Analysis of Kellis-Birrer-Landis

Benjamin Peter

May 26, 2015

1 Accelerated evolution

The authors constructed a phylogeny as follows: Let K denote the outgroup species, Y and Z the two S. c. paralogs. Let $D(K, Y|G)$, $D(K, Z|G)$ and $D(Y, Z|G)$ be the number of nucleotide substitutions observed between the pairs of gene sequences. Then, the transformed statistics

$$S(Z|G, K, Y) = \frac{1}{2}(D(K, Z|G) + D(Y, Z|G) - D(K, Y|G))$$
$$S(Y|G, K, Z) = \frac{1}{2}(D(K, Y|G) + D(Y, Z|G) - D(K, Z|G))$$
$$S(K|G, Y, Z) = \frac{1}{2}(D(K, Z|G) + D(K, Z|G) - D(Y, Z|G))$$

give the branch-specific mutations, i.e. the number of sites where two of the genome agree, but not the third. I'll omit the dependencies on the other two genotypes and just write $S(Z)$ for brevity.

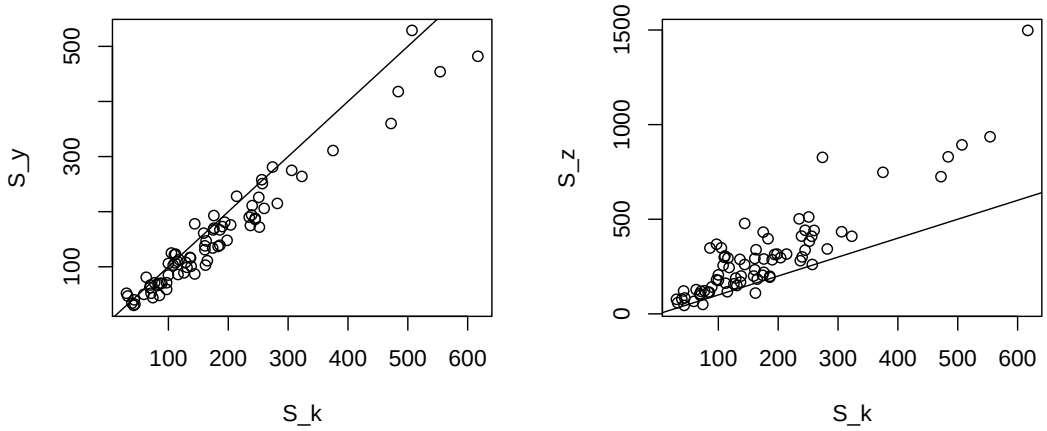
Then, the test of accelerated evolution is equivalent to testing the hypothesis that $S(K) = \frac{(S(Y)+S(Z)}{2}$, with alternative hypothesis $S(K) < \frac{(Z)+S(Y)}{2}$. The idea of the approach is that substitutions accumulate due to a molecular clock at a rate that is constant with time from the common ancestor of the three species. They reject the null hypothesis when

$$\frac{S(Z) + S(Y)}{2S(K)} > 1.5$$

or something similar. I could not reproduce that result from their data, however the set of 76 genes is given and can be found intersecting the tables included `S9_Trees/Duplicated_Pairs.xls` and `S9_Trees/Nucleotide_Divergence.xls` files from the supplement. The S values can be calculated from the alignments provided in the supplement to the paper, where I ignored loci that showed gaps in any of the three sequences.

The first figure shows that the data for these 76 gene more or less supports the claim in the paper that one gene shows similar divergence S_z than the outgroup gene S_K, whereas the other copy looks different (note that I ordered the gene copies such that $S_Y < S_Z$.

Figure 1: Seems reaonable..



For the test Lior asked about, we look at all pairs of genes that are identified as having an accelerated rate, and figure out if there are one or two rates of acceleration. We can formulate that in the following hypotheses

1. $\mathbf{H_0}$: $S_k < S_y, S_k < S_z$ (two different rates of accelerated evolution)
2. $\mathbf{H_1}$: $S_k = S_y, S_k < S_z$ (one rate of accelerated evolution)

Assuming that mutations are independent, S_k, S_y and S_z are distributed multinomially with likelihood

$$L = \log\left[\left(\sum S_i)\right)!\right] - \sum \log S_i + \sum S_i * p_i$$

where all sums are over (k, y, z). For $\mathbf{H_0}$, we find p_k, $p_y = p_k + p_y$ and $p_z = p_k + i_z$ subject to the constraints $p_z + p_k + p_y = 0, i_z \geq 0 i_y \geq 0$. For $\mathbf{H_1}$, we use the same likelihood, but with constraint $i_y = 0$. For a classical likelihood-ratio test, we would want to have the hypotheses switched, but looking at the data reveals that for 62 of the 76 samples, the maximum likelihood estimate of i_y is zero. So for these 62 samples, we would prefer $\mathbf{H_1}$ under pretty much any framework. Indeed, using BIC, we find that $\mathbf{H_0}$ is preferred over $\mathbf{H_1}$ for all samples. Similarly, looking at the statistic $D = 2 * L(\mathbf{H_0} - L(\mathbf{H_1})$ shows that only four of the values have a D greater than 2. As we would expect D to be χ^2_1 distributed, it is clear that the vast majority of genes are consistent with the claimed "only one of the genes shows accelerated evolution". It probably would be possible to wrangle that into a statement about p-values (and I fear if somebody did that, Lior might go bankrupt), but it does not seem worth the effort.

As a biologist, that is not really surprising; these are likely loss-of-function mutations and have probably little to do with any functionality.

Figure 2: Tests