

Supplementary Material

1. Hidden Markov model

In the HMM model for operon prediction problem, all the genes in a genome a_1, a_2, \dots, a_n can be considered as a series of adjacent gene pairs along the chromosome b_1, b_2, \dots, b_{n-1} , where $b_i = (a_i, a_{i+1})$. Each gene pair b_i corresponds to a hidden state h_i which is not directly observed. h_i has only two possible values, 0 and 1. $h_i = 0$ implies that the two adjacent genes belong to the same operon, $h_i = 1$ implies that the two belong to different operon. Different h_i also corresponds to different distribution for the observed data “emitted” by the hidden state. Two types of data are considered in this study: differences in phylogeny conservation— Δ_i and intergenic distances— d_i . The hidden layer $h = (h_1, h_2, \dots, h_{n-1})$ is a Markov chain. The 2×2 transition matrix is denoted by $\tau = (\tau_{kl})$, where $\tau_{kl} = P(h_i = k \rightarrow h_{i+1} = l)$. For HMM applied in section 2 and 3, constant transition probability $\tau_{kl} = 0.5$ was used. Different transition probabilities were used in section 4. A graphical illustration of the hidden Markov model is presented in Figure S1.

2. HMM for phylogenic conservation.

In this case, the observed data are phylogenic barcode stem from a number of species considered. For each adjacent gene pair $b_i = (a_i, a_{i+1})$, we assume the differences in phylogeny conservation— Δ_i 's follow two different binomial distributions:

$$\Delta_i | h_i = 0 \sim \text{Binomial}(n_i, \theta_0),$$

$$\Delta_i | h_i = 1 \sim \text{Binomial}(n_i, \theta_1).$$

The number of trials in the binomial distribution— n_i is defined as the number of species that orthologous genes were found for at least one gene. θ_0 and θ_1 represent the probabilities of seeing barcode differences for intra-operonic gene pair and inter-operonic gene pair respectively. Let X denotes the observed phylogenic barcodes, the prior distribution for θ_0 and θ_1 is chosen to be Beta:

$$\theta_0, \theta_1 \sim \text{Beta}(\alpha, \beta).$$

Then the joint posterior distribution is the follows:

$$p(\theta, h | X) \propto p(X | h, \theta) p(h | \theta) p(\theta).$$

Gibbs sampler was used to iterate the following two steps:

1. Draw $h^{(t+1)} \sim P(h | X, \theta^{(t)})$;
2. Draw $\theta^{(t+1)} \sim P(\theta | X, h^{(t+1)})$.

In the first step which is the imputation step, a HMM path was drawn from its posterior distribution with fixed parameter values $\theta_0^{(t)}$ and $\theta_1^{(t)}$ (After this point the headnotes (t)

were suppressed). Next, we let $F_i(h) = P(A_i = h)$, since $P(A_0 = 0) = 1$, $F_0(h) = \begin{cases} 1 & h = 0 \\ 0 & h = 1 \end{cases}$,

and the rest of the path was computed recursively:

$$F_i(h) = \sum_{h_{i-1}=0}^1 \left\{ F_{i-1}(h_{i-1}) \tau_{h_{i-1}h} P(X(i) | \theta_h) \right\}.$$

After all $F_i(h)$ were calculated, we used backward sampling to recursively sample path h . We drew h_n from the following:

$$P(h_n = a | X, \theta_1, \theta_2) = \frac{F_n(a)}{F_n(0) + F_n(1)}$$

Then drew h_i recursively backward from the distribution

$$P(h_i = a | h_{i+1}, X, \theta_1, \theta_2) = \frac{p_{ah_{i+1}} F_i(a)}{p_{0h_{i+1}} F_i(0) + p_{1h_{i+1}} F_i(1)}.$$

In the second step which is the posterior sampling step, all the predicted operon pairs and all the predicted non-operon pairs were put together separately, and the marginal posterior distributions for θ_1 and θ_2 were:

$$\theta_0 \sim \text{Beta}\left(\sum_{h_j^{(t+1)}=0} \Delta_j + \alpha, \sum_{h_j^{(t+1)}=0} (n_j - \Delta_j) + \beta\right),$$

$$\theta_1 \sim \text{Beta}\left(\sum_{h_j^{(t+1)}=1} \Delta_j + \alpha, \sum_{h_j^{(t+1)}=1} (n_j - \Delta_j) + \beta\right).$$

Initial values for these two parameters were taken to be $\theta_0 = 0.2$ and $\theta_1 = 0.5$.

