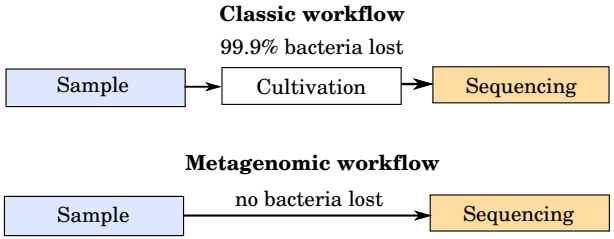
Detection of Enzymes in Metagenomic Data

Author: Ing. Stanislav Smatana, Supervisor: Ing. Jiří Hon

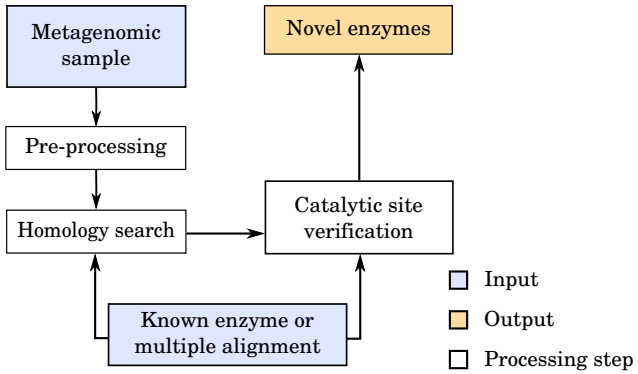
Enzymes represent a class of protein molecules, which is essential to all living beings. Although, it is not only living things who benefit from their abilities. In 2003, global sales of the enzyme-related industry reached 2.3 billion USD. Their wide industrial use results in constant hunger for novel enzymes with better properties. The main goal of my thesis is to assist in enzyme discovery and create tool for discovery of novel enzymes in metagenomic data.

Why metagenomics ?

Until recently, the search was limited to a small range of enzymes, which could be extracted from bacteria cultivated in laboratory conditions. It was estimated, that only 0.01% of bacteria could be grown in this way. Therefore, metagenomics, which allows extraction of genetic material without prior cultivation, unlocks the full potential for enzyme discovery.



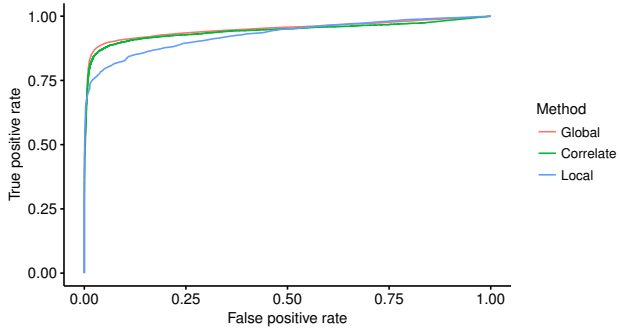
The Tool for Detection of Novel Enzymes



Given a sequence (or multiple alignment) of some known enzyme as the input, the toolset outputs novel enzymes with the same function, which are present in the given metagenomic sample. While their function is the same, found enzymes may have better properties. The toolset works in three main steps - pre-processing, search based on homology and catalytic site verification. The third step is neeccesary in order to guarantee that found enzymes really have the same function as the input, because function of sequentially similar enzymes may still be different. Developement of novel methods of catalytic function verification is the key contribution of my thesis.

Catalytic function verification

Two classes of verification methods were implemented in the system - methods based on alignment (both local and global) and a novel method based on normalized cross-correlation. While methods based on global alignment had the best performance, the novel correlation-based method was significantly faster with only small decrease in performance. The plot below shows ROC curves of the best variant of each method.



Conclusion

The tool was succesfuly implemented and experiments have shown, that proposed methods of catalytic verification reach sensitivity of 89%, specificity of 95% and the best average throughput of 1,203 verifications per second on regular personal computer. In conclusion, the tool, especially while using correlation-based verification, is suitable for practical enzyme discovery.

Brno University of Technology, Faculty of Information Technology
Loschmidt Laboratories