Abstract:
An important aspect of the functional annotation of enzymes is not only the type of reaction catalysed by an enzyme, but also the substrate specificity, which can vary widely within the same family. In many cases, prediction of family membership and even substrate specificity is possible from enzyme sequence alone, using a nearest neighbour classification rule. However, the combination of structural information and sequence information can improve the interpretability and accuracy of predictive models. The method presented here, Active Site Classification (ASC), automatically extracts the residues lining the active site from one representative three-dimensional structure and the corresponding residues from sequences of other members of the family. From a set of representatives with known substrate specificity, a Support Vector Machine (SVM) can then learn a model of substrate specificity. Applied to a sequence of unknown specificity, the SVM can then predict the most likely substrate. The models can also be analysed to reveal the underlying structural reasons determining substrate specificities and thus yield valuable insights into mechanisms of enzyme specificity. We illustrate the high prediction accuracy achieved on two benchmark data sets and the structural insights gained from ASC by a detailed analysis of the family of decarboxylating dehydrogenases. The ASC web service is available at http://asc.informatik.uni-tuebingen.de/.