DeNovo: Dissimilarity-Random-Sampling Algorithm

The Dissimilarity-Random-Sampling algorithm below takes as input a viral protein x, a set of human proteins $H(x)$ interacting with x, and a list L of $(y, H(y))$ pairs for different viral proteins y and their corresponding human interacting partner sets $H(y)$. An input parameter T serves as a cutoff value for the sequence dissimilarity between the viral proteins when selecting the less likely interacting human partners. The algorithm returns a set of n human proteins $N(x)$ to be paired with x as negative interactions.

Dissimilarity-Random-Sampling first calculates, for each viral protein in V, the $BitScore$ of globally aligning that protein with the viral protein in question x. Global alignment ($GlobalAlign$) was performed using the Needleman-Wunsch algorithm with the BLOSUM30 matrix to capture distant similarities. As V may contain some outliers (viral proteins with small bit score, and hence large dissimilarity, to all other viral proteins), $RemoveOutliers$ identifies these outliers and returns a reduced list of viral proteins V'. The bit scores are then normalized between 0 and 1 ($NormBitScore$) over all the bit scores of V' with x ($BitScore'$). $min(BitScore')$ and $max(BitScore')$ are the minimum and maximum values of $BitScore'$, respectively. One minus the normalized bit score for a viral protein y with x is used as a sequence dissimilarity distance between them ($Distance(x, y)$). If $Distance(x, y)$ is close to zero, then y is highly similar in sequence to x compared to the other viral proteins in V. If this dissimilarity distance is larger than the dissimilarity threshold T, then the human proteins interacting with y ($H(y)$) can be added to the set of unlikely interaction partners with x ($N(x)$). At $T = 0$, there is no dissimilarity enforced, and the dissimilarity-based method is equivalent to the non-constraint random sampling. Finally, n human proteins are randomly selected from $N(x)$ by $RandomPick$.

Dissimilarity-Random-Sampling(x, n, T)
1 // Selecting n negative interaction partners for viral protein
2 // x from interaction partners of other viral proteins that
3 // are more than T dissimilar in sequence to x.
4 // V is the set of all viral proteins in the data set from which
5 // negative samples are drawn.
6 // L is a list of viral proteins in V and their corresponding human
7 // interacting partners $H(V)$.

8 $N(x) = \emptyset$
9 **for** $y \in V$
10 $BitScore(y) = GlobalAlign(x, y)$
11 $[V', BitScore'] = RemoveOutliers(V, BitScore)$
12 **for** $y \in V'$
13 $NormBitScore(y) = \frac{BitScore'(y) - min(BitScore')}{max(BitScore') - min(BitScore')}$
14 $Distance(x, y) = 1 - NormBitScore(y)$
15 **if** $Distance(x, y) > T$
16 $N(x) = N(x) \cup H(y)$
17 $N(x) = N(x) \setminus H(x)$
18 $N(x) = RandomPick(N(x), n)$
19 **return** $N(x)$