Supplementary Material: Substitution Scores

For every species pair, there are two phylogenetic branches, or lineages, representing lines of evolutionary descent of extant species (except for mouse-rat) from the ancestral species. If one to one ortholog proteins exist in both species in a species pair, we infer the existence of an ancestral protein and compute a substitution score representing evolutionary selection strength in each of the two lineages leading to the ortholog proteins. These scores are then normalized (see below) and compared within pairs to determine whether the ancestral protein evolved significantly faster in one lineage than another. The substitution score of a protein in a branch is computed by retrieving the pairwise sequence alignment of the predicted ancestral protein and the ortholog protein sequences of the extant species (except for mouse-rat), and summing the Grantham score of each aligned residue. Since the Grantham matrix measures the physiochemical differences between each amino acid, our substitution score measures the dissimilarity between the ancestral and extant protein.

Since different lineages have different evolutionary rates, one cannot simply compare the number of residue substitutions or the scores computed based on the number of substitutions alone between two lineages sharing the same ancestor. Thus, to compare the evolutionary scores between two lineages, we had to normalise the evolutionary scores specifically for each experimental and control pairs.

In order to normalise the scores properly, let us assume that we are comparing the evolutionary pressure on genes or proteins in two branches (A and B) with different evolutionary rate. We would like to be able to detect whether one gene or protein is under heavier positive pressure in one branch compared to the other. Let $SP_A(p)$and $SP_B(p)$denote a measure of selective pressure in branch $A$and B respectively for a protein p. In our case, it is the sum of Grantham measure of all substitutions between the predicted ancestral sequence and each of the extant species. Since the two lineages represented by the branches may have different evolutionary or mutational rates, we can expect a higher measure in one branch, e.g. $SP_A(p)$ higher than $SP_B(p)$, for any protein p. Thus, comparing $SP_A(p)$ to $SP_B(p)$ directly would

heavily bias the analysis as proteins with high evolutionary pressure in branch A are more likely to be detected as having a significantly higher score than in branch B.

Instead, we need to capture this difference in evolutionary and mutational rate to normalise the two scores $SP_A(p)$ and $SP_B(p)$. The simplest solution we found was to normalise both scores by the expected number of substitutions each ancestral protein sequences given the protein sequence and linage. We first computed for each lineage *A* and *B*, the empirical probability distribution of amino acid residue substitutions using all the proteins with ancestral sequence prediction. Then, we normalised the selective pressure scores for each protein. This number will be an underestimate of the true number of substitutions in a lineage as there can be multiple substitutions at one residue. However, since the considered proteins are well conserved and the branches are generally short, this approximation should not alter the results in any significant way.

In mathematical terms, let $E(A,B)$ be the set of all ancestral sequences predicted for species *A* and *B*. The empirical probability that amino acid *R* is substituted for another amino acid in branch *A* is:

$$P(R;A,B) = \frac{\sum_{p\in E(A,B)} \sum_{i} \mathbf{1}_{\{aa_{ancestral}(p;i)=R, aa_A(p;i)\neq R\}}}{\sum_{p\in E(A,B)} \sum_{i} \mathbf{1}_{\{aa_{ancestral}(p;i)=R\}}}$$

where the first summation runs through all predicted ancestral sequences for species *A* and *B* while the second summation runs through the amino acid in each protein. $aa_{ancestral}(p;i)$ is the amino acid in the ancestral protein sequence *p* at position *i* of the alignment and $aa_A(i)$ is the amino acid sequence of the descendant of protein *p* of species *A* at position *i*.

Thus, the expected number of substitutions the ancestral protein sequence of a protein $p = aa_1aa_2...aa_n$ will have in lineage *A* is

$$N_A(p) = \sum_{i=1}^{n} E\{\mathbf{1}_{aa_{ancestral}(p;i) \neq aa_A(p;i)}\} = \sum_{i=1}^{n} P(aa_{ancestral}(p;i); A, B)$$

while the expected number of substitutions the ancestral protein sequence will have in lineage *B* is

$$N_B(p) = \sum_{i=1}^{n} P(aa_{ancestral}(p;i); B, A)$$

Thus, by normalising $SP_A(p)$ by $N_B(p)/N_A(p)$, we can compare the corrected evolutionary pressure scores $SP_A(p) \cdot N_B(p)/N_A(p)$ and $SP_B(p)$to infer in which lineage protein *p* was under heavier selection.