

SUPPLEMENTAL MATERIAL 1

for the paper “Quality estimation of multiple sequence alignments by Bayesian hypothesis testing”

A. Tomovic and E. J. Oakeley

The hypothesis test (3 in the paper) can also be converted into a test for independence in the following way. We introduce a random variable Q_i which takes values $\{1,2\}$. When $Q_i=1$ it represents the distribution X_i and when $Q_i=2$ it represents Y . A second variable B_i takes its values from the DNA sequence alphabet $\{a,c,g,t\}$ as shown in Table 1.

Table 1. Relationships between random variables B_i and Q_i

B_i \ Q_i	$B_i=a$	$B_i=c$	$B_i=g$	$B_i=t$
$Q_i=1$	$X(a,i)$	$X(c,i)$	$X(g,i)$	$X(t,i)$
$Q_i=2$	$Y(a)$	$Y(c)$	$Y(g)$	$Y(t)$

From this, we can test the following hypotheses (equivalent to (3)):

H_0 : Q_i and B_i are independent random variables

H_1 : otherwise

When Q_i and B_i are independent, X_i and Y will not share the same distribution. For Bayesian hypothesis testing we can calculate the Bayes factor $BF_i(H_0; H_1)$ as follows:

$$BF_i(H_0; H_1) = \frac{P_i(B_i, Q_i | H_0)P_i(H_0)}{P_i(B_i, Q_i | H_1)P_i(H_1)} \quad (1)$$

Assuming that the *a priori* probabilities for the models (hypotheses) H_0 and H_1 are equal ($P_i(H_0) = P_i(H_1) = 0.5$), the Bayes factor will be:

$$BF_i(H_0; H_1) = \frac{P_i(B_i, Q_i | H_0)}{P_i(B_i, Q_i | H_1)} \quad (2)$$

Because B_i and Q_i are independent under the null hypothesis, the Bayes factor will be:

$$BF_i(H_0; H_1) = \frac{P_i(B_i | H_0)P_i(Q_i | H_0)}{P_i(B_i, Q_i | H_1)} \quad (3)$$

The index i implies that this Bayes factor is calculated for column i . Then using the fact that:

$$P_i(B_i | H_0) = \int_{\bar{p}} P(B_i, \bar{p} | H_0) \quad (4)$$

where \bar{p} is a vector of $[P(a), P(c), P(g), P(t)]$.

A conjugate prior for \bar{p} is the Dirichlet distribution (5):

$$P(\bar{p} | \alpha) \sim Dir(\alpha_a, \alpha_c, \alpha_g, \alpha_t) = \frac{\Gamma(\sum_b \alpha_b)}{\prod_b \Gamma(\alpha_b)} \prod_b P(b)^{\alpha_b - 1} \quad (5)$$

where $P(b) > 0$ and $\sum_b P(b) = 1$. Given a Dirichlet prior, the joint distribution of B_i and \bar{p} :

$$P(B_i, \bar{p} | \alpha) = \frac{\Gamma(\sum_b \alpha_b)}{\prod_b \Gamma(\alpha_b)} \prod_b P(b)^{B_i(b) + \alpha_b - 1}$$

(6)

where $B_i(b) = X(b, i) + Y(b)$, i.e. the column sum.

And the posterior is:

$$P(\bar{p} | B_i, \alpha) \sim \text{Dir}(B_i(b) + \alpha_b) \quad (7)$$

And finally we can calculate

$$P_i(B_i | H_0) = \int_{\bar{p}} P(B_i, \bar{p} | H_0) = \frac{\Gamma(\sum_b \alpha_b)}{\Gamma(s + \sum_b \alpha_b)} \prod_b \frac{\Gamma(B_i(b) + \alpha_b)}{\Gamma(\alpha_b)} \quad (8)$$

where $s = \sum_b [X(b, i) + Y(b)] = 2n$, i.e. the table sum.

And likewise for $P_i(Q_i | H_0)$:

$$P_i(Q_i | H_0) = \int_{\bar{p}} P(Q_i, \bar{p} | H_0) = \frac{\Gamma(\sum_q \alpha_q)}{\Gamma(s + \sum_q \alpha_q)} \prod_q \frac{\Gamma(Q_i(q) + \alpha_q)}{\Gamma(\alpha_q)} \quad (9)$$

where $Q_i(q)$ is a row sum i.e. $Q_i(q) = \sum_b X(b, i)$ for $q=1$ and $Q_i(q) = \sum_b Y(b)$ for $q=2$

Then we need to calculate $P_i(B_i, Q_i | H_1)$ and this is:

$$P_i(B_i, Q_i | H_1) = \int_{\hat{p}} P(B_i, Q_i, \hat{p} | H_1) \quad (10)$$

where \hat{p} is a vector of $(P(a,1), P(c,1), \dots, P(t,2))$.

