Distinguishing counts by position along the oligopeptide

We use the asterisk sign for a “wildcard” residue: For each protein sequence s and amino-acid A, let $f_o(A*,s)$ be the number of times A is observed as the left (first) residue within a dipeptide, and let $f_o(*A,s)$ be the number of times A is observed as the right (second) residue within a dipeptide. $f_o(A*,s) = f_o(*A,s) = f_o(A,s)$ unless s begins ($f_o(*A,s) = f_o(A,s) - 1$) or terminates ($f_o(A*,s) = f_o(A,s) - 1$) with A. The quantities $f_o(A**,s)$, $f_o(*A*,s)$, $f_o(**A,s)$, $f_o(AB*,s)$ and $f_o(*AB,s)$ are analogously defined. Equations (2) and (3) are defined also for oligopeptides with wildcards.

Dipeptide expectations

We rewrite Equation (4), taking into account edge effects:

$$\begin{aligned} f_e(AB,s) &= \sum_{pair\ s[i]s[i+1]} \mathrm{Prob}(s[i]=A, s[i+1]=B) \\ &= (N_2(s)-2)\cdot\frac{\#\ possible\ mid\text{-}sequence\ AB\ pairs}{\#\ possible\ mid\text{-}sequence\ pairs} \\ &+ \frac{\#\ possible\ leading\ AB\ pairs}{\#\ possible\ leading\ pairs} + \frac{\#\ possible\ trailing\ AB\ pairs}{\#\ possible\ trailing\ pairs} \end{aligned} \tag{4'}$$

For amino-acids A and B, such as neither A, nor B is the leading/trailing amino acid of the sequence s, we use formulae similar to Equations (5), (6):

$$\begin{aligned} &\#\ possible\ mid\text{-}sequence\ AA\ pairs = f_o(A,s)(f_o(A,s)-1) \\ &\#\ possible\ leading\ AA\ pairs = 0 \\ &\#\ possible\ trailing\ AA\ pairs = 0 \end{aligned} \tag{5'}$$

$$\begin{aligned} &\#\ possible\ mid\text{-}sequence\ AB\ pairs = f_o(A,s)f_o(B,s) \\ &\#\ possible\ leading\ AB\ pairs = 0 \\ &\#\ possible\ trailing\ AB\ pairs = 0 \end{aligned} \tag{6'}$$

Otherwise, if A is the leading amino acid of the sequence s, but B is not the trailing one, we use formulae:

$$\begin{aligned} &\#\ possible\ mid\text{-}sequence\ AB\ pairs = (f_o(A,s)-1)f_o(B,s) \\ &\#\ possible\ leading\ AB\ pairs = f_o(B,s) \\ &\#\ possible\ trailing\ AB\ pairs = 0 \end{aligned} \tag{6''}$$

Otherwise, if B is the trailing amino acid of the sequence s, but A is not the leading one, we use formulae:

$$\#\,possible\ mid\text{-}sequence\ AB\ pairs = f_o(A,s)\left(f_o(B,s) - 1\right)$$
$$\#\,possible\ leading\ AB\ pairs = 0 \qquad (6''')$$
$$\#\,possible\ trailing\ AB\ pairs = f_o(A,s)$$

Otherwise, if both *A* is the leading amino acid of the sequence *s*, and *B* is the trailing one, we use formulae:

$$\#\,possible\ mid\text{-}sequence\ AA\ pairs = \left(f_o(A,s) - 2\right)\left(f_o(A,s) - 3\right)$$
$$\#\,possible\ leading\ AA\ pairs = f_o(A,s) - 2 \qquad (5'''')$$
$$\#\,possible\ trailing\ AA\ pairs = f_o(A,s) - 2$$

$$\#\,possible\ mid\text{-}sequence\ AB\ pairs = \left(f_o(A,s) - 1\right)\left(f_o(B,s) - 1\right)$$
$$\#\,possible\ leading\ AB\ pairs = f_o(B,s) - 1 \qquad (6'''')$$
$$\#\,possible\ trailing\ AB\ pairs = f_o(A,s) - 1$$

Equations (6')…(6'''') can be summarized by

$$\#\,possible\ mid\text{-}sequence\ AB\ pairs = f_o(*A,s)f_o(B*,s)$$
$$\#\,possible\ leading\ AB\ pairs = \left(f_o(A,s) - f_o(*A,s)\right)f_o(B*,s) \qquad (6*)$$
$$\#\,possible\ trailing\ AB\ pairs = \left(f_o(B,s) - f_o(B*,s)\right)f_o(*A,s)$$

The total number of pairs is obtained by summation over all *A*,*B*.

Expected tripeptide counts per-sequence

In the following sub-section, all formulae refer to a protein sequence *s*, which is omitted from notation.

