This script is designed to find the optimal positions for inserting an HA tag into a set of proteins. The HA tag is a short peptide sequence derived from the influenza hemagglutinin protein, often used in biochemistry and molecular biology to tag a protein of interest. The script attempts to choose insertion sites such that the HA tag’s insertion neither disrupts the protein's function nor overlaps with the signal peptide cleavage site, maximizing its accessibility for detection. The script imports and processes data from SignalP and NetSurfP predictions performed on sequences with and without the HA tags inserted at a series of sites downstream from the cleavage site, assesses potential HA tag insertion sites based on various predicted biophysical parameters, scores each site, and selects the optimal position for each protein. The results, along with detailed logs, are saved for further analysis. Adjustments to scoring weights and parameters can be made to refine the selection process.

***Cutoffs***

RSA, coil propensity, and disorder propensity were examined in the residue just before and just after the candidate HA insertion position in both the original sequencec and the sequence with the HA tag inserted (4 positions total). The cutoffs were applied to all four of these positions.

cut\_threshold\_abs (0.9):

It is the threshold for the absolute probability that a cut site exists in the protein sequence.

If the cut probability is below this threshold, the position is skipped, meaning it's not considered a viable option for HA tag insertion.

cut\_threshold\_loc (0.75):

Represents the threshold for the probability of the location of the cut site.

Similar to cut\_threshold\_abs, if the cut site location probability is below this threshold, the position is not considered.

disorder\_threshold (0.5):

It's the minimum required probability of disorder on both sides of the proposed HA insertion site.

Positions with a disorder probability below this threshold are skipped unless there are no positions above the threshold.

rsa\_threshold (0.25):

Indicates the minimum required RSA on both sides of the HA insertion site.

Positions below this threshold for RSA are not considered viable for HA tag insertion.

coil\_threshold (0.5):

Represents the minimum required probability of a coil structure on both sides of the HA insertion site.

Positions are skipped if the coil probability is below this threshold.

***Function Process***

The function iteratively goes through each protein transcript in the transcript\_list, evaluates various parameters at each possible HA insertion position, and calculates a score based on the weighted sum of these parameters. The position with the highest score is considered the best position for HA tag insertion.

***Important Points***

HA Cleavage Check: The function ensures that the new cut site does not cleave the inserted HA tag. If there's a risk of cleavage, the position is marked as 'HA\_CLEAVED' and skipped.

Fallback Mechanism: If no position meets all the threshold criteria, the function falls back to positions with a disorder probability above the threshold. If still none, it chooses the position with the highest cut site probability that doesn’t risk cleaving the HA tag.

Gaussian Function for Distance Scoring: The distance of the insertion site from a specified position (4 by default) is scored using a Gaussian function. It ensures that positions closer to this specified location receive higher scores.

Logging: All the intermediate steps, checks, and decisions are logged into a text file for detailed analysis and review.

***Summary***

The function utilizes multiple cutoffs to ensure that the selected HA tag insertion site is optimal based on the cut site probabilities and biophysical properties of the surrounding amino acids. It provides a robust mechanism to select the most suitable position, considering the constraints and criteria defined by the cutoffs, and implements fallback mechanisms to ensure a position is selected even if none meet all criteria.