

# Profiles

---

Profiles assigns an average of standard substitution scores from all the residues seen in the corresponding column.

Seq1 . . . V G A - - H A G E Y

Seq2 . . . V E A - - D V A G H

Seq3 . . . F N A - - N I P H K

Seq4 . . . I A G A D N G A G V

Aligning the character **a** to the first column will give an average score:

$$\frac{2}{4} \text{score}(V, a) + \frac{1}{4} \text{score}(F, a) + \frac{1}{4} \text{score}(I, a).$$

To complete a *scoring scheme*, we need to define the gap penalties.

## Position Specific Scoring Matrix(PSSM)

---

Let's assume that we have a model  $M$  for a set of seqs and a seq  $x$  of length  $L$ . Then, the probability of  $x$  given  $M$  can be written as

$$P(x|M) = \prod_{i=1}^L p(x_i)$$

However,  $P(x|M)$  is length dependent and we also want compare with the random model based on the background prob.

$$Score(x) = \sum_{i=1}^L \log \frac{p(x_i)}{q_{x_i}}$$

where  $q_{x_i}$  denotes a random probability based on the background prob.

*Why is it called a scoring matrix?*

To answer this question, you need to understand how a scoring matrix like BLOSUM 62 is constructed. (we will come back to this topic later.)