**Paper Summary**

PredCar-site: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue .

The experimental technologies are costly and time-consuming to detect the carbonylation sites in proteins. An accurate computational method for predicting carbonylation sites is an urgent issue which can be useful for drug development. In this study, a novel computational tool termed predCar-Site has been developed to predict protein carbonylation sites by Incorporating the sequence-coupled information into the general pseudo amino acid composition, By Different Error Costs method balancing the effect of skewed training dataset and As a classiﬁer constructing a predictor using support vector machine. The carbonylation is found as an irreversible post-translational modiﬁcation and considered a biomarker of oxidative stress. It also associated with some diseases such as Alzheimer's disease, diabetes, chronic renal failure, chronic lung disease and Parkinson's disease. Carbonylation has been considered as a biomarker for oxidative stress due to its some unique characteristics such as relatively early formation, Stability, Irreversibility. By The predCar-Site predictor achieves an average AUC (area under curve) score of 0.9959, 0.9999, 1, and 0.9997 in predicting the carbonylation sites of K, P, R, and T. One of the major challenges is to handle imbalance dataset problem as it is found in most of the dataset. For this kind of prediction, the number of negative subset is much larger than the corresponding positive subset. As the real world picture is that the non-carbonylation sites are always the majority compared with the carbonylation ones, so naturally the predictor should be biased to the non-carbonylation sites. Here the problem is that, for this type of predictors may interpret many carbonylation sites as non-carbonylation sites. But, the information about the carbonylation sites is mostly desired than non-carbonylation sites. As a result, it is crucial to ﬁnd an effective solution to balance this kind of bias consequence. Sequence sample, one of the critical problem in bioinformatics is how to extract vector from biological sequence with keeping considerable sequence characteristics. Challenge in imbalance problem is that the small classes are often more useful, but standard classiﬁers tend to be weighed down by the huge classes and ignore the tiny ones. For predictor support vector machine , features used in this predictor are extracted by vectorized sequence-coupling model, Data imbalance issue the Different Error Costs (DEC) method, To estimate the skill of machine learning models stratiﬁed 10-fold cross-validation and Implement system Matlab 2014. iCar-PseCp's benchmark dataset set has been used in this study. iCar-PseCp's dataset was derived from the 230 carbonylated protein sequences from human and 20 carbonylated protein sequences from Photobacterium and window size, < 30%. Probability values are derived from the negative benchmark dataset given in Supporting Information S1, S2, S3, and S4, respectively. In this paper, we have used a Different Error Costs(DEC)method to handle imbalance dataset problem of carbonylation sites prediction. The Different Error Costs (DEC) method is a cost-sensitive learning solution proposed to overcome the imbalance dataset problem for SVM. used k-fold cross-validation (subsampling) method to save the computational time. repeated the 10-fold cross-validation for 5 times. In this study value of C and sigma which appears most of the times as best model in 5 complete runs of the 10-fold cross-validation to train the system for the web. Carbonylation sites predictions. Apart from the metrics, they have calculated precision too for their system and got the average (±standard deviation) values of 83.19(±0.62)%, 97.52(±0.82)%, 93.95(±1.67)%, and 94.66(±1.19)% in predicting the carbonylation sites for K, P, R and T respectively.