HMM Weighting - From Bitscore to Probability

Chengze Shen

May 26, 2021

1 Notations

The following notations are extended from Min's HMM Weighting/Probability formulation.

Σ: The alphabet for sequences (e.g., $\Sigma = \{A, T, C, G\}$ for DNA, $|\Sigma| = 4$).

$\mathcal{H}$: The ensemble of HMMs. There are two major scenarios: 1) $\mathcal{H}$ has only disjoint HMMs (no overlapping sequences among HMMs); 2) $\mathcal{H}$ has overlapping HMMs (the UPP decomposition). In addition, we have $|\mathcal{H}| = h$.

H_i: The i-th HMM in $\mathcal{H}$.

$BS(H_i, q)$: The bitscore of a query sequence q on HMM H_i. $BS(H_i, q) = \log_2 \frac{P(q|H_i)}{P(q|H_0)}$ where H_0 is a null/random model.

w_i: The weighting of HMM $H_i \in \mathcal{H}$. In addition, we have $\sum_{i=1}^{h} w_i = 1$.

2 Disjoint HMMs Weighting

As written by Min, the weighting of HMMs from an ensemble of disjoint HMMs can be derived from their corresponding bitscores,

$$w_i = P(H_i|q) = \frac{1}{\sum_{j=1}^{h} 2^{BS(H_j,q)-BS(H_i,q)}} \tag{1}$$

We can easily prove that summing (1) over all HMMs yields 1. Also in this scenario, we assume our priors are the same ($P(H_i) = \frac{1}{h}$). That is: all HMMs are equal and the size of an HMM does not change how likely it is selected.

3 Disjoint HMMs Weighting - Different Sizes

When we have disjoint HMMs and their sizes may differ (i.e., HMMs may contain different numbers of sequences), we may need to relax our assumption on equally likely HMMs. For example, two HMMs have the same bitscore, but one is of size 1000 and one is of size 10. Then, the larger HMM should naturally be weighted more for the query sequence.

Assume we have S sequences in total, and each HMM H_i contains s_i sequences. As written by Min, the prior $P(H_i)$ can be expressed as the fraction of s_i over S,

$$P(H_i) = \frac{s_i}{S} \tag{2}$$

(2) guarantees that $P(\mathcal{H}) = \sum P(H_i) = 1$. Then, the weighting $P(H_i|q)$ can be expressed as,

$$\begin{aligned} P(H_i|q) &= \frac{P(q|H_i) \cdot P(H_i)}{P(q)} \\ &= \frac{P(q|H_i) \cdot P(H_i)}{\sum_{j=1}^{h} P(q|H_j) \cdot P(H_j)} \\ &= \frac{P(q|H_i)\frac{s_i}{S}}{\sum_{j=1}^{h} P(q|H_j)\frac{s_j}{S}} \\ &= \frac{P(q|H_i)s_i}{\sum_{j=1}^{h} P(q|H_j)s_j} \end{aligned} \tag{3}$$

The final expression from Min's formulation is,

$$w_i = P(H_i|q) = \frac{1}{\sum_{j=1}^{h} 2^{BS(H_j,q) - BS(H_i,q) + \log_2 \frac{s_j}{s_i}}} \tag{4}$$

We can show that (4) sums to 1 over all HMMs,

$$\begin{aligned}\sum_{i=1}^{h} w_i = \sum_{i=1}^{h} P(H_i|q) &= \sum_{i=1}^{h} \frac{P(q|H_i) \cdot P(H_i)}{\sum_{j=1}^{h} P(q|H_j) \cdot P(H_j)} \\ &= \frac{\sum_{i=1}^{h} P(q|H_i) \cdot P(H_i)}{\sum_{j=1}^{h} P(q|H_j) \cdot P(H_j)} \\ &= 1\end{aligned}$$

4 Overlapping HMMs Weighting - Different Sizes

It turns out that the difference between disjoint and overlapping HMMs is small. In the case of disjoint HMMs, our prior $P(H_i) = \frac{s_i}{S}$, where S is the total number of sequences that present in the HMMs (S sequences are all unique). In the case of overlapping HMMs, the priors are $P(H_i) = \frac{s_i}{S^*}$, where S^* is still $\sum_{i=1}^{h} s_i$, but we may have counted each unique sequence more than once. Therefore, the weighting expression for an HMM H_i given query q is left **unchanged** as defined in Section 3.

A simple example is shown below, where we have 3 nodes/HMMs. The root node has 100 sequences, and the subsequent children nodes have 10 and 90 sequences, respectively.

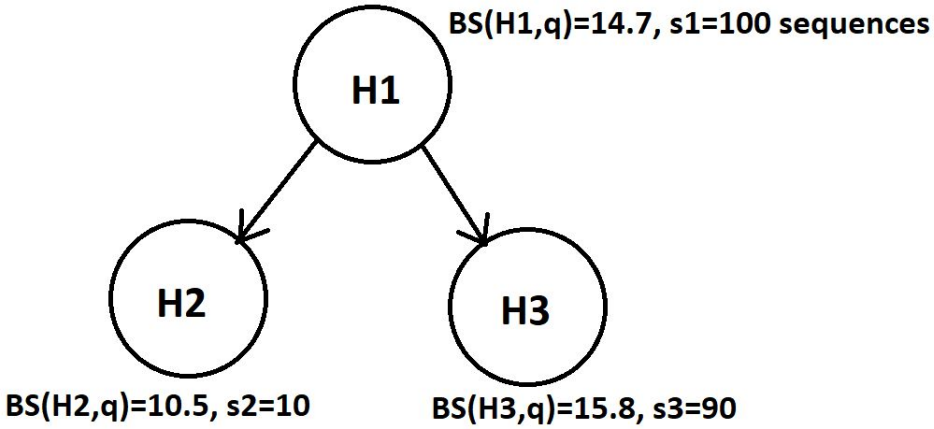

Figure 1: A simple example of overlapping HMMs weighting, where HMMs have different sizes.

Using (4) we calculate the weights for each of the HMMs:

$$w_1 = \frac{1}{1 + 2^{14.7-15.8+\log_2 \frac{10}{9}} + 2^{10.5-15.8+\log_2 \frac{1}{9}}} = 0.65739$$
$$w_2 = \frac{1}{1 + 2^{14.7-10.5+\log_2 10} + 2^{15.8-10.5+\log_2 9}} = 0.00185$$
$$w_3 = \frac{1}{1 + 2^{15.8-14.7+\log_2 0.9} + 2^{10.5-14.7+\log_2 0.1}} = 0.34076$$

and we have $w_1 + w_2 + w_3 = 1$.

5 Conclusion

I will be using (4) to calculate the weightings for HMMs.

First, I am going to examine how the weightings will be distributed. If the distribution is top-heavy (e.g., the first 10 HMMs in the ranking consist of 99% of the weights), then the top HMMs (with weighting) should have sufficient information to construct accurate backbone alignments for the GCM pipeline.

Secondly, based on the observation above, I am going to modify MAGUS so that when it reads in backbone alignments and register the edges in the alignment graph, it also considers the weight for each backbone alignment.