ISSN (Online) 2287-3406
Journal of Life Science **2021** Vol. 31. No. 9. 834~839 DOI : https://doi.org/10.5352/JLS.2021.31.9.834

Investigation of COGs (Clusters of Orthologous Groups of proteins) in 1,309 Species of Prokaryotes

Dong-Geun Lee and Sang-Hyeon Lee*

Department of Pharmaceutical Engineering, Silla University, Busan 617-736, Korea

Received August 2, 2021 /Revised August 19, 2021 /Accepted September 17, 2021

Authors previously reported the results of analyses of COGs (Clusters of Orthologous Groups of proteins) in 711 prokaryotes. The data of COGs were significantly updated for 2020 using 1,309 prokaryotic genomes. Here, we report the results of analyses of 3,455,853 proteins comprising 4,877 updated COGs in terms of COGs and prokaryotes. The numbers of COGs in each prokaryote ranged from 97 to 2,281, with an average of 1,430.0 and a standard deviation of 414.2. Mean numbers of COGs at the phylum level were minimal 497.86 for Mollicutes and maximal 1,642.90 for Cyanobacteria. The top 10 species with the highest COG retention numbers were all Proteobacteria, and 9 out of the bottom 10 were those that could not be cultured *in vitro*. The numbers of proteins belonging to each COG ranged from 2 to 22,048, with over 12,000 proteins up to the top 11. Five of the top 11 were COGs that bind to DNA and were involved in the gene expression, indicating the importance of regulating gene expression in prokaryotes in a changing environment. COG data are expected to be widely utilized as they can be used for the identification of genes included in the genome and the selection of genes for the strain improvement.

Key words : Clusters of Orthologous Groups of proteins, COGs, prokaryotes, prokaryotic genomes

서 론

생명체는 변화하는 환경에 맞추어 생명현상을 조절하며, 유전자의 유무와 발현 여부가 생존에 중요하다. 생명체가 지구에 나타난 이후 생물들은 유전자들을 얻거나 잃으면서 변화하는 환경에 적응하였을 것이다[1]. Fraser 등은 상대적으로 미시적 환경의 변화가 적고, 그러한 환경에서 살아가는 미생물들은 그들 조상과의 유전적 공통점을 오래 유지한 것으로 파악하였다[1].

공통조상 유전자(ancestral gene)는 종(species)들의 공통조상이 보유하던 것이며 종분화(speciation) 혹은 복사(duplication)로 각 종의 유전체에 분포한다[17]. 공통조상의 유전자가 종분화로 생물 종들에 분포할 때 이런 유전자들의 집합을 ortholog라고 하며, 동일한 ortholog의 구성원들은 서열이 유사하고 생성된 단백질의 기능이 동일하다[17]. Ortholog가 유전자의 집합인데 비해, COG (Cluster of Orthologous Groups of proteins)는 동일한 ortholog에서 발현된 단백질의 집합으로 동일한 기능과 유사한 구조를 갖는다[17]. 최초로 1997년도에 7종류의 게놈에서 유래한 720개의 COG가 보고된 후, 2014년도에 711종류의 게놈을 이용한 4,631개의 COG가 보고되었다[2]. Lee와 Lee는 2014년도의 COG를 기준으로 711종의 원핵생물 그리고 동일한 속(genus) 원핵생물의 보존적 유전자 등을 보고하였다[8, 9]. 한편 2020년도에 1,187개의 진정세균과 122개의 고세균 게놈 등 1,309개의 게놈을 이용한 데이터베이스로 업데이트되면서 200개 이상의 COG가 추가되어, 현재 총 4,877개의 COG가 있다[14]. 이 논문에서는 COG와 원핵생물 측면에서 업데이트된 4,877개의 COGs를 구성하는 3,455,853개의 단백질들에 대한 분석 결과를 보고한다.

