ISSN (Online) 2287-3406
Journal of Life Science **2022** Vol. 32. No. 6. 463~467 DOI : https://doi.org/10.5352/JLS.2022.32.6.463

- Note -

Conservative Genes among 1,309 Species of Prokaryotes

Dong-Geun Lee*

Department of Pharmaceutical Engineering, College of Medical and Life Science, Silla University, Busan 617-736, Korea

Received April 1, 2022 /Revised June 9, 2022 /Accepted June 16, 2022

As a result of applying the COG (Cluster of Orthologous Groups of Protein) algorithm to 1,309 species to confirm the conserved genes of prokaryotes, ribosomal protein S11 (COG0100) was identified. The numbers of conservative genes were 2, 5, 5, and 6 in 1,308, 1,307, 1,306, and 1,305 species, respectively. Twenty-nine genes were conserved in over 1,302 species, and they encoded 23 ribosomal proteins, 3 tRNA synthetases, 2 translation factors, and 1 RNA polymerase subunit. Most of them were related to protein production, suggesting the importance of protein expression in prokaryotes. The highest conservative COG was COG0048 (ribosomal protein S12) among the 29 COGs. The 29 conserved genes usually have one protein for each prokaryote. COG0090 (ribosomal protein L2) had not only the lowest conservation value but also the largest standard deviation of phylogenetic distance value. As COG0090 is not only a member of the ribosome, but also a regulator of replication and transcription, it could be inferred that prokaryotes have large variations in COG0090 to survive in various environments. This study could provide data necessary for basic science, tumor control, and development of antibacterial agents.

Key words : COG (Cluster of Orthologous Groups of proteins), conservative gene, ortholog, prokaryote

서 론

지구에 출현한 최초의 생명체는 고세균이며 뒤이어 진정세균이 출현하였다. 원핵생물은 현재도 식물 다음의 바이오매스 생성량과 다양한 기능으로 지구 생태계의 유지에 중요한 역할을 한다[3]. 지구환경은 원핵생물의 출현 후에도 변화하였을 것이며, 각자의 환경에 맞추어 원핵생물의 유전자들도 변화하였을 것이다. 각 생물종들의 공통 조상에 존재하던 공통 조상 유전자(ancestral gene)는 복제과정의 실수로 변이도 하였을 것이며, 종분화(speciation) 후에 환경에 맞추어 사라지거나 생기기도 하였을 것이다[13]. 공통 조상 유전자에서 유래한 유전자들이 현재의 생명체들에 분포한다면, 현재 지구의 환경에서 이 유전자들은 생명활동에 공통적으로 필요한 것일 수 있다. 모든 생명체에 공통적인 유전자와 각 분류 단위의 독특한 유전자들에 대한 이해가 현재의 생명체들을 이해하는데 필요하다[7].

***Corresponding author**
Tel : +82-51-999-6282, Fax : +82-51-999-5636
E-mail : ldg@silla.ac.kr

공통 조상 유전자에서 종분화로 서로 다른 종(species)들에 분포한 유전자들의 집합을 ortholog라고 하며, 같은 ortholog에서 발현되는 단백질들의 집합으로 하나의 공통 조상 유전자에서 유래하여 3가지 이상의 생물에 분포하는 단백질들의 집합을 COG (Cluster of Orthologous Groups of protein)로 정의한다[7]. 동일 ortholog에서 유래한 단백질들은 서열이 비슷하고 기능이 동일하므로[7] 게놈서열을 확보하면 COG 방법을 이용하여 유전자 파악과 ortholog 파악이 가능하며 유전체의 비교 등이 가능하다[7]. COG를 이용하여 711개의 원핵생물[13]과 동일한 속(genus)에 속하는 원핵생물[14]의 보존적 유전자 등이 보고되었고, 2020년에 업데이트된 COG 자료를 바탕으로 원핵생물 1,309종에 분포된 4,877개의 COGs도 분석되었다[15]. 이전의 COG database는 711개의 원핵생물을 이용하였고 동일한 종의 원핵생물들도 존재하였다[13]. 최신의 COG database는 각 종에 하나의 원핵생물만 이용하고 1,309종으로 원핵생물의 수가 대폭 확대되어[15] 보존적 유전자의 종류와 보존성 등에 변화가 있을 수 있다. 이 연구에서는 2022년 2월까지의 1,309종의 원핵생물 유전체에 존재하는 보존적 유전자들의 종류와 기능 그리고 보존성의 정도를 COG를 이용하여 파악하고자 하였다.

