Implementation protein sequence segmentation in AAC and DC as protein descriptors for improving a classification performance of acetylation prediction

Abstract - Post-Translational Modification (PTM) identification is carried out to determine the position of the PTM in protein. Acetylation in the lysine protein is one of the many types of PTM that play an important role in biological processes. In existing research, identification of lysine acetylation was developed by computational methods, using several available protein descriptors along with classification methods. Research on protein classification usually only uses the length of the protein sequence to describe the state of the whole protein, not its local state. Knowing the local state of the protein sequence will have a good effect on the classification results. To find out the situation, the protein sequence segmentation approach is done by adjacent and overlapped segments. Adjacent and overlapped segments divide the length of the protein into several segments, then numerical features will be calculated, so that information about the protein is also obtained locally. Calculation of numerical features using the Amino Acid Composition and Dipeptide Composition descriptors, then the data is classified with Support Vector Machine. The experimental results show that protein segmentation increases the performance of protein classification by 0.7-2.5%. Segmentation using adjacent and overlapped segments provides improved performance. In this research, it was found that protein segmentation affected the performance of protein classification, especially in overlapped segments.