재료 및 방법

재료

원핵생물의 각 유전체가 가지는 COG에 관한 데이터베이스는 2020년의 COG에 정리된 데이터베이스를 이용하였다[4]. 각 원핵생물이 가지는 COG 데이터베이스를 확보한 후, 1,309개의 원핵생물 모두가 보유하는 COG의 종류들을 파악하였다. Ftp에서 데이터베이스들을 다운로드 받은 후, 데이터베이스들의 누락을 확인하였다[4]. 이들은 2021년 7월 현재 1,309종의 원핵생물 유전체에 존재하는 3,455,853개의 유전자들이 4,877개의 COG 종류로 구성되어 있었다[14]. Table 1은 실제로 분석한 1,309종의 원핵생물을 문(phylum) 수준에서 각 문을 구성하는 종(species)의 개수를 나타내고 있다. Firmicutes와 Proteobacteria 문은 강(class) 수준으로 정리하였다.

***Corresponding author**
Tel : +82-51-999-5624, Fax : +82-51-999-5628
E-mail : slee@silla.ac.kr

Table 1. Numbers of studied species and numbers of COGs derived from the COGs database

Phylum Class	# of species	COG proteins / kinds	# of COGs				
			Average	Standard Deviation	Minimum	Maximum	Sum
Archaebacteria							
Crenarchaeota	25	1.45	947.08	414.21	783	1,119	23,677
Euryarchaeota	79	1.66	1,169.47	170.00	557	1,521	92,388
Thaumarchaeota	12	1.43	902.08	78.43	811	1,026	10,825
Other Archaea	6	1.32	661.33	511.98	332	1,289	3,968
Eubacteria							
Acidobacteria	7	2.18	1,638.86	267.11	1,403	1,908	11,472
Actinobacteria	155	2.02	1,399.35	290.20	586	1,979	216,899
Aquificae	9	1.31	1,119.89	128.41	1,042	1,229	10,079
Bacteroidetes	107	1.84	1,322.07	303.17	172	1,772	141,462
Chlamydiae	6	1.33	893.17	248.56	594	1,100	5,359
Chlorobi	5	1.53	1,300.40	50.40	1,188	1,378	6,502
Chloroflexi	14	2.02	1,385.57	318.39	864	1,807	19,398
Cyanobacteria	41	2.06	1,642.90	228.57	828	1,934	67,359
Deferribacteres	5	1.54	1,397.00	83.39	1,331	1,481	6,985
Deinococcus-Thermus	6	1.55	1,407.33	180.58	1,283	1,545	8,444
Firmicutes							
Bacilli	73	1.64	1,487.14	303.26	818	2,007	108,561
Clostridia	79	1.68	1,413.38	191.08	804	1,888	111,657
Negativicutes	10	1.56	1,352.00	289.46	934	1,877	13,520
Tissierellia	9	1.45	1,116.89	221.96	893	1,472	10,052
Other Firmicutes	4	1.55	1,171.75	265.33	944	1,602	4,687
Fusobacteria	6	1.50	1,140.83	328.74	770	1,478	6,845
Mollicutes	14	1.32	497.86	344.62	296	1,014	6,970
Planctomycetes	14	2.11	1,515.93	177.69	1,105	1,775	21,223
Proteobacteria							
Alpha-Proteobacteria	158	1.88	1,639.23	403.75	123	2,215	258,998
Beta-Proteobacteria	102	1.79	1,608.88	431.22	97	2,220	164,106
Delta-Proteobacteria	39	2.20	1,635.62	269.28	941	2,103	63,789
Epsilon-Proteobacteria	12	1.44	1,213.92	119.22	918	1,510	14,567
Gamma-Proteobacteria	224	1.64	1,642.92	463.29	152	2,281	368,015
Other Proteobacteria	6	1.59	1,394.50	329.87	1,309	1,462	8,367
Spirochaetes	11	1.67	1,162.36	330.83	574	1,588	12,786
Synergistetes	5	1.50	1,252.00	99.95	1,162	1,378	6,260
Thermotogae	9	1.48	1,157.44	120.94	1,113	1,194	10,417
Verrucomicrobia	9	1.72	1,320.78	305.34	647	1,648	11,887
Other eubacteria	48	1.46	923.98	432.79	141	1,744	44,351

원핵생물별 보유 COG 종류의 수와 COG에 속하는 단백질의 수

Perl과 엑셀 프로그램을 이용하여 전체 3,455,853개의 COG에 속하는 단백질들을 분류하여 분석대상 1,309종의 원핵생물 게놈 각각에 대하여 어떤 종류의 COG를 가지는 지, 그리고 원핵생물이 COG에 속하는 단백질들을 몇 개나 가지는 지를 파악하였다.

