RAPID COMMUNICATION

Genomics & Informatics Vol. 4(4) 182-187, December 2006

In Silico Identification of 6-Phosphogluconolactonase Genes that are Frequently Missing from Completely Sequenced Bacterial Genomes

Haeyoung Jeong[1]*, Jihyun F. Kim[2] and Hong-Seog Park[1]

[1]Genome Research Center, KRIBB, Guseong dong, Yuseong-gu, Daejeon 305-806, Korea, [2]Systems Microbiology Research Center, KRIBB, Guseong dong, Yuseong-gu, Daejeon 305-817, Korea

Abstract

6-Phosphogluconolactonase (6PGL) is one of the key enzymes in the ubiquitous pathways of central carbon metabolism, but bacterial 6PGL had been long known as a missing enzyme even after complete bacterial genome sequence information became available. Although recent experimental characterization suggests that there are two types of 6PGLs (DevB and YbhE), their phylogenetic distribution is severely biased. Here we present that proteins in COG group previously described as 3-carboxymuconate cyclase (COG2706) are actually the YbhE-type 6PGLs, which are widely distributed in *Proteobacteria* and *Firmicutes*. This case exemplifies how erroneous functional description of a member in the reference database commonly used in transitive genome annotation cause systematic problem in the prediction of genes even with universal cellular functions.

Keywords: 6-phosphogluconolactonase, 3-carboxymuconate cyclase, missing enzyme, pentose phosphate pathway, genome annotation

In the course of annotation processes of several *Firmicutes* genomes that had been completely sequenced by our institution, we found that genes encoding 6-phosphogluconolactonase (6PGL) are frequently missing throughout functional information transfer by homology from public amino acid databases, a *de facto* annotation strategy for completely sequenced genomes. 6PGL (EC 3.1.1.31), catalyzing the hydrolysis of δ-6-phosphogluconolactone into 6-phosphogluconate, plays a key role in the oxidative branch of the pentose phosphate pathway. Since this pathway provides major biosynthetic capacity as one of central carbon metabolism, life without functional pentose phosphate pathway is considered to be impossible. Cordwell reviewed several "missing" enzymes from genome sequence data that were expected to exist as members of well-studied biochemical pathways and indicated that no gene corresponding to 6PGL had been identified until that time (Cordwell, 1999). Lack of enzymes catalyzing major metabolic reactions is sometimes indicative of the presence of alternative pathway, but it is often the case that incorrect gene assignment in genome database and its propagation misleads accurate prediction of essential genes (Brenner, 1999; Devos and Valencia, 2001). These intriguing observations led us to study current knowledge of the functional assignment of bacterial 6PGL. In this paper, we investigated why genes encoding 6PGL are underrepresented in bacterial genomes, and tried to elucidate structural and familial relationship in 6PGL proteins.

Two protein families in bacterial 6PGLs

Genes encoding 6PGL was first characterized from human cDNA (6PGL_HUMAN) (Collard *et al.*, 1999). This protein is homologous to bacterial *devB* gene product that is often found in proximity to the gene encoding glucose-6-phosphate dehydrogenase (*zwf*) and is also related to glucosamine-6-phosphate deaminase (EC 3.5.99.6, a synonym with glucosamine-6-phosphate isomerase). They expressed the cDNA clone in *E. coli* system and showed its 6-phosphogluconolactonase activity. Furthermore, they also suggested that bacterial *devB* gene product is most likely 6PGL.

The latter assumption was proved later by characterizing *devB* homolog in *Pseudomonas aeruginosa* (6PGL_PSEAE) (Hager *et al.*, 2000). When they BLAST searched against NCBI database using 6PGL_PSEAE as a query sequence, a large number of related proteins previously identified as DevB or SOL homologs were found. Multiple alignments of DevB family proteins constitute a HMM designated as TIGR01198 in TIGRFAMs database (IPR005900 in InterPro database) (Haft *et al.*, 2003; Mulder *et al.*, 2005). However, this family does not represent entire 6PGL proteins from the bacterial kingdom, since it does not cover members from virtually all *Firmicutes* and enteric bacteria such as *E. coli*.

