I. BAYESIAN SEPARATION OF GENES WITH SELECTION ON 4-FOLD SITES

We assumed the distribution of the 4-fold sites conservation rate distribution was a mixture of two binomial processes – one corresponding to neutral conservation and the other corresponding to conservation due to purifying selection. This is certainly a simplification as purifying selection can in general be caused by multiple processes.

The conservation rates were estimated by maximizing the posterior probability $P(p_n, p_s, \lambda|D)$, where p_n is the neutral conservation rate, p_s is the conservation rate for ORFs under selection, λ is the percentage of ORFs under selection and D is the observed conservation rates of 4-fold sites for each ORF. Bayes' rule gives

$$P(p_n, p_s, \lambda|D) = \frac{P(p_n, p_s, \lambda)P(D|p_n, p_s, \lambda)}{P(D)} \tag{1}$$

We used a uniform prior $P(p_n, p_s, \lambda)$ for $p_n < p_s$ and $P(p_n, p_s, \lambda) = 0$ for $p_s > p_n$. Thus, maximizing the posterior is equivalent to maximizing the likelihood $P(D|p_n, p_s, \lambda)$ in the region $p_n < p_s$. The likelihood is

$$P(D|p_n, p_s, \lambda) = \prod_i^N (1-\lambda) p_n^{c_i} (1-p_n)^{m_i} + \lambda p_s^{c_i} (1-p_s)^{m_i}, \tag{2}$$

where c_i and m_i are the number of conserved and non-conserved 4-fold sites in the ith ORF and N is the total number of ORFs. An iterative process was used to search for the maximum over p_n,p_s and λ. The likelihood that the ith ORF is under selection is s_i, where

$$s_i = \frac{\lambda p_s^{c_i} (1-p_s)^{m_i}}{(1-\lambda) p_n^{c_i} (1-p_n)^{m_i} + \lambda p_s^{c_i} (1-p_s)^{m_i}}. \tag{3}$$

p_n and p_s were updated by the relations

$$p_n = \frac{\sum_i^N (1-s_i) c_i}{\sum_i^N (1-s_i)(c_i + m_i)} \tag{4}$$

$$p_s = \frac{\sum_i^N s_i c_i}{\sum_i^N s_i (c_i + m_i)}. \tag{5}$$

After p_n and p_s were updated, we updated λ by finding the value of λ that maximizes the likelihood $P(D|p_n, p_s, \lambda)$. We tested this algorithm on simulated data, and it correctly recovered the p_n,p_s, and λ used to generate the data.

This algorithm yielded p_n consistent with the mode neutral conservation rates in the *S. cerevisiae-S. paradoxus* (0.74) and 4 species (0.33) comparisons. It was also used to calculate the neutral rates for the *S. cerevisiae-S. bayanus* and *S. cerevisiae-S. mikatae* comparisons.