LOCP: Methods

Let X be a vector of ones and zeros, where ones stand for pilus related sequences and zeros stand for sequences not related to pili. Let N denote the size of X and let K denote the total number of ones in X. Let w, window, denote a subset of consecutive entries in X, let n denote the size of w, let k denote the total number of ones in w and let *max gap* denote the maximum number of consecutive zeros in w. Let as define a subset of all possible windows by:

$$W(X,M) = \{w \mid w\_starts\_and\_ends\_with\_1, \max gap(w) \leq M\}$$

where M is a user defined variable (by default equal to five).
For each w in W(X,M), LOCP computes the P-value, P(w) using a one-tailed Fisher's Exact Test:

$$P(w) = \sum_{k'=k}^{n} \binom{K}{k'} \binom{N-K}{n-k'} \Big/ \binom{N}{n} \qquad (1)$$

To estimate P_adj, LOCP runs 1000 simulation rounds. On each simulation round i ($1 \leq i \leq 1000$), LOCP samples a random permutation of X, denoted X_i. After sampling X_i, LOCP calculates the P-value for each w in $W(X_i,M)$ using (1). Then, for each simulation i LOCP determines

$$P_{i,\min} = \min(\{P_i(w) \mid w \in W(X_i, M)\})$$

Finally, $P_{adj}(w)$ is estimated as the fraction of these $P_{i,\min}$ that are less than or equal to P(w).