*Corresponding author: E-mail hyjeong@kribb.re.kr, Tel +82-42-879-8137, Fax +82-42-879-8139
Accepted 30 Nov 2006

The COG database has one orthologous cluster related to DevB-type 6PGL (COG0363), but it also includes related but functionally diverged protein such as glucosamine-6-phosphate deaminase (Tatusov *et al.*, 2003).

It was not until very recently that *ybhE* was assigned to be the gene encoding 6PGL in *E. coli* by two independent research groups, although the locus was mapped at 17.2 min more than 30 years ago with unknown function (Thomason *et al.*, 2004; Zimenkov *et al.*, 2005). 6PGL in *E. coli* is totally unrelated from DevB-type enzyme. This is a good example of non-orthologous gene displacement, in which unrelated (or distantly related) enzymes catalyzes the equivalent biochemical reaction (Galperin and Koonin, 1998). Among the 260 proteins described as 6PGLs proteins in current UniProt KnowledgeBase Release 9.3 (Wu *et al.*, 2006), 196 entries are found to be members of TIGR01198 (true DevB family).

COG2706 (3-carboxymuconate cyclase) and 6PGL

YbhE protein belongs to both COG2706 (16 members, 3-carboxymuconate cyclase) and MF_01605 HAMAP family (17 members, 6PGL) (Gattiker *et al.*, 2003). Because COG clustering is based on similarity with a less stringent cutoff, some COGs may contain ortholgous proteins with diverged, heterogeneous functions as shown in COG0363 mentioned above. This apparent discrepancy of functional descriptions on a single protein can be thus elucidated by assuming that 3-carboxymuconate cyclase and 6PGL were so similar that they were classified into a single COG group. This sounds plausible, since there were no members of COG2706 proved to be 6PGL either by experiment or sequence analysis when COG database was built.

3-carboxymuconate cyclase (carboxy-*cis, cis*-muconate cyclase or 3-carboxy-*cis,cis*-muconate cyclase, EC 5.5.1.5) catalyzes the interconversion between 3-carboxy- *cis, cis*-muconate and 3-carboxy-2,5-dihydro-5-oxofuran-2-acetate. This reaction constitutes the β-ketoadipate pathway in some microorganisms, which is responsible for the catabolism of aromatic compounds such as protocatechuate and catechol into TCA cycle intermediates. At present, 3-carboxymuconate cyclase is characterized only from eukaryotic microorganism (CMLE_NEUCR) (Mazur *et al.*, 1994). The same substrate can be lactonized to a similar compound, 2-carboxy-2,5-dihydro-5-oxofuran-2-acetate by the prokaryotic enzyme 3-carboxy-*cis,cis*-muconate cycloisomerase (EC 5.5.1.2) (Lorite *et al.*, 1998; Murakami *et al.*, 2004). These two enzymes, collectively called CMLEs (3-carboxymuconate lactonizing enzymes), are not related to each other in spite of their common substrate.

Erroneous functional description of COG2706

Is 3-carboxymuconate cyclase best describing COG2706? Interestingly, none of the 16 original members constituting COG2706 were characterized as 3-carboxymuconate cyclase (see Table 1). On the other hand, six of them are true 6PGLs and also belong to MF_01605. Because the only 3-carboxymuconate cyclase ever characterized (CMLE_NEUCR) was not included in the 66 completely sequenced genomes for COG construction, function of COG2706 might have been assigned according to their overall sequence similarity to CMLE_NEUCR. When CMLE_NEUCR was subjected to BLAST search against COG database, however, similarities to COG2706 members were too low. For example, BS_ykgB (*Bacillus subtilis*) in COG database is the best hit for CMLE_NEUCR, but the BLASTP bit score is only 54. This observation indicates that assigning 3-carboxymuconate cyclase to COG2706 is inappropriate since it was based on weak sequence similarity. We also suggest that description of COG2706 should be changed into 6PGL since one member (YbhE in *E. coli*) of this cluster already has characterized biochemical functions. Association of EC 3.1.1.31 to COG2706 in KEGG Ortholgy definition file also supports our findings regarding this COG cluster (Kanehisa *et al.*, 2006). As mentioned above, this error might have been inevitable because no member of COG2706 was described as 6PGL when COGs were constructed. Once such errors occur in the reference genome database, they tend to be propagated by rather careless transitive annotation, and what is worse, they are seldom corrected. If prokaryotic protein database is searched with *E. coli* YbhE protein as a query, we can easily find strong positive hits throughout whole prokaryotic classes except for archaeal groups. Several hits have 6PGL as their names, but most of them are given either hypothetical protein or 3-carboxymuconate cyclase. Lack of 6PGL in *Firmicutes* thus results from misnomer of genuine 6PGL enzymes. A recent report suggested that Bfl341 protein from *Blochmannia floridanus* formerly classified into COG2706 should be 6PGL due to its sequence similarity with *E. coli* YbhE, but they did not explore the possibility of inaccurate functional description given to COG2706 (Gaudermann *et al.*, 2006)