A conjugate prior for \hat{p} is the Dirichlet distribution:

$$P(\hat{p} | \alpha) \sim \text{Dir}(\alpha_{a,1}, \alpha_{c,1}, \dots, \alpha_{t,2}) = \frac{\Gamma(\sum_{b,q} \alpha_{b,q})}{\prod_{b,q} \Gamma(\alpha_{b,q})} \prod_{b,q} P(b, q)^{\alpha_{b,q} - 1} \quad (11)$$

where $P(b,q) > 0$ and $\sum_{b,q} P(b, q) = 1$. Given a Dirichlet prior, the joint distribution of B_i, Q_i and \hat{p} :

$$P(B_i, Q_i, \hat{p} | \alpha) = \frac{\Gamma(\sum_{b,q} \alpha_{b,q})}{\prod_{b,q} \Gamma(\alpha_{b,q})} \prod_{b,q} P(b, q)^{T(b,q) + \alpha_{b,q} - 1} \quad (12)$$

where $T(b, q) = X(b, i)\delta(q=1) + Y(b)\delta(q=2)$ (δ is Kronecker's symbol).

And posterior is

$$P(\hat{p} | B_i, Q_i, \alpha) \sim \text{Dir}(T(b, q) + \alpha_{b,q}) \quad (13)$$

And finally we can calculate:

$$P_i(B_i, Q_i | H_1) = \int_{\hat{p}} P(B_i, Q_i, \hat{p} | H_1) = \frac{\Gamma(\sum_{b,q} \alpha_{b,q})}{\Gamma(s + \sum_{b,q} \alpha_{b,q})} \prod_{b,q} \frac{\Gamma(T(b, q) + \alpha_{b,q})}{\Gamma(\alpha_{b,q})} \quad (14)$$

If we define $\alpha_b = \sum_q \alpha_{b,q}$, $\alpha_q = \sum_b \alpha_{b,q}$ we are left with:

$$BF_i(H_0; H_1) = \frac{\Gamma(\sum_{b,q} \alpha_{b,q})}{\Gamma(s + \sum_{b,q} \alpha_{b,q})} \prod_b \frac{\Gamma(B(b) + \alpha_b)}{\Gamma(\alpha_b)} * \\ * \prod_q \frac{\Gamma(Q(q) + \alpha_q)}{\Gamma(\alpha_q)} \prod_{b,q} \frac{\Gamma(\alpha_{b,q})}{\Gamma(T(b,q) + \alpha_{b,q})} \quad (15)$$

After calculating $BF_i(H_0; H_1)$ for each $i=1, L$ (for each column), because we assume independence between columns, we can calculate:

$$BF = \prod_{i=1}^L BF_i(H_0, H_1) \quad (16)$$

In addition for the whole alignment in terms of B and Q (i.e. X and Y) we can calculate the posterior probability of the null model (hypothesis) in the following way:

$$P(H_0 | B, Q) = \frac{P(B, Q | H_0)P(H_0)}{P(B, Q | H_0)P(H_0) + P(B, Q | H_1)P(H_1)} \quad (17)$$

Using $P(H_0) = P(H_1) = 0.5$ and the fact that B and Q are independent under the null hypothesis:

$$P(H_0 | B, Q) = \frac{\prod_i P_i(B_i|H_0)P_i(Q_i|H_0)}{\prod_i P_i(B_i|H_0)P_i(Q_i|H_0) + \prod_i P_i(B_i, Q_i|H_1)} \quad (18)$$

Formula (18) can be calculated using (8),(9) and (14).

We can define all priors: $\alpha_{b,q} = 0.5$ as the so called Jeffreys' prior (the other, more conservative,

alternative would be uniform $\alpha_{b,q} = 1$. After comparing the results, we found that Jeffrey's prior gave the best results, when we evaluated the method against a series of different alignments (a set of random and biologically derived alignments) using correlation coefficient as a parameter of success.

The final scores (16) and (18) act as estimates of multiple sequence alignment significance. The alignment of multiple sequences is more significant when BF is small (much smaller than 1) and when $P(H_0 | B, Q)$ (i.e. $P(H_0 | X, Y)$) is small (smaller probability of the null model for the alignment, i.e. a smaller probability that alignment is random alignment). Jeffreys' scale (Jeffreys, 1961) of evidence for Bayes factors is given in Table 2.

Table 2. Jeffreys' scale for the interpretation of Bayes factors (BF) (Jeffreys, 1961).

Bayes Factor (BF) range	Evidence
$BF \geq 1$	Null hypothesis (model)* is supported
$0.3 \leq BF < 1$	Minimal evidence against null hypothesis (model)*
$0.1 \leq BF < 0.3$	Substantial evidence against null hypothesis (model)*
$0.01 \leq BF < 0.1$	Strong evidence against null hypothesis (model)*
$BF < 0.01$	Decisive evidence against null hypothesis (model)*

* Null model/hypothesis: alignment is random (not biological relevant).

Background frequencies can be specified by the user or estimated from the input sequence as the observed frequencies of each letter.