COG별 단백질 구성원들의 수

Perl과 엑셀 프로그램을 이용하여 전체 3,455,853개의 COG에 속하는 단백질들을 분류하여 4,877개의 각 COG에 속하는 단백질들로 구분하였다.

결과 및 고찰

원핵생물별 보유 COG 종류의 수와 COG에 속하는 단백질의 수

1,309개의 원핵생물에 4,877개의 COG가 분포하였고, 원핵생물마다 보유한 COG 종류의 수와 동일한 COG라도 COG에 속하는 단백질의 수는 원핵생물마다 달랐다. Tatusov 등[17]은 각 분류단위에 독특한 유전자 그리고 모든 생명체에 공통적인 유전자에 대한 이해로 현재 지구상에 존재하는 생명을

이해할 수 있다고 하였다. 원핵생물이 보유하고 있는 COG 종류의 수는 97개(*Candidatus* Nasuia deltocephalinicola NAS-ALF)에서 2,281개(gamma-Proteobacteria 강의 *Metakosakonia* MRY16-398)의 범위였고, 평균 1,430.0개, 표준편차 414.2개였다. *Candidatus*는 게놈분석 등으로 특성 파악이 어느 정도 되었지만 배양이 불가능한 세균을 명명할 때에 사용된다[16]. 각 생물 혹은 분류단계가 보이는 생명현상은 전체 생물에 공통적인 것과 각 생물 혹은 분류단계에 특이적인 것이 모여 나타나는데, *Metakosakonia* MRY16-398는 다른 원핵생물들과 보이는 공통현상이 많은 것으로 판단할 수 있었다. 각 속(genus)에 공통되는 COG 종류의 수, 각 균주의 보유 COG 종류의 수, 각 균주의 전체 단백질들의 수가 다름이 보고되었다[9].

하나의 원핵생물이 가지는 COG에 속하는 단백질들의 수는 98개(*Candidatus* Nasuia deltocephalinicola NAS-ALF)에서 8527개(Actinobacteria 문의 *Nonomuraea* ATCC 55076)의 범위였고, 평균 2,640.0개 표준편차 1,253.2개였다. *Candidatus* Nasuia deltocephalinicola NAS-ALF는 COG 종류의 수가 94개이고 COG에 속하는 단백질은 98개인데, COG0004 (Ammonia channel protein AmtB)에 속하는 단백질 3개와 COG0006 (Xaa-Pro aminopeptidase)에 속하는 단백질 2개를 제외한 나머지 COG들은 1개의 단백질만을 가졌다. COG에 속하는 단백질들의 총 수를 보유한 COG 종류의 수로 나눈 비율은 1,979 종류의 COG를 가진 *Nonomuraea* ATCC 55076가 4.309로 최대였고, 123 종류의 COG를 가진 *Candidatus* Hodgkinia cicadicola Dsem이 1.008로 최소였다. 각 COG에 속하는 단백질 구성원들의 수(Table 1의 sum)를 분류수준을 구성하는 종들의 수(Table 1의 # of species)로 나눈 분류수준별 COG 구성 단백질 수의 평균은 670.36~3745.38의 범위였다. Mollicutes 문이 최소였고 Proteobacteria 문의 delta-Proteobacteria 강이 최대였다. "COG의 단백질 구성원 수 / COG의 종류"를 각 원핵생물별로 구한 후에 분류수준으로 파악한 Table 1의 COG proteins/kinds의 범위는 1.31~2.20였고, 평균 1.65, 표준편차 0.26이었다. Aquificae 문이 최소이고 delta-Proteobacteria 강이 최대였다. 하나의 원핵생물이 동일한 COG에 단백질, 즉 유전자가 많으면 높은 단백질량을 발현시킬 수 있거나 하나의 유전자가 문제가 생겨도 다른 유전자가 기능을 할 수 있으므로 돌연변이에 대한 저항성이 올라갈 것이다.