Relationship between COG2706 and MF_01695

Table 1 summarizes the basic properties of 27 proteins belonging either to COG2706 or to MF_01695. They are similar in sizes and contain no known PFAM/PROSITE motifs. They do not overlap apparently with each other

Table 1. Combined list of proteins for COG2706 and MF_01695 family.

UniProt ID (COG ID)	Class[1]	Size (aa)	Organism	Description
6PGL_BUCAI (BU293)	C, H	334	*Buchnera* sp. APS	6-phosphogluconolactonase
6PGL_ECO57 (Ecs0795)	C, H	331	Escherichia coli O157:H7[2]	6-phosphogluconolactonase
6PGL_SALTY (STM0785)	C, H	331	*Salmonella typhimurium* LT2	6-phosphogluconolactonase
6PGL_ECOLI (ybhE)	C, H	331	*Escherichia coli* K12	6-phosphogluconolactonase
6PGL_YERPE (YPO1149)	C, H	334	*Yersinia pestis*	6-phosphogluconolactonase
YKGB_BACSU (BS_ykgB)	C	349	*Bacillus subtilis*	Hypothetical protein ykgB
Q97MV5_CLOAB (CAC0086)	C	359	*Clostridium acetobutylicum*	Muconate cycloisomerase related protein, ortholog of YKGB *B.subtilis*
YWCC_LACLA (L11851)	C	341	*Lactococcus lactis*	Hypothetical protein ywcC
Q92E93_LISIN (lin0567)	C	346	*Listeria innocua*	Lin0567 protein
Q9HWH7_PSEAE (PA4204)	C	388	*Pseudomonas aeruginosa*	Hypothetical protein
Q8Y3D1_RALSO (RSc0049)	C	418	*Ralstonia solanacearum*	Putative hemagglutinin-related protein
Q8XTD0_RALSO (RSp0183)	C	410	*Ralstonia solanacearum*	Putative hemagglutinin-related protein
Q97PT8_STRPN (SP1506)	C	337	*Streptococcus pneumoniae* TIGR4	Hypothetical protein
YBS1_SCHPO (SPBC18E5.01)	C	342	*Schizosaccharomyces pombe*	Hypothetical protein C18E5.01 in chromosome II
Q7A4P4_STAAN (SA1737)	C	342	*Staphylococcus aureus* N315	Hypothetical protein SA1737
6PGL_BLOFL	H	338	*Blochmannia floridanus*	6-phosphogluconolactonase
6PGL_BUCAP	H	333	*Buchnera aphidicola*	6-phosphogluconolactonase
6PGL_BUCBP	H	331	*Buchnera aphidicola*	6-phosphogluconolactonase
6PGL_ECOL6	H	331	*Escherichia coli* O6	6-phosphogluconolactonase
6PGL_ERWCT	H	332	*Erwinia carotovora*	6-phosphogluconolactonase
6PGL_PHOLL	H	328	*Photorhabdus luminescens*	6-phosphogluconolactonase
6PGL_SALPA	H	331	*Salmonella paratyphi*	6-phosphogluconolactonase
6PGL_SALTI	H	331	*Salmonella typhi*	6-phosphogluconolactonase
6PGL_SHIFL	H	331	*Shigella flexineri*	6-phosphogluconolactonase
6PGL_YERPS	H	334	*Yersinia pseudotuberculosis*	6-phosphogluconolactonase
Q3Z422_SHISS	H	331	*Shigella sonnei*	Putative isomerase
Q57RH2_SALCH	H	331	*Salmonella choleraesuis*	Putative 3-carboxymuconate cyclase

[1]H, HAMAP MF_01695 family; C, COG2706.

