Evolution and Classification of GH101 Family Glycoside Hydrolases: Four Years Later

Daniil G. Naumoff

S. N. Winogradsky Institute of Microbiology, Russian Academy of Sciences, Prospekt 60-letiya Oktyabrya 7/2, Moscow 117312, Russia; daniil_naumoff@yahoo.com

Recent progress in genome sequencing resulted in a rapid growth of the sequence databases. The number of known representatives of all protein families has increased dramatically. It naturally raises a question if the existing protein classifications are still stable or appearance of «the intermediate forms» washed out the clear borders between protein families or subfamilies. As a case study, we choose family GH101 of glycoside hydrolases.

On the basis of sequence similarity of the catalytic domains all glycoside hydrolases have been grouped into more than 100 families (GH1-GH131). Family GH101 includes retaining endo-α-*N*-acetylgalactosaminidases (EC 3.2.1.97) and their uncharacterized homologues (totally 100 proteins) [1]. This family was described in 2005 [2]. Four years ago we revealed 95 non-identical protein sequences of GH101 domains from GenPept database using the blast algorithm [3]. These sequences represented 18 genera of bacteria from *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, and *Verrucomicrobia*. Pairwise sequence alignment and phylogenetic analysis allowed us to distinguish six subfamilies (101a-101f) in the GH101 family. Iterative screening of the protein database by PSI-BLAST revealed the closest relationship of GH101 domains with GH129 (or GHL1) domains. More distant similarity was found with some proteins from GH13, GH20, GH27, GH29, GH31, GH36, GH66, GH70, GH97, COG1306, and COG1649 families [3-5].

Recently we did another analysis of the GenPept database. Iterative screening of the database allowed us to reveal more than 15,000 proteins homologous to GH101 domains. They represent several families of glycoside hydrolases having the TIM-barrel type catalytic domains. Particularly we found 345 proteins containing the GH101 domain. This family is still quite distinct and its closest neighbor is GH129 family of α-N-acetylgalactosaminidases

(EC 3.2.1.49). In order to additionally increase the number of proteins for phylogenetic analysis we also used 17 GH101-containing proteins found in the Integrated Microbial Genomes database [6]. All obtained proteins belong to the same four bacterial phyla as four years ago with clear domination of representatives from *Actinobacteria* and *Firmicutes*. We found no GH101-containing proteins from *Proteobacteria* despite the fact that almost a half of all bacterial genome projects corresponds to this phylum [7]. According to the phylogenetic analysis, the subfamily structure retains stable: almost all proteins can be easily classified in six subfamilies described by us four years ago [3]. All subfamilies compose stable clusters on the tree. In contrast to many other glycoside hydrolase families, the role of lateral gene transfers and gene duplications was minor during the evolution of genes encoding GH101-containing proteins.

1. The Carbohydrate-Active Enzymes Database [http://www.cazy.org/].

2. K. Fujita et al. (2005) Identification and molecular cloning of a novel glycoside hydrolase family of core 1 type *O*-glycan-specific endo-α-*N*-acetylgalactosaminidase from *Bifidobacterium longum*, *J. Biol. Chem.*, **280**:37415–37422.

3. D.G. Naumoff (2009) Sequence analysis of endo-α-*N*-acetylgalactosaminidases and their homologues, *Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB'09), July 20-23, 2009, Moscow, Russia*, P.251–252 [http://mccmb.belozersky.msu.ru/2009/MCCMB09_Proceedings.pdf].

4. D.G. Naumoff (2009) Sequence analysis of endo-α-*N*-acetylgalactosaminidases, *Glycoconjugate J.*, **26**:847.

5. D.G. Naumoff (2010) GH101 family of glycoside hydrolases: Subfamily structure and evolutionary connections with other families, *J. Bioinform. Comput. Biol.*, **8**:437–451.

6. The Integrated Microbial Genomes [https://img.jgi.doe.gov/].

7. The Genomes On-Line Database [http://www.genomesonline.org/].