Table 1에는 분석된 원핵생물의 수를 문(phylum) 혹은 강(class) 분류수준에 속하는 원핵생물의 수로 표시하였으며, 각 분류수준이 가지는 COG 종류의 수를 나타내었다. 각 분류수준을 구성하는 원핵생물의 수가 서로 다르지만 각 분류수준이 갖는 COG 종류의 평균으로 비교하면, Mollicutes 문이 497.86개로 최소였고, Cyanobacteria 문이 1,642.90개로 최대였다. Lee와 Lee [9]는 유전자 보유 정도가 원핵생물이 각 서식지에 적응하는 정도를 나타내고 원핵생물 진화의 역사 혹은 현재 지구의 원핵생물 서식지 범위를 나타내는 것일 수도 있다고 하였다. 이에 따르면 Mollicutes 문의 서식지보다 Cyanobacteria 문의 서식지가 훨씬 넓다고 할 수 있었다. 각 분류수준 내에서 COG 종류의 범위(최대-최소)는 81~2,184개였고, 평균 849.36개, 표준편차 586.01개였다. 9개의 원핵생물로 구성된 Themotogae 문이 최소였고, 541개의 원핵생물로 구성된 Proteobacteria 문이 최대였다. 각 문을 구성하는 구성원의 수와 COG 종류 개수의 범위를 선형관계로 나타낼 때 결정계수(R^2)가 0.5171로 둘 사이에 큰 관계는 없었다.

분류수준별 COG 구성 단백질 개수의 표준편차를 보면 Chlorobi 문이 50.40으로 최소였고, gamma-Proteobacteria 강이 463.29로 최대였다. Other Archaea는 511.98로 최대였지만 분류가 완전하지 않아 제외하였다. 단순한 표준편차 해석에 오류가 발생할 수 있어, 표준편차를 평균으로 나눈 변동계수(coefficient of variation)를 비교하여 상대적인 산포도를 비교하였다. Other Archaea의 0.77을 제외하고는 Mollicutes 문이 0.69로 최대였고 Chlorobi 문이 0.04로 최소였다. 분류수준별 구성원의 수에 따라 변동계수를 비교하기 위해 (분류수준, 구성원의 수, 변동계수)를 보면 구성원이 많은 순서로 (gamma-Proteobacteria, 224, 0.28), (alpha-Proteobacteria, 158, 0.25), (Actinobacteria, 155, 0.21), (Bacteroidetes, 107, 0.23) 등이었고, 구성원이 적은 순서로 (other Firmicutes, 4, 0.23), (Defembacteres, 5, 0.06), (Chlorobi, 5, 0.04), (Fusobacteria, 6, 0.29) 등이었다. 변동계수의 값이 작으면 자료들이 평균 주위에 많이 분포해 있는데, Fusobacteria 문과 other Firmicutes 문은 구성원의 수가 각각 6, 4개라도 구성원의 수가 각각 155, 107인 Actinobacteria와 Bacteroidetes 문보다 높았다. 변동계수와 구성원의 수 사이의 결정계수(R^2)가 0.0161로, 각 분류수준의 구성원 수와 보유한 COG 종류 수의 표준편차 사이의 관계는 낮았다.

각 생물종이 보유하고 있는 COG 종류의 수는 97~2,281개의 범위였다. COG 종류(원핵생물의 수)를 보면 2,100 종류 이상(42개), 1,800~2,099 종류(217개), 1,500~1,799 종류(339개), 1,200~1,499 종류(378개), 900~1,199 종류(202개), 600~899 종류(62개), 300~599 종류(56개), 299종류 이하(13개) 였다. 1,200~1,799 종류의 COG를 가진 원핵생물의 수가 54.78%로 과반을 넘었다. Table 2에는 각 생물종이 보유하고 있는 COG 종류의 수를 기준으로 상위와 하위 10개의 생물종을 나타내었다. 상위 10위까지 모두 Proteobacteria 문으로 8위까지는 gamma-Proteobacteria 강(class)이고 9위와 10위는 각각 beta-와 alpha-Proteobacteria 강에 속한다. 하위 10개의 생물종에서 9개가 *Candidatus*로 시작되는데, *Candidatus*는 시험관에서 배양이 불가능하다[16]. 2020년 현재의 COG 데이터베이스는 97개의 *Candidatus*가 있으며, Lee [7]는 단독배양이 가능한 원핵생물 중 최소의 게놈이며 367개의 COG를 가진 *Mycoplasma genitalium*보다 적은 수의 COG를 가지는 14개의 원핵생물들을 비교하였는데, *Candidatus*가 13개였고 모두 세포내에서 기생/공