재료 및 방법

재료

2022년 2월 현재 COG database [9]에는 1,309종의 원핵생물 유전체에 존재하는 3,455,853개의 유전자에서 유래한 단백질들이 4,877개의 COGs로 분류되어 있어[8], 1,309종의 원핵생물 유전체가 갖는 공통 유전자 COGs 자료를 확보하였다. 분석에 사용된 원핵생물의 분류와 종의 개수 등은 이전 논문을 참조하였다[15].

아미노산 서열 분석

분석대상 원핵생물에 공통적으로 존재하는 COGs에 속하는 단백질들의 서열을 NCBI의 공개 데이터베이스에서 추출하고, 각 COG에 속하는 아미노산 서열들은 ClustalX (ver. 2.1) 프로그램으로 다중서열비교를 수행한 후 bootstrap NJ method (n=1,000)를 통해 distance value를 담고 있는 '*.phb' 파일을 생성하였다[12]. Dendroscope 프로그램[10]을 이용하여 각 단백질의 distance value를 구하였고, distance value 자료의 분석과 정리에는 엑셀 프로그램을 이용하였다.

보존적 유전자에서 유래한 단백질의 보존성

1,303종 이상에 공통적으로 존재하는 COGs로 보존적 유전자인 orthologs를 파악하였다. 분석대상 원핵생물에 공통적인 COGs (Table 1)가 나타내는 distance value의 평균과 표준편차로 각 보존적 단백질의 보존성을 분석하였다.

Table 1. The range of species in 1,309 prokaryotes and the number of conservative COGs

Range of species	# of COGs
1,300~1,309	39
1,200~1,299	105
1,100~1,199	224
1,000~1,099	156
900~999	168
800~899	150
700~799	188
600~699	177
500~599	239
400~499	259
300~399	375
200~299	548
100~199	1,015
0~99	1,234
Sum	4,877

결과 및 고찰

보존적 유전자의 분포

원핵생물의 분류단위별로 보유하는 보존적 유전자인 orthologs에서 유래한 COGs의 개수를 파악하였다. 분류단위별의 평균은 497.86개(Mollicutes 문) ~ 1,642.90개(Cyanobacteria 문) 사이였고[15], 각 원핵생물별로 97~2,281개의 보존적 유전자를 가졌다(자료미제시). Table 1에 원핵생물의 종의 수에 따른 보존적 COGs의 개수를 나타내었다. 1,309종의 원핵생물에 4,877개의 COGs가 존재하는데 199종 이하의 원핵생물에 분포하는 보존적 COGs가 2,249로 전체 4,877개의 46.1%였고, 1,300종 이상의 원핵생물에 전체 COGs의 0.8%인 39개의 COG가 보존적이었다. 종의 수가 증가함에 따라 보존적 COGs의 개수가 감소하였다.

1,309종의 원핵생물을 대상으로 보존적 COGs를 파악한 후, 1,303종 이상의 원핵생물에 보존적인 COGs의 종류와 각 COG에 속하는 단백질의 수를 Table 2에 나타내었다. 1,303종 이상의 세균에 보존적인 COGs는 29개였다. 보존적 COGs의 수는 43종과 66종의 미생물에서 각각 72개[11]와 62개[16]였고, 711개의 원핵생물에서 1개[13]였다. 본 연구에서는 1,309종의 원핵생물에서 1개의 보존적 COG가 파악되었다. 29개의 보존적 COGs를 모두 보유한 원핵생물은 1,239종이었고, 50종이 28개, 11종이 27개, 6종이 26개, 2종이 25개, 1종이 24개를 가졌다(자료미제시). 전체 COGs 개수가 97개로 가장 적은 *Candidatus* Nasuia deltocephalinicola NAS-ALF와 2,281개로 가장 많은 *Metakosakonia* MRY16-398는[15] 1,303종 이상에 보존적인 COGs가 각각 25개와 29개로, 전체 COG 개수 대비 25.77%와 1.27%였다.

