

Automatic classification of P-type ATPases using Structured Logistic Regression

P-type ATPases are a very large family of ATP-driven membrane pumps involved in transmembrane transport of charged substrates. The family can be divided into 5 groups and 11 subgroups which appears specific for particular classes of substrates. The manually curated P-Type ATPase database (PATbase, Alexsen and Palmgren, 1998; Møller et al. 2008) currently contains 490 P-Type ATPase sequences grouped in these subgroups. However the large and rapidly growing number of possible P-type ATPase sequences calls for an automated procedure to facilitate analysis of the distribution of pumps into different subgroups.

Using a newly developed text categorization tool, Structured Logistic Regression (SLR) (Ifrim et al. 2008), trained on PATbase, we have constructed a classifier based on binary SLR categorizers. The classifier can identify and distinguish between the eleven subfamilies with high accuracy, and requires no user input other than the sequence itself. This final classifier is available via: <http://www.birc.dk/p-classifier/>.

To evaluate our method, we used our classifier on the UniProt Knowledgebase (Swiss-Prot 56.8/TrEMBL 39.9) containing 7.754.276 sequences. Of these 8.292 were found by the SLR classifier to be P-type ATPases and 6.624 could be uniquely classified to a specific subgroup. The classification of 7.7 million sequences took around 50 minutes. As a sanity check, we built a NJ-tree of the 6.624 sequences and found large agreement between our groupings and subtrees. Currently detailed data analysis is being undertaken e.g. to search for new subgroups and for bacterial homologous to mammalian sequences in the known subgroups.