Table 2. The top 10 and bottom 10 organisms by numbers of containing COGs

Top 10		Bottom 10	
# of COGs	Organism	# of COGs	Organism
2281	*Metakosakonia* MRY16-398	97	*Candidatus* Nasuia deltocephalinicola NAS-ALF
2270	*Phytobacter ursingii* CAV1151	123	*Candidatus* Hodgkinia cicadicola Dsem
2266	*Pseudomonas aeruginosa* PAO1	125	*Candidatus* Vidania fulgoroideae OLIH
2262	*Citrobacter freundii* CFNIH1	141	bacterium AB1
2262	*Pluralibacter gergoviae* FB2	152	*Candidatus* Carsonella ruddii DC
2250	*Klebsiella variicola* 342	163	*Candidatus* Tremblaya phenacola PAVE
2228	*Escherichia coli* O157 H7 Sakai	172	*Candidatus* Sulcia muelleri PUNC
2221	*Enterobacter cloacae* ATCC 13047	186	*Candidatus* Zinderia insecticola CARI
2220	*Burkholderia cepacia* ATCC 25416	215	*Candidatus* Uzinura diaspidicola ASNER
2215	*Mesorhizobium japonicum* MAFF 303099	234	*Candidatus* Walczuchella monophlebidarum

생을 한다. 하위 10개와 Lee [7]가 분석한 원핵생물을 비교하면 *Candidatus* Walczuchella monophlebidarum와 bacterium AB1가 추가되었는데, 전자는 식물의 세포내에서 공생을 하고 [15] 후자는 환경시료 유래의 게놈만 알려져 있다.

생물종 1개에 단백질이 하나인 COG는 총 501개로 전체 4,877개 COG의 10.27%로, COG에 따라서 분포된 생물종의 개수는 2~1307의 범위였다(자료미제시). 이들은 경우에 따라 분류에 활용될 수도 있을 것이다.

COG별 단백질 구성원의 수

Table 3에는 4,877개의 COG들을 구성하는 단백질 구성원들의 수를 전체 연구대상 1,309개의 원핵생물로 나눈 비율(R), COG 종류의 수, 그리고 각 COG의 단백질 구성원 수의 범위를 표시하였다. 비율 R을 보면 1 미만인 COG의 수는 4,177개로 전체 4,877개의 85.65%였다. Table 3에서는 1 이하인 R의 구간을 세분화하였다. 즉, R이 0.25 미만인 COG의 종류는 전체 4,877개의 48.60%인 2,370개이며, 0.25 미만인 R을 0.05 단위로 보면 단백질의 개수가 66~130인 COG의 종류가 738개로 최대였다. R이 0.05 이상 0.10 미만이며 66~130개의 단백질을 갖는 COG의 종류가 많다는 의미이다. 각 생물이 보이는 생명현상은 전체 생물에 공통적인 것과 각 생물에 특이적인 것이 모여 나타나는데, 유전자의 관점에서는 공통유전자와 각 생물 특이유전자에 의한다고 할 수 있다[8].

Table 4에는 각 COG가 보유하고 있는 단백질들의 수를 기준으로 상위와 하위 11개의 COG를 나타내었다. 4,877개의 각 COG별 단백질 구성원들의 수는 COG5307이 2개로 최소였고 COG0583이 22,048개로 최대였으며, COG의 종류에 따라 차이가 매우 많았다. 평균은 708.60개였고, 표준편차는 1,269.79개였다.

상위 11위까지는 12,000개 이상의 단백질 구성원이 있었고 가장 많은 수의 단백질을 보유한 COG0583은 전사조절인자의 LysR 계열이고, 1,068개의 세균에서 22,048개의 단백질이 존재한다. LysR은 전사인자의 한 종류이며 LTTR (LysR type transcriptional regulator family)은 negative auto-regulation에 관여하는 전사촉진자이다[11]. Table 2의 11개 COG 중에서 COG0583, COG1309, COG 0745, COG1595, COG2207 등 5개의 COG가 DNA에 결합하여 유전자 발현에 관여하는데, 세균이 변화하는 환경에서 생존을 위해 다양하고 정교하게 유전자 발현을 조절해야 하므로 이와 관련된 조절인자가 많아야 할 것인데, 이것이 해당 COG의 구성 단백질들이 많은 것과 연관이 있을 것으로 사료된다.