1,303종 이상의 원핵생물에 보존적인 각 COG를 구성하는 단백질의 수를 원핵생물의 수로 나눈 비율은 COG0008 (Glutamyl- or glutaminyl-tRNA synthetase)이 2.008, COG0480 (Translation elongation factor EF-G, a GTPase)이 1.610으로 높았고, 리보솜을 형성하는 각 COG들은 거의 원핵생물 한 종에 단백질 하나였다(Table 2). 각 원핵생물당 하나의 단백질을 갖는 COG는 COG0096, COG0090, COG0092, COG0185, COG0256, COG0081이었다(Table 2). 1,308종의 원핵생물에 공통적인 COG0049와 COG0051은 각각 α-Proteobacteria 강의 *Methylocystis* sp. SC2와 δ-Proteobacteria 강의 *Desulfovibrio vulgaris* Hildenborough에만 없었다.

원핵생물 711개 연구에서 COG0080 (Ribosomal protein L11)이 모든 원핵생물에 분포하고 COG0100 (Ribosomal protein S11)은 703개의 원핵생물에 분포하였는데[13], 본 연구에서는 COG0100이 1,309종에 그리고 COG0080

Table 2. Conserved COGs in more than 1,302 out of 1,309 species of prokaryotes

# of prokaryote	COG ID	COG name (Function)	# of proteins
1,309	COG0100	Ribosomal protein S11	1,312
1,308	COG0049	Ribosomal protein S7	1,313
	COG0051	Ribosomal protein S10	1,342
1,307	COG0088	Ribosomal protein L4	1,309
	COG0094	Ribosomal protein L5	1,310
	COG0096	Ribosomal protein S8	1,307
	COG0098	Ribosomal protein S5	1,308
	COG0197	Ribosomal protein L16/L10AE	1,308
1,306	COG0008	Glutamyl- or glutaminyl-tRNA synthetase	2,623
	COG0080	Ribosomal protein L11	1,317
	COG0090	Ribosomal protein L2	1,306
	COG0093	Ribosomal protein L14	1,307
	COG0186	Ribosomal protein S17	1,308
1,305	COG0048	Ribosomal protein S12	1,308
	COG0087	Ribosomal protein L3	1,308
	COG0092	Ribosomal protein S3	1,305
	COG0097	Ribosomal protein L6P/L9E	1,306
	COG0099	Ribosomal protein S13	1,313
	COG0185	Ribosomal protein S19	1,305
1,304	COG0103	Ribosomal protein S9	1,309
	COG0172	Seryl-tRNA synthetase	1,374
	COG0180	Tryptophanyl-tRNA synthetase	1,422
	COG0202	DNA-directed RNA polymerase, alpha subunit/40 kD subunit	1,367
	COG0256	Ribosomal protein L18	1,304
	COG0361	Translation initiation factor IF-1	1,456
	COG0480	Translation elongation factor EF-G, a GTPase	2,100
	COG0522	Ribosomal protein S4 or related protein	1,362
1,303	COG0081	Ribosomal protein L1	1,303
	COG0244	Ribosomal protein L10	1,304

은 1,306종에 분포하였다. 이것은 COG 자료의 최신 업데이트에서 동일한 종의 원핵생물은 하나만 남기는 과정에서 COG0100이 없는 원핵생물들이 제거된 결과 등으로 판단되었다.

보존적 유전자의 기능

Table 2에 연구대상 1,309종의 원핵생물 중에서 1,303종 이상의 원핵생물에서 발견되는 29개의 COGs들의 기능을 나타내었다. 29개 중에서 리보솜 구성에 관계된 것이 23개로 최대였고 tRNA synthetase 3개, 전사와 번역 개시 및 번역 지속 관련 COG가 각 1개였다. 1,309종의 원핵생물 모두에 공통적인 COG는 COG0100 (Ribosomal protein S11) 1개였고, 1,308 종의 원핵생물에 2개, 1,307 종의 원핵생물에 5개 등 8개의 COGs는 모두 리보솜 구성에 관계된 COGs였다. 다른 연구들도 보존적 COGs중에서 ribosomal subunit를 구성하는 COG들의 보존 비율이 높았다[13, 16].

