

Consider,

a set of sequences $S = S_1 \dots S_k$, where $S_i = s_{i1} \dots s_{in}$, s_{ij} is the j -th amino acid in sequence S_i .The probability of observing each amino acid $X = p(X)$

$$p(X) = \sum_{i=1}^k c(X_j, S_i) / \sum_{j=1}^{20} \sum_{i=1}^k c(X_j, S_i)$$

where $c(X_j, S_i)$ is the count of amino acid X_j in sequence S_i

$$P(X_l, X_m | \text{random}) = 2p(X_l) \cdot p(X_m) \text{ or } p(X_l)^2, \text{ if } l = m$$

A factor of 2 because a given set of $\{X_l, X_m\}$ can be selected from the alignments as (X_l, X_m) or (X_m, X_l) as long as $l \neq m$.

 $P(X_l, X_m | \text{data}) = \# \text{ of } X_l, X_m \text{ combinations} / \text{all possible pairwise combinations}$

finally take the log2-odds likelihood ratio and multiply by 2 for easy representation

Note: all possible pairwise combinations = $k \cdot \{n(n-1)/2\}$, since $k =$

Number of columns in the alignment and,

$$\frac{n(n-1)}{2} = \text{Number of different pairs in a given column}$$

Now to accommodate for substitutions we do not observe, we add a pseudo count to each possible substitution pair

we have $\frac{20(20-1)}{2} + 20$ such pairs

$$= 210$$

Coming from heterologous pairs like AT, AC, ... etc.

Coming from pairs like AA, TT, YY -- etc

add 1 in numerator, add 210 (20*19+20) in denominator

$$\text{ie } P(X_i, X_m | \text{data}) = \frac{\text{count}(X_i, X_m) + 1}{k \cdot \{n(n-1)/2\} + 210}$$

$\lambda \cdot \log_2 \frac{P(X_i, X_m | \text{data})}{P(X_i, X_m | \text{random})} = \text{subs.matrix}$

Handwritten notes: "Add 1 pseudo count" (pointing to the +1 in the numerator), "All pseudo counts" (pointing to the +210 in the denominator).

λ being a scaling factor for representation

Same ^{concept} as log-odds for PAM but to the base of 2.