[2]COG2706 also includes ZybhE from *Escherichia coli* O157:H7 EDL933 strain. It is intentionally excluded from this list as it is essentially identical to 6PGL_ECO57 except for one amino acid designated as 'x'. UniProt database does not record ZybhE as a separate entry, either.

family, as only five proteins (BU293, Ecs0795, STM0785, ybhE, and YPO1149) found common in both groups. But this observation is feasible since clustering strategies are entirely different for two protein family database. When we scanned 16 COG2706 sequences against the entire HAMAP families, most of them matched MF_01695 at different score levels. Furthermore, 16 COG2706 proteins also constitute a PSSM-based family in CDD database (Marchler-Bauer *et al.*, 2005). These imply the two independently built protein families are essentially the same such that they can be given a single functional description.

Phylogenetic tree of the 27 proteins constructed by neighbor-joining method showed members belonging to COG group only tend to form a somewhat distant subgroup. Among them, *Ralstonia solanacearum* (two proteins), *Pseudomonas aeruginosa*, and *Schizosaccharomyces pombe* have 6PGL of aforementioned DevB family. These species may be currently undergoing non-orthologous gene displacement. We are not going to investigate this phenomenon here, as detailed analysis would go beyond the scope of this report.

Consequence of errors in the reference databases

COG is one of the major reference databases commonly used in microbial genome annotation. The utility of COG-based annotation is that it could increase the number of functional assignment with a newly sequenced genome because a single functional description can be given to all members of a COG cluster rather reliably even if the individual member was not yet assigned function (Natale *et al.*, 2000), and its functional classification category is very widely used in bacterial genomics study. The most highly annotated protein database such as Swiss-Prot, which is the first choice for transitive annotation, contains 6PGLs mostly from DevB family. If a newly found YbhE-type 6PGL from bacterial phylum other than *Proteobacteria* was processed, it would not get any functional clues. In

most cases, COG-based annotation is the next measure to taken when pairwise search against a protein database does not give discernable functional predictions. Erroneous annotation is resulted from adopting functional descriptions that are associated with the matched COG at this step. To make the matter worse, protein data set such as NCBI-NR or UniProt/TrEMBL, the last resource utilized when no functional information was obtained through the previous search against databases with higher accuracy, already contains probable YbhE-type 6PGL that named after the description of COG2706, 3-carboxymuconate cyclase. For example, 98.9% of 553 bacterial 3-carboxymuconate cyclase found in NCBI-NR dataset (as of Dec 2006) was named after their COG affiliation.

Although this kind of errors can be corrected to some extent in the secondary (or 'derived') protein databases, its effect is minimal. For example, 31 proteins in YbhE-type 6PGL in KEGG Orthology category do not fully cover this type of proteins in prokaryotes. This means that once the original description other than '(conservative) hypothetical protein' was given, it seems very reluctant to correct it no matter how correct the information might be.

In conclusion, we compiled briefly present knowledge on prokaryotic 6PGL with regard to genome annotation, and showed that one specific case of missing enzyme was

