Calculation of p-values for sub-HMM/PROSITE overlaps

Let $q_{ij}$ be the probability that a query fragment of length $F_j$ overlaps a PROSITE fragment of length $P_j$ on a protein of length $S_j$ in PFAM family $i$ by a fraction of at least $x$:

$$q_{ij} \leq \min \left(1, \frac{P_j + F_j - 2x \min (P_j, F_j) + 1}{S_j - F_j + 1}\right) \tag{1}$$

Then we want to compute the probability, $D_i$, that a certain number of overlaps occurs between a PFAM family $i$ and a PROSITE family. In particular, given that at least 50% of the members of either family lie in the intersection, we want the probability that 95% of the sequences in the intersection have an overlapping fragment.

Let $F$ be a PFAM family and $P$ be a PROSITE family. We define $R$ as the set of all subsets of $F \cap P$ which contain at least 95% of the intersection:

$$R = \{ R | R \subseteq F \cap P \land |R| \geq 0.95n \} \tag{2}$$

where $n = |F \cap P|$. Let $p_{ij} = \{ q_{ij} | j \in F \cap P \}$, then

$$D_i = \sum_{R \in R} \left( \prod_{j \in R} p_{ij} \prod_{j \in (F \cap P) \setminus R} (1 - p_{ij}) \right) \tag{3}$$

Since this would require enumerating every set in $R$, this would take too long to calculate, so we approximate it with an upper bound. Let $j^* = \arg\max_j p_{ij}$ and $R^* = \arg\min_{R' \in R} (|R'|) \forall R' \in R$. Then we have

$$D_i \leq \sum_{R \in R} \prod_{j \in R} p_{ij} \tag{4}$$

$$\leq \sum_{R \in R} |R|^{p_{ij}} \tag{5}$$

$$\leq |R|^{p_{ij}^{R^*}} \tag{6}$$

$$= \left( \sum_{k = [0.95n]}^{n} \binom{n}{k} \right) p_{ij}^{\lfloor 0.95n\rfloor} \tag{7}$$

This bound is often too loose in practice however. This is because for large values of $p_{ij}^*$, the last term in equation 3 makes that term very small, whereas the corresponding term in our bound would still be large. Therefore, we adopt a method of removing these large outliers to get a tighter bound.

$$D_i = \sum_{R \in R} \left( \prod_{j \in R} p_{ij} \prod_{j \in (F \cap P) \setminus R} (1 - p_{ij}) \right) \tag{8}$$

$$= \sum_{R \in R} \left( \prod_{j \in R} p_{ij} \prod_{j \in (F \cap P) \setminus R} (1 - p_{ij}) \right) \forall n' \in [1, n] \tag{9}$$

$$= \min_{n' \in [1, n]} \sum_{R \in R} \left( \prod_{j \in R} p_{ij} \prod_{j \in (F \cap P) \setminus R} (1 - p_{ij}) \right) \tag{10}$$

To simplify the notation, we re-write this in terms of the following sets:
\[ U = F \cap P \]  
\[ U_- = \{ x \in U : p_{ix} \leq p_{in'} \} \]  
\[ U_+ = U \setminus U_- \]  
\[ R_- = \{ S : S \subset U_- \wedge |S| \geq n' - 0.05n \} \]  
\[ R_+ = 2^{U_+} \]

These essentially divide \( U \) and \( R \) into their corresponding sets for elements less than \( p_{in'} \) and elements greater than \( p_{in'} \). Now we can rewrite Equation (10) as:

\[
\min_{n' \in [1,n]} \sum_{S \in R_-} \sum_{S^+ \in R_+} \left( \prod_{j \in S^-} p_{ij} \prod_{j \in S^+} 1 - p_{ij} \prod_{j \in U_- \setminus S^-} 1 - p_{ij} \right) \leq \min_{n' \in [1,n]} \sum_{S \in R_-} \prod_{j \in S^-} p_{ij}
\]

\[ \leq \min_{n' \in [1,n]} \sum_{S \in R_-} (p_{in'})^{|S^-|} \]  
\[ \leq \min_{n' \in [1,n]} \sum_{k=n' - 0.05n}^{n'} \binom{n'}{k} p_{in'}^k \]  

In equation (21), we replace the sum in the previous equation with a sum over the possible sizes of \( R \). For each size, the binomial term gives the number of sets of size \( k \), and the last term gives the probability of a set of size \( k \).