리보솜을 구성하는 단백질들은 리보솜 구성 단백질 이외의 기능도 수행하는데[2] 1,309종의 원핵생물 모두에서 발견되는 ribosomal protein S11은 topoisomerase II를 목표로 하는 뇌종양 등의 암치료제에 반응을 하며[1], 바이러스의 증식과 감염 등과도 연관이 있다[19]. 1,308종의 원핵생물에서 발견되는 COG0049 (Ribosomal protein S7)은 과발현되면 p53을 통해 암세포의 세포자살을 유도한다[4]. 따라서 COG에 대한 연구를 통해 항생제 개발과 함께 종양의 치료 등에도 사용할 수 있을 것이다. COG0172 (Seryl-tRNA synthetase)를 대상으로 하는 항생물질[18]과 함께 seryl-tRNA synthetase는 상피세포의 카제인 합성과도 연관이 있다[6]. COG0480은 ribosome-interacting GTPase로 리보솜 50S 소단위체의 구조 변화에 관여하여 리보솜의 생성 및 세포의 생장 조절에도 관여한다[20].

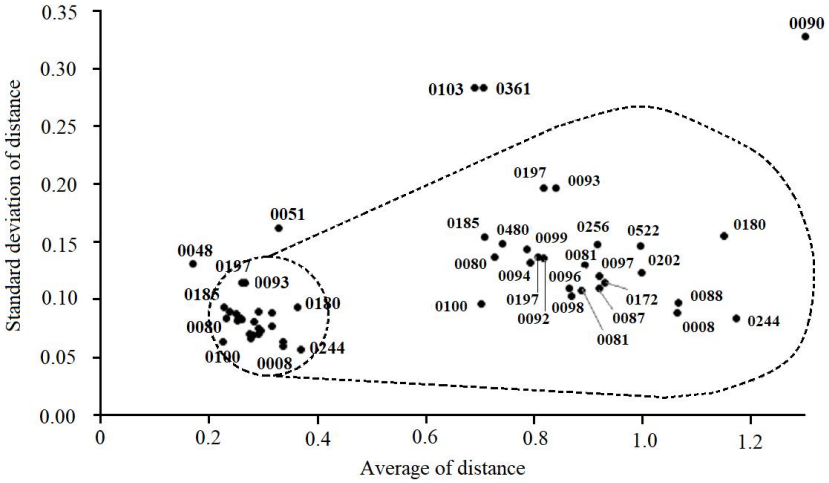

Fig. 1. Average and standard deviation of distance values obtained from phylogenetic trees of 29 COGs distributed among 1,309 prokaryotes. Inside the dotted line is the enlarged part of the cluster of dots and prefix COG prior to each number was omitted.

보존적 유전자에서 유래한 단백질의 보존성

분석대상 모든 원핵생물에서 발견되는 COG0100은 Clostridia 강의 *Caldanaerobacter subterraneus* MB4와 *Carboxydocella thermautotrophica* 41 그리고 α-Proteobacteria 강의 *Croceicoccus marinus* E4A9에서만 단백질이 2개씩 발견되고 나머지 1,306종의 원핵생물에서는 각 하나만 존재하였다. 2개씩의 단백질이 있을 때 서열정렬로 최소의 distance value를 보이는 하나만 선택하여 보존성을 파악하였다.

Fig. 1에 1,303종 이상의 원핵생물에 보존적인 각 COG의 구성원들의(Table 2) 아미노산 서열로 계통수를 작성할 때 생성되는 distance value들의 평균과 표준편차로 각 COG의 분포를 나타내었다. Fig. 1에서 점선부분은 밀집된 부분을 확대한 것이다. 낮은 평균은 각 COG를 구성하는 단백질들의 아미노산 서열들 사이가 유사성이 높다는 것이며, 높은 표준편차는 서열의 유사성이 낮은 것들과 높은 것들이 넓게 분포한다는 것으로 진화과정에서 각 원핵생물별로 혹은 분류단위의 구성원 별로 변이 속도가 다르거나 수평적 유전자 전달의 사례로 해석할 수도 있다. 분석대상 29개의 COGs 중에서 보존성이 가장 높은 COG는 COG0048 (Ribosomal protein S12) 이었고, 보존성이 가장 낮은 COG는 COG0090 (Ribosomal protein L2) 이었다. Ribosomal protein L2는 대장균에서 DNA 복제와 전사의 조절자로도 역할을 하는데[5, 17] 다양한 환경에 존재하는 원핵생물에서 조절기능을 한다면, 각 환경에 맞게 적응되어 아미노산 서열 차이가 클 것이므로 높은 평균 및 표준편차를 나타내는 것과 관계가 있을 것이다. COG0244 (Ribosomal protein L10)는 가장 낮은 표준편차를 보였다. 표준편차가 낮은 것은 서열들의 유사성이 높거나 혹은 낮거나 하면서 distance value가 상대적으로 일정한 수치에 밀집되어 있다는 것이다.

