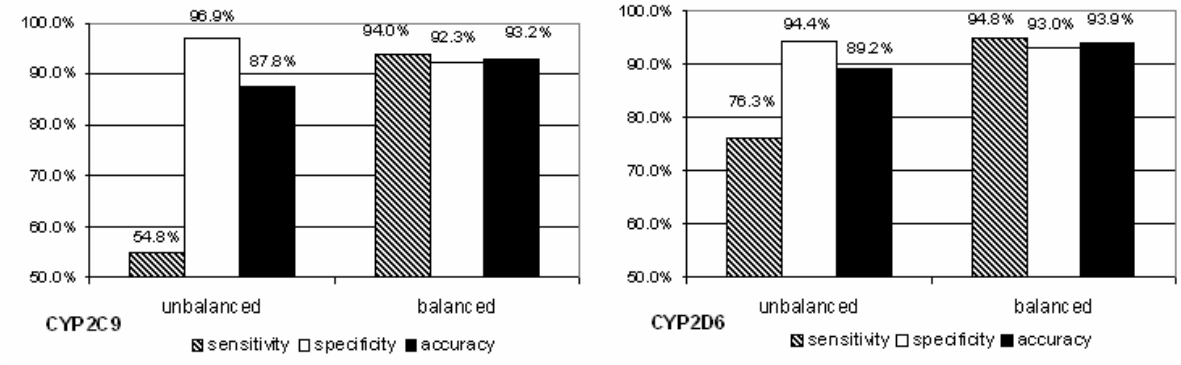
Classification of metabolic CYP P450 substrates: Improved sensitivity due to balanced data sets

Miriam Carbon-Mangels, Michael C. Hutter

Center for Bioinformatics, Saarland University, Saarbrücken, Germany



The prediction of metabolic stability of pharmaceutical substances is of substantial interest during the preclinical development of drugs. [1] Particularly enzymes from the superfamily of Cytochrome P450 are involved in hydroxylation and other redox reactions. The isoforms CYP2D6, CYP3A4, and CYP2C9 together contribute to more than 95% of all biotransformations of xenobiotics. We collected a data set of compounds that are known to be either substrates or nonsubstrates of these enzymes. To determine the most relevant molecular descriptors decision trees and support vector machines were used for binary classification into substrates and nonsubstrates. Furthermore, the LASSO method (least absolute sum of squared operators) [2] was applied for the selection of relevant descriptors. We found that the ratio of substrates to nonsubstrates within the training sets strongly influences the sensitivity (percentage of correctly classified substrates) and specificity (percentage of correctly classified nonsubstrates), despite similar accuracy (percentage of all correctly classified compounds). In general, balanced data sets, in which the number of compounds in either class is approximately equal, yield more adequate results.

[1] M.C. Hutter, *Curr. Med. Chem.*, **2009**, *16*, 189-202.
[2] R Development Core Team, *R: A Language and Environment for Statistical Computing*, **2007**.