COG5307의 구성 단백질이 2개로 최소였는데, 기능은 Guanine-nucleotide exchange factor YEL1 (contains Sec7 domain)이다. Alpha-Proteobacteria 강(class)의 *Rickettsia prowa-*

Table 3. The counts of COGs and range of numbers of proteins according to ratio of the numbers of proteins in 1,309 bacteria

Ratio (R) N/1309	Count of COGs	Range of numbers of protein (N)
10≦R	9	13,192~22,048
9≦R<10	2	12,226~12,355
8≦R<9	1	10,835
7≦R<8	8	9,235~10,449
6≦R<7	6	7,974~9,032
5≦R<6	10	6,860~7,839
4≦R<5	15	5,257~6,444
3≦R<4	46	3,978~5,214
2≦R<3	118	2,623~3,864
1≦R<2	485	1,309~2,606
0.75≦R<1	438	982~1,308
0.5≦R<0.75	437	655~981
0.25≦R<0.5	795	328~654
0.20≦R<0.25	279	262~327
0.15≦R<0.20	347	197~261
0.10≦R<0.15	475	131~196
0.05≦R<0.10	738	66~130
0.00≦R<0.05	667	2~65
Total	4,877	

Table 4. The top 11 and bottom 11 COGs by numbers of containing proteins and their distributions in 1,309 prokaryotes

Top 11			Bottom 11		
COG ID	# of proteins	# of prokaryotes	COG ID	# of proteins	# of prokaryotes
COG0583	22,048	1,068	COG5307	2	2
COG1028	20,857	1,248	COG5164	4	4
COG0642	19,080	1,184	COG5160	4	4
COG1309	19,004	1,157	COG5153	4	4
COG0745	17,928	1,195	COG5148	4	4
COG0438	16,517	1,243	COG5023	4	4
COG2202	16,257	967	COG4904	4	4
COG2814	16,105	1,229	COG5371	5	4
COG0596	13,192	1,186	COG5275	5	5
COG1595	12,355	1,036	COG5206	5	4
COG2207	12,226	951	COG5055	5	5

zekii Madrid E 균주와 gamma-Proteobacteria 강(class)의 *Legionella pneumophila* sub pneumophila Philadelphia 1 균주에서 각각 1개씩의 단백질이 존재하였다. 포유동물의 Arf6 단백질은 G-단백질의 일종으로 세포내이입, 액틴의 리모델링, 세포부착 등에 필요하며 효모에서 이러한 역할을 수행하는 ortholog가 Yel1 (yeast EFA6-like-1) 단백질이다[3]. Kang 등[5]은 2002년에 *Saccharomyces cerevisiae*와 원핵생물 42종에서 보존적인 COG를 보고하였다. COG 데이터베이스가 2014년에 업데이트 되면서 원핵생물의 COG와 진핵생물의 KOG로 분리되었다. 2014년의 COG에서 진핵생물이 분리될 때 1, 2개의 원핵생물에 존재하는 64개의 COG도 제거되었다[2]. *Gardnerella vaginalis* 15개 균주, *Buchnera aphidicola* 14개 균주, *Rickettsia prowazekii* 10개 균주 등 하나의 종(species)에 여러 균주가 존재하는 것을 2020년의 COG 데이터베이스는 정리하여 각 종당 하나씩만 남겨두었다[12]. COG는 3종류 이상의 생물종에 분포해야 하는데 COG5307은 2종의 세균에만 분포하므로 추후의 업데이트에서 변화가 있을 것으로 판단된다.

Table 4에서 COG5307, COG5164, COG5153 등은 모두 *S. cerevisiae* ATCC204508에서 확인된 COG들이다. COG5164는 전사 신장(elongation)인자인 SPT5인데, Spt4-Spt5의 이종이량체 복합체는 전사 개시 이후의 모든 단계에 참여하며 *S. cerevisiae*의 Spt4-Spt5는 RNA 중합효소I 및 II 모두에 대한 신장인자로 기능하고 전사체의 5′ 캡핑, 스플라이싱 및 3′-말단 처리와 연관되어 있다[13]. COG5160인 Ulp1은 Ubiquitin-like-specific protease 1로 역시 *S. cerevisiae*에서 발견되고 세포주기의 G2와 M 시기에서 중요한 역할을 한다[10].