Parameters for the prior distributions are set to be $\alpha = \beta = 1$. By monitoring the joint likelihood, it was clear that the Gibbs sampler procedure converged rather rapidly; hence iterations were only performed 120 times, with the first 20 treated as burn-in and discarded. Probabilities were calculated as the average path assignment in the 100 recorded iterations.

Note that it is believed that two adjacent genes located on different strands do not belong to the same operon. Hence in our model, we automatically assign adjacent gene pairs that are not on the same strands as belonging to different operon.

3. HMM for intergenic distance.

In an alternative model, phylogenic conservation is replaced by intergenic distance-- d_i , which is defined as number of bases that separating adjacent genes. It is widely accepted that intergenic distance between intra-operonic gene pair is much shorter than that of an inter-operonic gene pair. We use two Gamma distributions to model the two types of intergenic distance. It has been observed that sometimes the DNA sequence of two adjacent genes overlapped with each other, which results in negative intergenic distances. To reduce the incidence of negative intergenic distance, 10 bps were added to the intergenic distances of all gene pairs. All gene pairs with intergenic distances less than -10 bps were automatically classified as inter-operonic. From now on, the new intergenic distances with 10 bps added were treated as the intergenic distances d_i .

$$d_i | h_i = 0 \sim \text{Gamma}(\gamma_0, \theta_0),$$

$$d_i | h_i = 1 \sim \text{Gamma}(\gamma_1, \theta_1).$$

To avoid unidentifiability problem, initial values were chosen such that the mean and variance for the first Gamma distribution is smaller than those of the second Gamma

distribution. Noninformative prior is chosen for the parameters in the two Gamma distributions.

4. Inhomogeneous HMM that combines both intergenic distance and phylogenetic conservation.

In regular Hidden Markov model such as the two we described before, the transition probabilities from one state to another are assumed to be fixed. In the current model, we extended this by allowing the transition probabilities to be depended on the properties of each state. To be specific, we treated the transition probabilities as a function of the intergenic distances between neighboring genes. From past experience (e.g., the inferred empirical distributions displayed in De Hoon 2004), we use two different Gamma distributions to model the two types of intergenic distances, note that 10 bp were added to every intergenic distances to reduce the incidence of negative cases.

$$P(d_i | h_i = 0) \sim \text{Gamma}(2, 15)$$

$$P(d_i | h_i = 1) \sim \text{Gamma}(5, 50)$$

The two Gamma distributions have mean 30 and 250, standard deviation 21 and 112 respectively, which closely matched the mean and variance of the inferred empirical distributions shown in Figure 3 of De Hoon 2004. Figure S2 illustrated the two Gamma distributions.

Using Bayes' theorem, we can write down the transition probabilities for a pair of adjacent genes with intergenic distance of d bps as:

$$P(h_i = 1 | d_i = d) = \frac{p(h_i = 1)p(d | h_i = 1)}{p(h_i = 1)p(d | h_i = 1) + p(h_i = 0)p(d | h_i = 0)} = \frac{1}{1 + p(d | h_i = 0) / p(d | h_i = 1)}$$

$$P(h_i = 0 | d_i = d) = 1 - P(h_i = 1 | d_i = d).$$

Since Gamma distribution is continuous, probability $P(d | h_i = k)$, $k = 0, 1$ was taken to be the difference of the cumulative density function evaluated at d and $d + 1$:

$$p(d_i = d | h_i = k) = \int_d^{d+1} \frac{1}{\theta_i^{\gamma_i} \Gamma(\gamma_i)} x^{\gamma_i-1} e^{-x/\theta_i} dx,$$

$$\gamma_1 = 2, \quad \gamma_2 = 5,$$

$$\theta_1 = 15, \quad \theta_2 = 50.$$

For adjacent gene pairs that still having negative intergenic distances after 10 bps were added, we define:

$$p(h_i = k | d_i = d) = p(h_i = k | d_i = 0) \quad \text{when } d < 0, \quad i = 1, 2.$$

It can be think of the intergenic distances (with the addition of 10 bps) were truncated at 0.

Finally since we do not have much *a priori* information as how likely a adjacent gene pair is an operon pair or a non-operon pair, we set equal probability to $p(h_i = 0)$ and $p(h_i = 1)$:

$$p(h_i = 0) = p(h_i = 1) = 0.5.$$

Figure S1. Schematic illustration of the HMM used in this study.

Figure S2. Probability density function of two Gamma distributions used to model the inter-operonic and intra-operonic integenic distances: Gamma (2, 15) and Gamma (5, 10).