본 연구에서 파악된 29개의 보존적 COGs는 1,309종의 원핵생물이 다양한 환경에 분포하더라도, 현재 존재하는 원핵생물의 생명현상에 공통적으로 필요한 것으로 판단된다. 현재의 생물들에 보존적 유전자들은 원시생명체가 종분화(speciation)하기 전부터 존재하였거나, 생존 환경이 바뀜에 따라 유전자를 추가 혹은 제거하면서 적응되었거나, COG 구성원이 아닌 다른 유전자에 의한 유전자의 기능대체현상(gene displacement)에 의한 것으로 토의되었다[13]. 현재의 보존적 유전자는 현재 지구 환경에서 생명현상에 필요한 유전자로 유추되었다. 본 연구는 원핵생물의 진화 과정에서 보존된 유전자들의 종류와 기능 등 기초적인 자료의 제공과 함께 항균제 및 종양 연구와 약물개발[1, 4]에도 활용할 수 있을 것이다.

The Conflict of Interest Statement

The authors declare that they have no conflicts of interest with the contents of this article.

References

1. Awah, C. U., Chen, L., Bansal, M., Mahajan, A., Winter, J., Lad, M., Warnke, L., Gonzalez-Buendia, E., Park, C., Zhang, D., Feldstein, E., Yu, D., Zannikou, M., Balyasnikova, I. V., Martuscello, R., Konerman, S., Győrffy, B., Burdett, K. B., Scholtens, D. M., Stupp, R., Ahmed, A., Hsu, P. and Sonabend, A. M. 2020. Ribosomal protein S11 influences glioma response to TOP2 poisons. *Oncogene* **39**, 5068-5081.
2. Ba, Q., Li, X., Huang, C., Li, J., Fu, Y., Chen, P., Duan, J., Hao, M., Zhang, Y., Li, J., Sun, C., Ying, H., Song, H., Zhang, R., Shen, Z. and Wang, H. 2017. BCCIPβ modulates the ribosomal and extraribosomal function of S7 through a direct interaction. *J. Mol. Cell Biol.* **9**, 209-219.