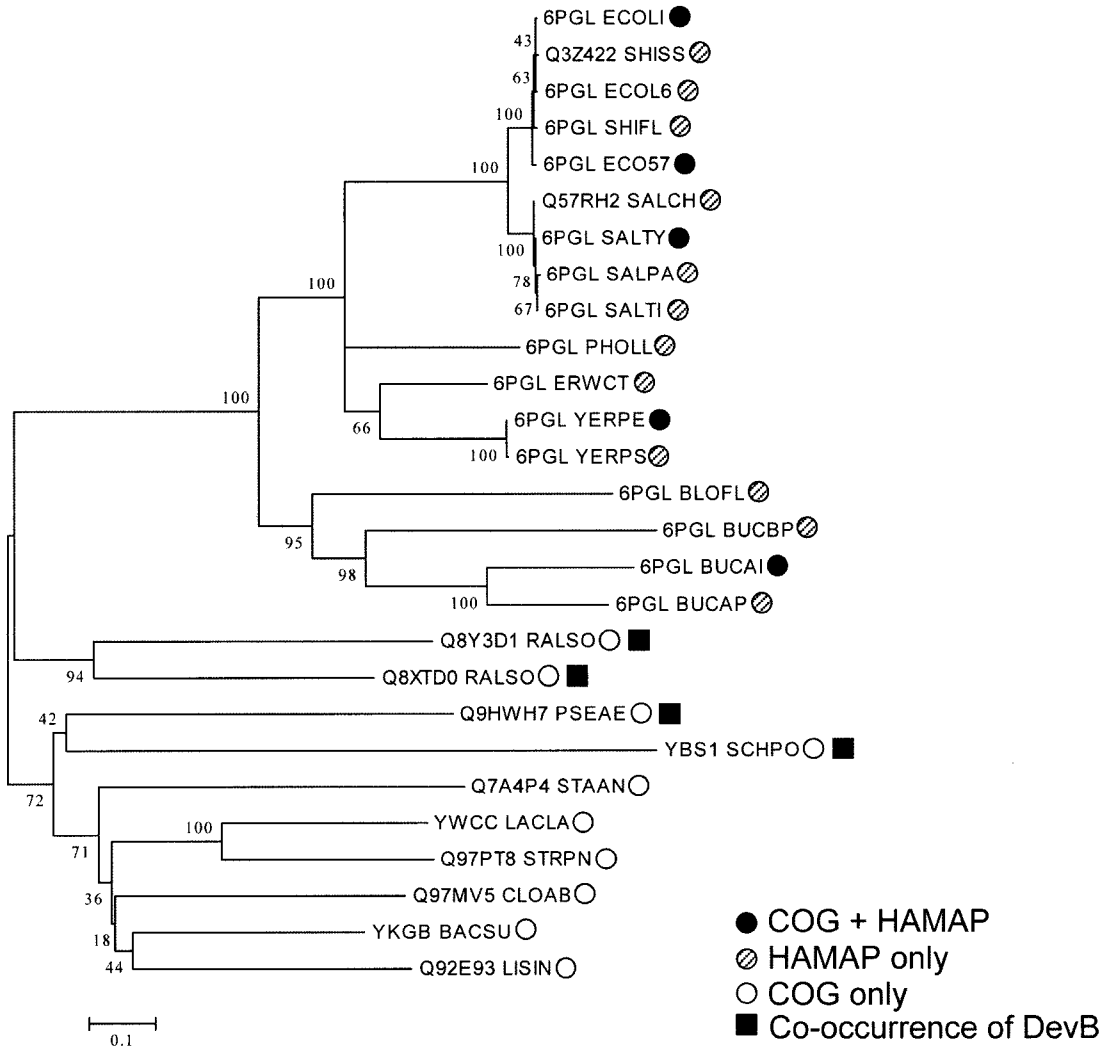

Fig. 1. Phylogenetic tree of 27 proteins in COG2706 and MF_01695. Multiple alignment and tree construction by neighbor joining method was carried out by MEGA (Kumar *et al.*, 2004). Poisson model was used to estimate the evolutionary distances between protein sequences. Bootstrap values for 500 iterations are shown and the scale bar represents number of substitutions per site.

caused by wrongly assigned functional information in the reference database. Considering that 6PGL was brought into focus due to its indispensability in cellular metabolism, it is obvious that there must be more erroneous functional assignments yet undiscovered. To help eliminate error propagation during genome annotation, we should be sure of the authenticity of original functional assignment on the reference sequence. This is important not only for the accuracy of our data, but also for avoiding propagation of errors into the scientific community.

