Bitscores

Minhyuk Park

April 2021

1 Notations

Some notations used in the formulas:

Σ: Σ is the alphabet where $|\Sigma| = n$ ($n = 4$ for DNA, RNA and $n = 20$ for Amino Acids)

p_m: p_m are the match probabilities where $p_m[x][y]$ is the match state emission probability at state x of character y. There are n emission probabilities.

p_i: p_i are the insertion probabilities where $p_i[x][y]$ is the match state emission probability at state x of character y. There are n emission probabilities.

$t_{a \to b}$: 7 transition probabilities ($m \to m, m \to i, m \to d, i \to m, i \to i, d \to m, d \to d$) where m is match, d is deletion, and i is insertion states) where $t_{a \to b}[][y]$ is the .

$BS(H, q)$: Bitscore of query sequence q on the hmm H

2 Assigning Probability Weights to Two HMMs Given a Query Sequence Using Their Bitscores

Say our query sequence is q, and the two input hmms are H_1 and H_2.We can get the weights of each HMMs by doing the following.

$$\begin{aligned}
weight_{H_1} &= \frac{1}{2^{BS(H_1 q)-BS(H_1,q)} + 2^{BS(H_2,q)-BS(H_1,q)}} \\
&= \frac{1}{1 + 2^{BS(H_2,q)-BS(H_1,q)}} \\
weight_{H_2} &= \frac{1}{2^{BS(H_1 q)-BS(H_2,q)} + 2^{BS(H_2,q)-BS(H_2,q)}} \\
&= \frac{1}{2^{BS(H_1,q)-BS(H_2,q)} + 1}
\end{aligned}$$

Let's use an example with real bitscores taken from outputs of HMMSearch before we do the proof. Suppose $BS(H_1, q) = 716.3$ and $BS(H_2, q) = 721.5$, then

$$\begin{aligned}
weight_{H_1} &= \frac{1}{2^{716.5-716.5} + 2^{721.5-716.5}} \\
&= \frac{1}{1 + 2^5} \\
&= \frac{1}{33} \\
weight_{H_2} &= \frac{1}{2^{716.5-721.5} + 2^{721.5-721.5}} \\
&= \frac{1}{2^{-5} + 1} \\
&= \frac{32}{33} \\
weight_{H_1} + weight_{H_2} &= 1
\end{aligned}$$

$$\begin{aligned}
weight_{H_1} + weight_{H_2} &= \frac{1}{1+2^{BS(H_2,q)-BS(H_1,q)}} + \frac{1}{2^{BS(H_1,q)-BS(H_2,q)}+1} \\
&= \frac{1}{1+2^{BS(H_2,q)-BS(H_1,q)}} + \frac{1}{\frac{1}{2^{BS(H_2,q)-BS(H_1,q)}}+1} \\
&= \frac{1}{1+2^{BS(H_2,q)-BS(H_1,q)}} + \frac{1}{\frac{1+2^{BS(H_2,q)-BS(H_1,q)}}{2^{BS(H_2,q)-BS(H_1,q)}}} \\
&= \frac{1}{1+2^{BS(H_2,q)-BS(H_1,q)}} + \frac{2^{BS(H_2,q)-BS(H_1,q)}}{1+2^{BS(H_2,q)-BS(H_1,q)}} \\
&= \frac{1+2^{BS(H_2,q)-BS(H_1,q)}}{1+2^{BS(H_2,q)-BS(H_1,q)}} \\
&= 1
\end{aligned}$$

Now we know how do weight two HMMs given a query sequence using the bitscores obtained from HMMSearch!

3 Bitscore to Probability When HMMs are Different Sizes

The bitscore of a query sequence given a HMMER HMM is $\log_2 \frac{P(q|H)}{P(q|H_0)}$ where H is the HMM, q is the query sequence, and H_0 is the null model, or the random model.

Using Bayes' theroem, we arrive at the probability of H_i generating sequence q as follows.

$$\begin{aligned}
P(H_i|q) &= \frac{P(q|H_i)\cdot P(H_i)}{P(q)} && (1) \\
&= \frac{P(q|H_i)\cdot P(H_i)}{\Sigma_{j=1}^{n} P(q|H_j)\cdot P(H_j)} && (2)
\end{aligned}$$

where n is the number of HMMs ($H_i ... H_n$).

If we assume that the more sequences the HMM is trained on, the more likely the HMM is to output a sequence, then we can transform the above into the following.

$$\begin{aligned}
P(H_i|q) &= \frac{P(q|H_i)\cdot \frac{s_i}{S}}{\Sigma_{j=1}^{n} P(q|H_j)\cdot \frac{s_j}{S}} && (3) \\
&= \frac{1}{\Sigma_{j=1}^{n} \frac{P(q|H_j)\cdot s_j}{P(q|H_i)\cdot s_i}} && (4) \\
&= \frac{1}{\Sigma_{j=1}^{n} 2^{\log_2 \frac{P(q|H_j)\cdot s_j}{P(q|H_i)\cdot s_i}}} && (5)
\end{aligned}$$

where s_i is the number of sequences that HMM H_i was trained on and S is the total number of sequences that the HMMs were trained on.

From the definition of Bitscores, we can derive the following.

$$BS(H_j) - BS(H_i) = \log_2 \frac{P(q|H_j)}{P(q|H_0)} - \log_2 \frac{P(q|H_i)}{P(q|H_0)} \tag{6}$$

$$= \log_2 \frac{P(q|H_j)}{P(q|H_i)} \tag{7}$$

So

$$P(H_i|q) = \frac{1}{\Sigma_{j=1}^{n} 2^{\log_2 \frac{P(q|H_j)\cdot s_j}{P(q|H_i)\cdot s_i}}} \tag{8}$$

$$= \frac{1}{\Sigma_{j=1}^{n} 2^{BS(H_j)-BS(H_i)+\log_2 \frac{s_j}{s_i}}} \tag{9}$$

(10)