Integrated approaches for the discovery of novel enzymatic activities

Guillaume REBOUL, Maria SOROKINA, Jonathan MERCIER, Karine BASTARD, Mark STAM, David VALLENET, Claudine MEDIGUE – CEA, Genoscope, LABGeM

**Summary**

Next Generation Sequencing technology has dramatically increased the number of available sequences in public databases. At the same time, many enzymatic activities (~22%) are orphans of protein sequence (Sorokina et al., 2014). The large amount of available protein sequences is an opportunity to discover enzymes associated to new reactions. We present here an integrated bioinformatics approach to reduce this lack of knowledge in metabolism and to propose new activity/protein associations for experimental validation. With this objective, the "New Enzymatic Activity" group of the LABGeM team is developing several methods. The CanOE method combines genomic and metabolic contexts to predict candidate genes for orphan enzymes (Smith et al., 2012). Currently, this approach is extended to the detection of conserved chemical transformation motifs in the metabolism (Sorokina et al., submitted).

From a structural point of view, the ASMC (Active Site Modeling and Clustering) method finds and compares active site pockets to classify enzymes of a family and detects important residues for substrate specificity (de Mele-Minardi et al., 2010). These methods were successful applied to elucidate the enzymatic diversity of a protein family of unknown function (Bastard et al., 2014). Their results, associated with present knowledge, must be unified in a database allowing the elaboration of strategies for the selection of enzymatic families of interest.

This work is supported by genomic and metabolic network data from MicroScope, a platform for microbial genome analyses (Vallenet et al., 2013).

**Literature references**


Mercier, J., Valiente, D. GROOLS: Reactive Graph Reasoning for Genomes Annotation. RuMIS 2015 Conference.