PHYML technical documentation

March 1, 2004

Substitution models in PHYML correspond to homogeneous, stationary and time-reversible Markov processes. Therefore, the likelihood does not depend on the position of the root of the phylogeny. Let r be this root, and R denotes the tree. u and v are the roots of subtrees U and V respectively. u is a tip of R that is separated from v by a single branch of length l. π_h is the equilibrium frequency of state h. $P_{hh'}(l)$ is the probability for the state h to be replaced by state h' after l substitution events. The substitution rate at each site is distributed as a discretized gamma distribution with G categories. r_g is the relative rate for cateogory g and p_g is its probability. $\mathcal{A}$ is the state space to be considered. We define the conditional likelihood $L(s = h|U)$ as the probability of data at site s given that node u has state h. $L(s = h|V)$ has the same meaning when V (and v) replaces U (and u). $L^*(s = h|V) = L(s = h|V)/Z_s(V)$ and $L^*(s = h|U) = L(s = h|U)/Z_s(U)$ are scaled conditional likelihoods. $Z_s(V)$ and $Z_s(U)$ are scale factors that are used to avoid numerical underflows when computing very small values of conditional likelihoods. We first consider the case where the state observed at the leaf u is not ambiguous. The scaled-likelihood at site s is then :

$$L^*(s) = \sum_{g}^{G} \sum_{h \in \mathcal{A}} p_g \pi_h L^*(s = h|V) P_{hh'}(l \times r_g) \tag{1}$$

If the state at tip u is ambiguous, each potential state has to be considered (e.g., for DNA the states 'A' and 'G' are the potential states that are considered when the observed state is

'R'). The scaled-likelihood is then :

$$L^*(s) = \sum_{g}^{G} \sum_{h \in \mathcal{A}} \sum_{h' \in \mathcal{A}} p_g \pi_h L^*(s = h|V) P_{hh'}(l \times r_g) L^*(s = h'|U) \tag{2}$$

The (unscaled) log likelihood is then :

$$ln(L(s)) = ln(L^*(s)) + ln(Z_s(U)) + ln(Z_s(V)) \tag{3}$$

Let ν be the expected frequency of invariable sites or invariants. Invariant are peculiar sites that do not sustain any mutation. Let $L(s|inv = 1)$ be the probability of site s given that it is an invariant. $L(s|inv = 0)$ is the probability of site s given that it is not an invariant. We have :

$$ln(L(s)) = ln(L(s|inv = 0) \times (1 - \nu) + L(s|inv = 1) \times \nu)$$

$L(s|inv = 0)$ is the logarithm of the right-hand side of equation 4. If no polymorphism is observed at site s and that the state at this site is h, the log likelihood is then :

$$ln(L(s)) = ln(L(s|inv = 0) \times (1 - \nu) + \pi_h \times \nu) \tag{4}$$

If polymorphism is observed, the site can not be an invariant. The log likelihood is therefore :

$$ln(L(s)) = ln(L(s|inv = 0) \times (1 - \nu) \tag{5}$$