significant as Mutation

- *Mutation*: Suppose background mutation rate is 1% per 10^6 years, with each of A, C, G, and T equally likely to mutate.
- *Selection*: Suppose that after 10^5 years, the expected number of descendents of a C genotype (or G or T genotype) is 1% less than the expected number of A genotype descendents.

Change in Equilibrium

With a statistical model that incorporates selection pressures, we can compute the population equilibrium exactly:

$$(A, C, G, T) \approx (0.90, 0.03, 0.03, 0.03) \ . \tag{1}$$

No Change in Mutation Rate

However, even with these selection pressures and this skew equilibrium, we expect approximately one in 10^8 nucleotides to mutate each year.

Conclusion?

"Conservation" would better be used to indicate the nonuniformity of an equilibrium distribution rather than a reduced rate of substitution.

But This Isn't the Popular Definition

So why do folks reduce the rate of mutation?

Parsimony: Counting Mismatches

A skew distribution reduces the number of mismatches.

Example

With equilibrium distribution, *e.g.*, $(A, C, G, T) = (0.7, 0.1, 0.1, 0.1)$, infinitely evolutionarily distant species show joint probability distribution of

$$\begin{pmatrix} 0.49 & 0.07 & 0.07 & 0.07 \\ 0.07 & 0.01 & 0.01 & 0.01 \\ 0.07 & 0.01 & 0.01 & 0.01 \\ 0.07 & 0.01 & 0.01 & 0.01 \end{pmatrix}, \tag{2}$$

regardless of the nucleotide mutation model.

This is 48% mismatches, compared to $\approx 75\%$ for neutral sites.

Based upon parsimony criteria the species appear closer!

Statistician would say previous example shows statistical independence → still infinitely distant. But, ...

Statistical Phylogeny of Mixed Distributions

First codon positions are conserved. Suppose first codon positions come in four kinds: *A predominant* with equilibrium $(0.7, 0.1, 0.1, 0.1)$ and also *C predominant*, *G predominant*, and *T predominant*.

For infinitely evolutionarily distant species, the joint probability distribution for the *A predominant* kind is as before:

$$\begin{pmatrix} 0.49 & 0.07 & 0.07 & 0.07 \\ 0.07 & 0.01 & 0.01 & 0.01 \\ 0.07 & 0.01 & 0.01 & 0.01 \\ 0.07 & 0.01 & 0.01 & 0.01 \end{pmatrix}, \tag{3}$$

regardless of the nucleotide mutation model, and likewise for *C predominant*, *G predominant*, and *T predominant* with rows and columns appropriately permuted.

Statistical Phylogeny of Mixed Distributions, cont'd

If each of the four kinds is equally likely than the mixed joint distribution is:

$$\begin{pmatrix} 0.13 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.13 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.13 & 0.04 \\ 0.04 & 0.04 & 0.04 & 0.13 \end{pmatrix} . \tag{4}$$

This is the joint distribution one would get from:

- the model of Jukes & Cantor (1969); or
- the model of Felsenstein (1981) with uniform nucleotide distribution; or
- the model of Hasegawa *et al.* (1985) with uniform nucleotide distribution and a transition / transverion ratio of $\kappa = 1$.

Regardless, the implied phylogenetic distance is ≈ 0.7662, *not* infinite.

Based upon statistical criteria the species appear closer!

Conclusion?

Because of

- a focus on mismatch counts in evolutionarily distant species; and/or
- a focus on mixtures of distributions

folks have been *misled* (?!) into believing that conservation reduces the rate of mutation.

Where's the Math? (Newberg, 2005)

Mutation Model

In a short generation time ϵ the nucleotide substitution matrix won't be very different from the identity:

$$I + \epsilon R \ . \tag{5}$$

For example, $R_{AC} = R_{AG} = R_{AT} = 10^{-8}/3$ and $\epsilon = 0.02$.

Selection Model

In a short generation time ϵ the selection model matrix won't be very different from the identity:

$$I + \epsilon S \ . \tag{6}$$

For example, $S_{CC} = S_{GG} = S_{TT} = -10^{-7}$ and $\epsilon = 0.02$.

