EXPERIMENT-BASED COMPUTATIONAL METHOD FOR PROPER ANNOTATION OF THE MOLECULAR FUNCTION OF ENZYMES

Véronique de Berardinis, CEA/Genoscope/UMR8030 “Génomique Métabolique”
vberard@genoscope.cns.fr

Key Words: Sequence and function-based discovery, high-throughput screening, structural modeling, genomic context

The rate of protein functional elucidation lags far behind the rate of gene and protein sequence discovery, leading to an accumulation of proteins with no known function. Millions of protein database entries are not assigned reliable functions, preventing the full understanding of chemical diversity in living organisms. Pfam contains over 16,712 families, among which more than 3,919 families are of unknown function (DUF families). An additional difficulty, often underestimated, is that only a tiny fraction of enzymes have experimentally established functions and in most cases, function is extrapolated from a small number of characterized proteins to all members of a family leading to over-annotation[1,2]. Here, two examples of an integrated strategy for the discovery of various enzymatic activities catalyzed within protein families will be presented. This approach relies with a high-throughput enzymatic screening on representatives, structural and modeling investigations, analysis of genomic and metabolic context. The structural analysis is in both cases based on the Active Site Clustering Method[3] developed at Genoscope. We investigated the protein family with no known function, DUF849 Pfam family, and unearthed 14 potential new enzymatic activities, leading to the designation of these proteins as □-keto acid cleavage enzymes[4]. In addition, we propose an in vivo role for four enzymatic activities and suggest key residues for guiding further functional annotation. The second study will illustrate that proteins with high sequence similarity might not have the same function. We determined the enzymatic activities of 100 *O*-acyl-L-homoserine transferases representative of the biodiversity of the two unrelated families, MetX and MetA, involved in the first step of the methionine biosynthesis and assumed to always use acetyl-CoA and succinyl-CoA, respectively. We interpreted the results by structural classification of active sites based on protein structure modeling. We identified the specific determining positions responsible for acyl-CoA specificity in the active sites of MetX and MetA enzymes, actually iso-functional for both activities. We then predict that >60% of the 10,000 sequences from these families currently in databases are incorrectly annotated. Finally, we uncovered a divergent subgroup of MetX enzymes in fungi that participate only in L-cysteine biosynthesis as O-succinyl-L-serine transferases[5]. Our results show that the functional diversity within a family may be largely underestimated. The extension of this strategy to other families will improve our knowledge of the enzymatic landscape and the chemical capabilities of biodiversity.

References:

1 de Crecy-Lagard, V. Quality Annotations, a Key Frontier in the Microbial Sciences. Microbe 11, 303-310 (2016).
2 Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5, e1000605 (2009).
3 de Melo-Minardi, R. C., Bastard, K. & Artiguenave, F. Identification of subfamily-specific sites based on active sites modeling and clustering. Bioinformatics 26, 3075-3082, doi:10.1093/bioinformatics/btq595 (2010).
4 Bastard, K. et al. Revealing the hidden functional diversity of an enzyme family. Nature chemical biology 10, 42-49, doi:10.1038/nchembio.1387 (2014).
5 Bastard, K. et al. Parallel evolution of non-homologous isofunctional enzymes in methionine biosynthesis. Nature chemical biology june (2017).