Acknowledgements

This work was supported by the 21C Frontier Microbial Genomics and Applications Center Program, Ministry of Science and Technology, Korea.

References

Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* 15, 132-133.

Collard, F., Collet, J. F., Gerin, I., Veiga-da-Cunha, M., and Van Schaftingen, E. (1999). Identification of the cDNA encoding human 6-phosphogluconolactonase, the enzyme catalyzing the second step of the pentose phosphate pathway(1). *FEBS Lett.* 459, 223-226.

Cordwell, S. J. (1999). Microbial genomes and "Missing" Enzymes: Redefining biochemical pathways. *Arch. Microbiol.* 172, 269-279.

Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genet.* 17, 429-431.

Galperin, M. Y. and Koonin, E. V. (1998). Sources of systematic error in functional annotation of genomes: Domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.* 1, 55-67.

Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A. H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C. J., Lachaize, C., Veuthey, A. L., Gasteiger, E., and Bairoch, A.(2003). Automated annotation of microbial proteomes in Swiss-Prot. *Comput. Biol. Chem.* 27, 49-58.

Gaudermann, P., Vogl, I., Zientz, E., Silvar, F. J., Moya, A., Gross, R., and Dandeka, T. (2006). Analysis of and function predictions for previously conserved hypothetical or putative proteins in *Blochmannia floridanus*. *BMC Microbiol.* 6, 1.

Haft, D. H., Selengut, J. D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res.* 31, 371-373.

Hager, P. W., Calfee, M. W., and Phibbs, P. V. (2000). The *Pseudomonas aeruginosa* devB/SOL homolog, pgl, is a member of the hex regulon and encodes 6-phosphogluconolactonase. *J. Bacteriol.* 182, 3934-3941.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* 34, D354-357.

Kumar, S., Tamura, K., and Nei, M. (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* 5, 150-163.

Lorite, M. J., Sanjuan, J., Velasco, L., Olivares, J., and Bedmar, E. J. (1998). Characterization of *Bradyrhizobium japonicum* pcaBDC genes involved in 4-hydroxybenzoate degradation. *Biochim. Biophys. Acta* 1397, 257-261.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Marchler, G. H., Mullokandov, M., Shoemaker, B. A., Simonyan, V., Song, J. S. *et al.* (2005). CDD: A conserved domain database for protein classification. *Nucleic Acids Res.* 33, D192-196.

Mazur, P., Henzel, W. J., Mattoo, S., and Kozarich, J. W. (1994). 3-carboxy-cis,cis-muconate lactonizing enzyme from *Neurospora crassa*: An alternate cycloisomerase motif. *J. Bacteriol.* 176, 1718-1728.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R.,Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J. *et al.* (2005). InterPro, progress and status in 2005. *Nucleic Acids Res.* 33, D201-205.

Murakami, S., Kohsaka, C., Okuno, T., Takenaka, S., and Aoki, K. (2004). Purification, characterization, and gene cloning of cis,cis-muconate cycloisomerase from benzamide-assimilating *Arthrobacter* sp. Ba-5-17. *FEMS Microbiol. Lett.* 231, 119-124.

Natale, D. A., Shankavaram, U. T., Galperin, M. Y., Wolf, Y. I., Aravind, L., and Koonin, E. V. (2000). Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol.* 1, RESEARCH0009.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4, 41.

Thomason, L. C., Court, D. L., Datta, A. R., Khanna, R., and Rosner, J. L. (2004). Identification of the *Escherichia coli* K-12 ybhE gene as pgl, encoding 6-phosphogluco-

nolactonase. *J. Bacteriol.* 186, 8248-8253.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. (2006). The Universal Protein Resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res.* 34, D187-191.

Zimenkov, D., Gulevich, A., Skorokhodova, A., Biriukova, I., Kozlov, Y., and Mashko, S. (2005). *Escherichia coli* ORF ybhE is pgl gene encoding 6-phosphogluconolactonase (EC 3.1.1.31) that has no homology with known 6PGLs from other organisms. *FEMS Microbiol. Lett.* 244, 275-280.