For *ABC* tripeptides, we rely on the number of right/left/mid-sequence runs (of length 1 or more) of an amino acid *B*:

$$R_L(B) = f_o(*B*) - f_o(BB*)$$
$$R_R(B) = f_o(*B*) - f_o(*BB)$$
$$R_M(B) = R_L(B) + R_R(B) - R(B)$$

$R_L(B)$, $R_R(B)$, $R_M(B)$ and $R(B)$ differ only if *B* is the leading/trailing residue. Denote the number of mid-sequence *BB* occurances $f_M(BB) = f_o(BB*) + f_o(*BB) - f_o(BB)$.

Let $U_M(B)$ be the number of mid-sequence singleton *B*-s (*B* runs consisting of a single residue, which is not leading nor trailing). If $f_o(*B*)=1$, then also $U_M(B)=1$. Otherwise, $U_M(B) = f_M(*B*) - f_M(BB)$ is a random variable. Let $E(U_M(B))$ be the expectation of this random variable. Let $U_1(B), U_2(B), \ldots, U_{R_M}(B)$ be the binary random variables that are 1 if the respective run is a singleton.

$$\begin{aligned} E\left(U_M(B)\right) &= \sum_{i=1}^{R_M(B)} E\left(U_i(B)\right) \\ &= \sum_{i=1}^{R_M(B)} \mathrm{Prob}\left(U_i(B)=1\right) \\ &= \sum_{i=1}^{R_M(B)} \mathrm{Prob}\left(\text{the } i^{\text{th}} \text{ run is a singleton}\right) \\ &= R_M(B)\cdot \mathrm{Prob}\left(\text{a specific run is a singleton}\right) \end{aligned} \qquad (7')$$

The process of choosing a random sequence that preserves the $f_o(*B*)$, $f_o(BB*)$, $f_o(*BB)$ and $f_M(BB)$ counts involves partitioning the $f_M(BB)$ mid-sequence BB dimers into the $R_M(B)$ runs.

The number of such partitions is

$\binom{R_M(B)+f_M(BB)-1}{R_M(B)-1}$, out of which $\binom{R_M(B)+f_M(BB)-2}{R_M(B)-2}$ partitions have a specific singleton run. The probability of a specific singleton run is therefore

$$\frac{\binom{R_M(B)+f_M(BB)-2}{R_M(B)-2}}{\binom{R_M(B)+f_M(BB)-1}{R_M(B)-1}} = \frac{R_M(B)-1}{R_M(B)+f_M(BB)-1} = \frac{R_M(B)-1}{f_M(B)-1} \qquad (8')$$

For each residue B with ($f_M(B)>1$) the expected number of singletons is therefore implied by equations (7') and (8'):

$$E\left(U_M(B)\right) = \frac{R_M(B)\left(R_M(B)-1\right)}{f_M(B)-1} \qquad (9')$$

Analogously to the distinction between homodipeptides and heterodipeptides, we now need to distinguish several cases, as follows:

1. For each heterotripeptide ABC ($A\neq B$ and $C\neq B$) the expected count $f_e(ABC)$ is the product of $E(U_M(B))$ by the estimated probability of C following a run of B's and A preceding such a run:

$$f_e\left(ABC\right) = E\left(U_M(B)\right)\cdot\frac{f_o\left(AB*\right)}{f_o\left(*B*\right)-f_o\left(BB*\right)}\cdot\frac{f_o\left(*BC\right)}{f_o\left(*B*\right)-f_o\left(*BB\right)} \qquad (10')$$

2. For semi homotripeptides ABB or BBC, one needs to consider only positions that are beginnings or ends, respectively, of non-singleton runs. There are expected to be $R_L(B)$-$E(U_M(B))$, or, respectively $R_R(B)$-$E(U_M(B))$ such

positions. Thus, one needs to multiply this number by the probability of encountering the non-B residue:

$$
\begin{aligned}
f_e(ABB) &= \left(R_L \cdot E\left(U_M(B)\right)\right) \cdot \frac{f_o(AB)}{f_o(*B) \cdot f_o(BB)} \\
f_e(BBC) &= \left(R_R \cdot E\left(U_M(B)\right)\right) \cdot \frac{f_o(BC)}{f_o(B*) \cdot f_o(BB)}
\end{aligned}
\qquad (11')
$$

3. For homotripetides BBB, the raw count is a direct function of $U_M(B)$, $R_M(B)$, and $f_M(BB)$:

   $$f_o(BBB) = f_M(BB)\text{-}(R_M(B)\text{-}U_M(B)) = 2f_M(BB) - f_o(*B*)+U_M(B)$$

   Therefore:

   $$f_e(BBB) = 2f_M(BB) - f_o(*B*)+E(U_M(B)) \qquad (12')$$