

1 Profiles

We first looked at profiles derived from alignment of a set of sequences, where the alignment does not have any spaces. In that case, there is a natural alignment of any new sequence (of the same length) to the profile: the character in position j of the sequence is aligned with column j of the profile. But how do we score that alignment? Let $S(x, j)$ be the score of a character x to the set of characters in column j . The simplest way to assign $S(x, j)$ is to sum the score of aligning x to each character found in column j .

Then, the score of an alignment of a sequence to a profile of length n is

$$\sum_{j=1}^{j=n} S(x, j),$$

where x is the character of the sequence aligned to column j .

Often $S(x, j)$ is chosen to be $\log_2[p(x, j)/p(x)]$, where $p(x, j)$ is the fraction of characters in column j that are character x , and $p(x)$ is the fraction of all characters in the sequences (or in some larger database) that are x . How do we explain this choice of $S(x, j)$?

First, why the ratio $p(x, j)/p(x)$? Let's look at the two extreme cases. Suppose character x is very common in column j , i.e. $p(x, j)$ is large. If x is very rare in the set of sequences, then the fact that it is common in column j seems a more significant statement about column j , than if x is a very common character in the sequences. Some biological or chemical force is encouraging character x to appear in that position, even though character x is rare in the sequences overall. Conversely, suppose $p(x, j)$ is small, but $p(x)$ is large. That case also seems a significant statement about column j (and the associated molecular position) - there is some biological or chemical force that is excluding x from column j , even though x is common in the sequences overall. If we only looked at $p(x, j)$ and did not relate it to $p(x)$, we would lose these important distinctions. If we only used $p(x, j)$ to characterize the occurrence of x in column j , we would not distinguish between the cases when $p(x)$ is large or small. By taking the ratio of $p(x, j)$ to $p(x)$, we are able to increase or decrease the importance we attach to character x in column j .

Next, why take the log of the ratio? There are several ways to motivate this. The simplest is to notice that when $p(x, j) < p(x)$ then $p(x, j)/p(x) < 1$ so $\log[p(x, j)/p(x)] < 0$, and when $p(x, j) > p(x)$ then $p(x, j)/p(x) > 1$ so

$\log[p(x, j)/p(x)] > 0$. This is true no matter what the base of the logarithm is. So we get a positive $S(x, j)$ when x is more common in column j than it is in the sequences overall, and a negative $S(x, j)$ when it is less common, and a 0 when x is as common in column j as it is overall. That is a sensible, clean way to assign scores to characters in columns, and one simple way to justify taking the log.

Another explanation is that the log nicely displays how many orders of magnitude different are $p(x, j)$ and $p(x)$. For example, if the log is base 10, and $p(x, j)$ is 100 times bigger than $p(x)$ (i.e. 10^2 times larger), then the log of the ratio is 2, so $p(x, j)$ is two orders of magnitude larger than $p(x)$. This makes it easy to specify how much larger we want $p(x, j)$ to be than $p(x)$ in order to conclude that the score is significant.

A deeper explanation of the use of the log comes from a simple Markov Model viewpoint of how the sequences are generated. In the Markov Model view, we imagine that the aligned set of sequences we have in hand are only a sample from a larger set of unknown sequences, of roughly the same lengths. We multiply align the sample sequences in order to estimate the frequency of each character in each position in the sequences (corresponding to columns in the multiple alignment). In the Markov Model, the frequencies in any one position are assumed to be independent from the frequencies in any other position. This is the most vital (and questionable) assumption in Markov Models. In this simple approach, all we know now about the unknown large set of sequences are the estimated frequencies of different characters at each position. That is our model of the set. Any sequence might be in the set, but with quite varying frequencies. Then, given a specific sequence S_1 , we are interested in the probability S_1 would be picked if we drew a sequence at random from that set. Equivalently, we imagine generating a sequence by randomly picking a character for each position i , independent of the character choice for any other position. And, when picking the character for position i , we pick with probabilities equal to the observed character frequencies in column i . Then, we are interested in the probability that sequence S_1 is generated in this way.

So given the frequencies from the multiple alignment (which we view as probabilities) we can examine a query sequence S_1 not in the sample, and ask what is the probability that it would be the randomly generated sequence? The higher the probability, the more S_1 seems consistent with the profile?

For simplicity now, let us assume that S_1 has length equal to the length

of the multiple alignment, i.e. it has as many characters as the alignment has columns. Let n denote that length. By the assumption of independence at each position, the probability that S_1 is generated as defined above, is calculated by taking the product of the probability of each character in the sequence, in its respective position. Introducing some notation, let $S_1(i)$ be the i 'th character of S_1 . Then the probability that character $S_1(i)$ occurs in position i in the larger set of sequences is estimated to be $p(S_1(i), i)$. Hence the probability that S_1 is generated in the random sequence generation is

$$\prod_{i=1}^n p(S_1(i), i).$$

(Recall that \prod means taking the product of a set of numbers, and is the multiplication analog of Σ .)

Now that probability by itself is not so informative. Instead, we want to compare it to the probability that S_1 is a randomly generated using a distribution that does not distinguish different columns. The probability that S_1 is a random sequence is taken as

$$\prod_{i=1}^n p(S_1(i)).$$

Then we compare the two probabilities by taking the ratio

$$\frac{\prod_{i=1}^n p(S_1(i), i)}{\prod_{i=1}^n p(S_1(i))}.$$

But that is also

$$\prod_{i=1}^n \frac{p(S_1(i), i)}{p(S_1(i))}.$$

Now here is where the log comes in. First, the log function is a monotonically increasing function, so the relative order between two numbers, or two ratios, is preserved if we take the log of those two numbers. Second, multiplication is somewhat difficult (and was more so in the days before computers), but $\log(p \times q) = \log(p) + \log(q)$, so

$$\log \prod_{i=1}^n \frac{p(S_1(i), i)}{p(S_1(i))} = \sum_{i=1}^n \log \frac{p(S_1(i), i)}{p(S_1(i))}.$$

Hence if we are willing to take the log of the ratio we really are interested in, we have an addition problem, rather than a multiplication problem. And

since \log is monotonic, the order of the ratios is the same after taking the \log , as it was before taking the \log . So it seems worthwhile to simplify the computational problem by taking logs.

Now note that $\log \frac{p(S_1(i), i)}{p(S_1(i))}$ is just $S(S_1(i), i)$, and it is just this form for S that we have been trying to explain. So that is our final explanation for setting $S(x, j)$ to $\log_2[p(x, j)/p(x)]$.

There is actually a much deeper explanation for setting $S(x, j)$ to $\log_2[p(x, j)/p(x)]$, that comes from *Information Theory*. We may touch on this later when we talk more about (Hidden) Markov Model explicitly.