# Predicting Protein Subcellular Localization Using Abstract Sequential Features

Adrian Silvescu, Cornelia Caragea, and Vasant Honavar

Determination of protein subcellular localization plays an important role in understanding protein function [1]. Knowledge of subcellular localization is also essential for genome annotation and drug discovery. Moreover, abnormal subcellular localization has been correlated with several diseases, such as cancer and Alzheimer's disease. Experimental determination of protein subcellular localization is expensive and laborious. Hence, there is significant interest in the development of computational methods for reliable prediction of protein subcellular localization.

In this study, we present a machine learning approach to predicting protein subcellular localization from amino acid sequences. Our approach exploits the complementary strengths of *feature construction* (constructing complex features by combining existing features) and *feature abstraction* (grouping of similar features to generate a more abstract feature) or *feature selection* to adapt the data representation used by the learner. In particular, consider a special case of topologically constrained feature construction, namely, super-structuring. Super-structuring provides a way to increase the predictive accuracy of the learned models by enriching the data representation [2] (and hence increasing the complexity of the learned model) whereas abstraction or selection [3] help reduce the model size by simplifying the data representation.

We performed experiments that compare Naive Bayes Multinomial classifiers constructed using the original features with those constructed using feature selection, feature abstraction, and the combination of abstraction and super-structuring and feature selection and super-structuring on two datasets from the bioinformatics domain: **Eukaryotes** and **Prokaryotes**.

The results of our experiments on both data sets show that adapting data representation by combining abstraction and super-structuring makes possible to construct predictive models that use significantly smaller (1-3 orders of magnitude) number of features than those that are obtained using super-structuring alone without sacrificing predictive accuracy (Fig.1).

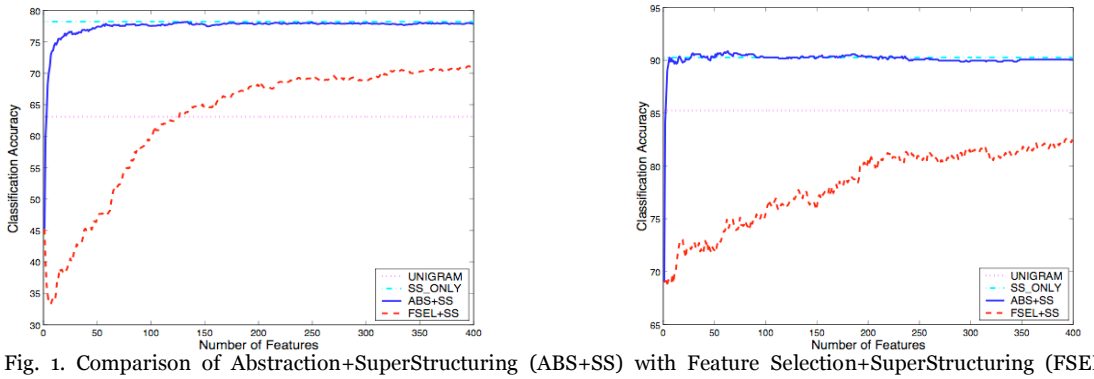


Fig. 1. Comparison of Abstraction+SuperStructuring (ABS+SS) with Feature Selection+SuperStructuring (FSEL+SS), SuperStructuring only (SS_ONLY), and Unigram (UNIGRAM) on the **Eukaryotes** (left) and **Prokaryotes** (right) data sets using unigrams and 3-grams. For the same number of features used to train the classifiers, ABS+SS is superior in performance to FSEL+SS, and UNIGRAM. After a relatively small number of features, ABS+SS achieves the performance of SS_ONLY. For a small drop in performance on both **Eukaryotes** and **Prokaryotes**, we obtain a reduction of model sizes by three orders of magnitude.

References

[1] Yu, C.S., Chen, Y.C., Lu, C.H., Hwang, J.K. (2006). Prediction of protein subcellular localization. *Proteins* 64:643-51.
[2] Liu, H., and Motoda, H. (1998). *Feature Extraction, Construction and Selection.* Springer.
[3] Zhang, J.; Kang, D.-K.; Silvescu, A.; and Honavar, V. 2006. Learning accurate and concise naive bayes classifiers from attribute value taxonomies and data. *Knowledge and Information Systems* 9(2):157–179.