Classification of Enzymes via Machine Learning Approaches

Neetika Nath*, John B. O. Mitchell

Biomedical Sciences Research Complex and School of Chemistry, University of St Andrews, KY16 9ST
*nn223@st-andrews.ac.uk

Keywords: Enzyme Classification, Machine Learning, R, Protein Function Prediction.

Abstract: We compare enzyme mechanistic descriptors derived from the MACiE (Mechanism, Annotation and Classification in Enzymes) database [Holliday et al., 2006] and use multivariate statistical analysis for assessment of enzyme classification. Each enzyme has an Enzyme Commission (EC) number, a numerical code designed to classify enzymes by describing the overall chemistry of the enzymatic reaction. The EC number system was devised five decades ago, in a pre-bioinformatics age. As the volume of available information is increasing, a large number of informatics groups have tried to use protein sequence and structural information to understand and reproduce the classification, some of which have been successful [Cai et al., 2004]. Other groups have effected automatic EC classification using chemoinformatics descriptions of the underlying reactions. Our objective is to develop a computational protocol using the *R* package **CARET** [Kuhun *et al.*, 2007] to predict EC number from MACiE-derived descriptors. We evaluate 260 well annotated chemical reaction mechanisms of enzymes using machine learning methods, placing them into the six top level EC classes. Moreover, we compare the classification performances of three supervised learning techniques, Support Vector Machine (SVM) [Vapnik, 1998], Random Forest (RF) [Breiman, 2001] and K Nearest Neighbour (kNN), for the reaction mechanism classification task using five different descriptor sets from MACiE data. **Results:** We found that all classifiers performed similarly in terms of overall accuracy with the exception of K Nearest Neighbour analysis, which has the lowest performance. The best performance was achieved by the Random Forest classifier.

References

Holliday, G.L., et al.. (2006) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms http://www.ebi.ac.uk/thornton-srv/databases/MACiE/

Cai, C., L. Han, et al. (2004). "Enzyme family classification by support vector machines." *Proteins* 55: 66 - 76.

Kuhn, M., Wing, J., Weston, S, Williams A., Keefer, C. & Engelhardt, A. (2007) caret: Classification and Regression Training, http://cran.r-project.org/web/packages/caret/

Vapnik, V. N. (1998) Statistical Learning Theory. New York: John Wiley and Sons.

Breiman, L. (2001) “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5– 32.