Other Time Periods

Each generation has a chance to mutate and then a chance to be selected out.

Repeating for a time t gives

$$M_t = [(I + \epsilon R)(I + \epsilon S)]^{t/\epsilon} \ . \tag{7}$$

Starting with an ancestor with distribution $\vec{\beta}$, the joint distribution with a descendent is given by

$$J_t = \frac{D_{\vec{\beta}} M_t}{\vec{1} D_{\vec{\beta}} M_t \vec{1}^T} \ , \tag{8}$$

where

$$D_{\vec{\beta}} = \begin{pmatrix} \beta_A & 0 & 0 & 0 \\ 0 & \beta_C & 0 & 0 \\ 0 & 0 & \beta_G & 0 \\ 0 & 0 & 0 & \beta_T \end{pmatrix} \text{, and} \tag{9}$$

$$\vec{1} = (1, 1, 1, 1) \ . \tag{10}$$

Off-Diagonal Elements of J_t

For closely related species, the expected number of nucleotide mismatches is proportional to the evolutionary distance. This constant of proportionality is the mutation rate relative to background.

Equilibrium from Selection Pressures

If, due to selection pressures, the equilibrium changes from $\vec{\beta}$ to $\vec{\theta}$, then we can show that

$$\text{ods}\left(\left.\frac{\partial J_t}{\partial t}\right|_{t=0}\right) \approx \text{ods}\left(D_{\vec{\theta}}R\right) \ , \tag{11}$$

where ods$(\cdot)$ means off-diagonal sum. (Note, ϵ and S drop out.)

OrthoGibbs, PhyloScan, etc.

Note that even if S is not known, so long as $\vec{\theta}$ is known (or estimated) we can calculate Formula 11. (Recall that R depends on only the background model.)

Calculating the Mutation Rate

$$\text{ods}\left(\left.\frac{\partial J_t}{\partial t}\right|_{t=0}\right) \approx \text{ods}\left(D_{\vec{\theta}}R\right) \ , \tag{12}$$

- If R is the model of Jukes & Cantor (1969) then this gives 1, regardless of the selection matrix S and the selection-sensitive distribution $\vec{\theta}$.
- With the model of Hasegawa *et al.* (1985), when the junk-DNA equilibrium, $(\beta_A, \beta_T, \beta_C, \beta_G)$, equals $(0.3, 0.3, 0.2, 0.2)$ and κ equals 3, the overall instantaneous rate of Formula 12 will fall in the interval

  $$[0.901, 1.148] \ , \tag{13}$$

  regardless of the selection matrix S and the selection-sensitive distribution $\vec{\theta}$.

For reasonable *background* models, the number of mismatches between closely related species is nearly the same when considering functional *vs.* junk positions

Extreme Selection

The analysis above discusses a fitness time scale of 10^7 years.

Q. What if the selection time scale is one to a few generations?

A. Mutation rates can go down.

The formula for the mutation rate, with error term, is:

$$\text{ods}\left(\left.\frac{\partial J_t}{\partial t}\right|_{t=0}\right) = \text{ods}\left(D_{\vec{\theta}}R\right)\left[1+\mathcal{O}\left(\epsilon(R+S)\right)\right] \ , \tag{14}$$

Applicable to TFBSs?

Extreme selection also gives an extreme distribution for $\vec{\theta}$, *e.g.*,

$$\vec{\theta} = (0.9999990, 0.0000003, 0.0000003, 0.0000003) \ . \tag{15}$$

Do we see that?

Conclusion?

For TFBSs, selective pressures are sufficiently subtle, and conservation does not significantly affect the mutation rate.

References

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol,* **17** (6), 368–376. PubMed 7288891.

Hasegawa, M., Kishino, H. & Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol,* **22** (2), 160–174. PubMed 3934395.

Jukes, T. H. & Cantor, C. (1969) Evolution of protein molecules. In *Mammalian Protein Metabolism*, (Munro, H. M., ed.), vol. 3,. Academic Press. New York, NY pp. 21–132.

Newberg, L. A. Selection pressures do not significantly affect mutation rates. In preparation.