3. Bar-On, Y. M., Phillips, R. and Milo, R. 2018. The biomass distribution on earth. *PNAS.* **115**, 6506-6511.
4. Chen, D., Zhang, Z., Li, M., Wang, W., Li, Y., Rayburn, E. R., Hill, D. L., Wang, H. and Zhang, R. 2007. Ribosomal protein S7 as a novel modulator of p53-MDM2 interaction: binding to MDM2, stabilization of p53 protein, and activation of p53 function. *Oncogene* **26**, 5029-5037.
5. Chodavarapu, S., Felczak, M. M. and Kaguni, J. M. 2011. Two forms of ribosomal protein L2 of *Escherichia coli* that inhibit DnaA in DNA replication. *Nucleic Acids Res*. **39**, 4180-4191.
6. Dai, W., Zhao, F., Liu, J. and Liu, H. 2020. Seryl-tRNA synthetase is involved in methionine stimulation of β-casein synthesis in bovine mammary epithelial cells. *Br. J. Nutr*. **123**, 489-498.
7. Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I. and Koonin, E. V. 2019. Microbial genome analysis: the COG approach. *Brief Bioinform*. **20**, 1063-1070.
8. Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Alvarez, R. V., Landsman, D. and Koonin, E. V. 2020. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res*. **49**, D274-D281.
9. https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/
10. Huson, D. H. and Scornavacca, C. 2012. Dendroscope 3 : An interactive viewer for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061-1067.
11. Kang, H. Y., Shin, C. J., Kang, B. C., Park, J. H., Shin, D. H., Choi, J. H., Cho, H. G., Cha, J. H., Lee, D. G., Lee, J. H., Park, H. K. and Kim, C. M. 2002. Investigation of conserved gene in microbial genomes using *in silico* analysis. *J. Life Sci.* **5**, 610-621.
12. Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press.
13. Lee, D. G. and Lee, S. H. 2015. Investigation of conservative genes in 711 prokaryotes. *J. Life Sci.* **9**, 1007-1013.
14. Lee, D. G. and Lee, S. H. 2019. Conserved genes and metabolic pathways in prokaryotes of the same genus. *J. Life Sci.* **1**, 123-128.
15. Lee, D. G. and Lee, S. H. 2021. Investigation of COGs (Clusters of Orthologous Groups of proteins) in 1,309 species of prokaryotes. *J. Life Sci.* **9**, 834-839.
16. Lee, D. G., Lee, J. H., Lee, S. H., Ha, B. J., Kim, C. M., Shim, D. H., Park, E. K., Kim, J. W., Li, H. Y., Nam, C. S., Kim, N. Y., Lee, E. J., Back, J. W. and Ha, J. M. 2005. Investigation of conserved genes in microorganism. *J. Life Sci.* **15**, 261-266.
17. Rippa, V., Cirulli, C., Di Palo, B., Doti, N., Amoresano A. and Duilio, A. 2010. The ribosomal protein L2 interacts with the RNA polymerase alpha subunit and acts as a transcription modulator in *Escherichia coli*. *J. Bacteriol.* **192**, 1882-1889.
18. Saha, A., Dutta, S. and Nandi, N. 2020. Inhibition of seryl tRNA synthetase by seryl nucleoside moiety (SB-217452) of albomycin antibiotic. *J. Biomol. Struct. Dyn.* **38**, 2440-2454.
19. Wang, R., Du, Z., Bai, Z. and Liang, Z. 2017. The interaction between endogenous 30S ribosomal subunit protein S11 and Cucumber mosaic virus LS2b protein affects viral replication, infection and gene silencing suppressor activity. *PLoS One* **12**, e0182459.
20. Zhang, X., Yan, K., Zhang, Y., Li, N., Ma, C., Li, Z., Zhang, Y., Feng, B., Liu, J., Sun, Y., Xu, Y., Lei, J. and Gao, N. 2014. Structural insights into the function of a unique tandem GTPase EngA in bacterial ribosome assembly. *Nucleic Acids Res*. **42**, 13430-13439.

초록 : 원핵생물 1,309종의 보존적 유전자

이동근*
(신라대학교 제약공학과)

원핵생물 1,309종(species)에 보존적인 유전자(ortholog)를 파악하기 위해 1,309종을 대상으로 COG (Cluster of Orthologous Groups of proteins) 기법을 적용하였으며, 그 결과 ribosome protein S11 (COG0100)을 확인하였다. 1,308, 1,307, 1,306 및 1,305종에서 보존된 ortholog의 수는 각각 2, 5, 5 및 6개였다. 1,303종 이상에서 보존된 유전자는 29개였고, 이들은 23개의 리보솜 단백질, 3개의 tRNA 합성효소, 2개의 번역 인자 및 1개의 RNA 중합효소 소단위체 유전자였다. 대부분이 단백질 합성과 연관되어 원핵생물에서 단백질 발현이 중요한 것으로 판단되었다. 29개의 COG 중에서 ribosome protein S12 (COG0048)가 보존성이 가장 높았다. 29개의 보존된 COG는 대개 하나의 원핵생물에 하나의 단백질이 분포하였다. COG0090은 보존성이 가장 낮았으며 phylogenetic distance value의 표준편차도 가장 컸다. COG0090은 리보솜의 구성원 기능 외에 복제와 전사의 조절자 역할을 하기에, 각 원핵생물이 다양한 환경에서 생존하기 위해 변이가 큰 것으로 추론되었다. 이 연구는 기초 과학과 종양 조절 및 항균제 개발에 필요한 데이터를 제공할 수 있을 것이다.