DeNovo: Feature-Extraction

The feature extraction scheme first clusters the 20 amino acids based on similarities of physiochemical properties known to drive most PPIs (dipoles and volumes of side chains) . The amino acid clusters are $\{A, V, G\}$, $\{I, L, F, P\}$, $\{Y, M, T, S\}$, $\{H, N, Q, W\}$, $\{R, K\}$, $\{D, E\}$, and $\{C\}$.

The amino acid sequence of each protein is mapped to the corresponding cluster numbers. For example, $ACDHNYA$ is mapped to the vector $(1, 7, 6, 4, 4, 3, 1)$, for A is in Cluster 1, C in Cluster 7, and so on.

The frequency of each possible 3-mer was calculated in each mapped protein, generating a feature vector of length $7^3 = 343$ that was then normalized for each protein independently. If $F = (f_1, f_2, ..., f_{343})$ is a protein feature vector, then each normalized element f_i' is calculated as:

$$f_i' = \frac{f_i - min(F)}{max(F) - min(F)} \tag{1}$$

The two normalized vectors of an interacting (or negative) pair were concatenated into a single feature vector representing the interaction, with the viral protein vector coming first.

For each of the (R, S) training-testing pairs, the values of each feature (f_i') were independently normalized to lie between 0 and 1 across all feature vectors in the training set. The mean and variance for each feature were carried out to standardize the values of the corresponding feature in the testing samples. At this stage, each subset (R or S) was represented by a set of feature vectors for the positive and negative interaction along with their labels.