COG는 기능유전체학(functional genomics)과 비교유전체학(comparative genomics)에 널리 사용되고 새로운 게놈의 서열 분석 후에 유전자들의 기능 파악에 널리 사용된다[12]. COG는 기초과학적 기능 외에, 의약품의 대량생산을 위한 균주개량을 위해 돌연변이 유전자의 선택[6] 등에 사용되므로 활용가치가 높다고 할 것이다.

The Conflict of Interest Statement

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

References

1. Fraser, C. M., Eisen, J. A. and Salzberg, S. L. 2000. Microbial genome sequencing. *Nature* **406**, 799-803.
2. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. and Koonin, E. V. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261-D269.
3. Gillingham, A. K. and Munro, S. 2007. Identification of a guanine nucleotide exchange factor for Arf3, the yeast orthologue of mammalian Arf6. *PLoS One* **2**, e842.
4. https://ftp.ncbi.nih.gov/pub/COG/COG2020/data/
5. 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.
6. Klein-Marcuschamer, D., Santos, C. N., Yu, H. and Stephanopoulos, G. 2009. Mutagenesis of the bacterial RNA polymerase alpha subunit for improvement of complex phenotypes. *Appl. Environ. Microbiol.* **75**, 2705-2711.
7. Lee, D. G. 2017. Conservative genes of less orthologous prokaryotes. *J. Life Sci.* **27**, 694-701.
8. Lee, D. G. and Lee, S. H. 2015. Investigation of conservative genes in 711 prokaryotes. *J. Life Sci.* **9**, 1007-1013.
9. 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.
10. Li, S. J. and Hochstrasser, M. 1999. A new protease required for cell-cycle progression in yeast. *Nature* **398**, 246-251.
11. Maddocks, S. E. and Oyston, P. C. 2008. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* **154**, 3609-3623.

12. Makarova, K. S., Wolf, Y. I. and Koonin, E. V. 2015. Archaeal clusters of orthologous genes (arCOGs): An update and application for analysis of shared features between *Thermococcales*, *Methanococcales*, and *Methanobacteriales*. *Life (Basel)* **5**, 818-840.
13. Meyer, P. A., Li, S., Zhang, M., Yamada, K., Takagi, Y., Hartzog, G. A. and Fu, J. 2015. Structures and functions of the multiple KOW domains of transcription elongation factor Spt5. *Mol. Cell. Biol.* **35**, 3354-3369.
14. 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.
15. Rosas-Pérez, T., Rosenblueth. M., Rincón-Rosales, R., Mora, J. and Martínez-Romero, E. 2014. Genome sequence of *Candidatus* Walczuchella monophlebidarum the flavobacterial endosymbiont of *Llaveia axin axin* (Hemiptera: Coccoidea: Monophlebidae). *Genome Biol. Evol.* **6**, 714-726.
16. Stackebrandt, E. 2002. Report of the *ad hoc* committee for the re-evaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* **52**, 1043-1047.
17. Tatusov, R. L., Koonin, E. V. and Lipman, D. L. 1997. A genomic perspective on protein families. *Science* **278**, 631-637.

초록 : 원핵생물 1,309종에 분포된 COGs (Clusters of Orthologous Groups of proteins) 연구

이동근 · 이상현*
(신라대학교 제약공학과)

저자들은 이전에 711개의 원핵생물에서 COG (Clusters of Orthologous Groups of proteins)를 분석한 결과를 보고하였다. COG 데이터베이스는 2020년에 1,309개의 원핵생물 게놈들을 사용하여 대폭 업데이트되었다. 이에 COG와 원핵생물 측면에서 업데이트된 4,877개의 COG를 구성하는 3,455,853개의 단백질들에 대한 분석 결과를 보고한다. 각 원핵생물이 보유한 COG 종류의 수는 97에서 2,281개의 사이였으며, 평균은 1,430.0개이고 표준편차는 414.2개였다. 문(phylum) 수준에서 보유 COG의 평균 수는 Mollicutes가 497.86개로 최소였고, Cyanobacteria가 1,642.90개로 최대였다. 가장 높은 보유 COG 개수를 가진 상위 10개 종은 모두 Proteobacteria였으며, 하위 10개 중 9개는 시험관 내에서 배양할 수 없는 *Candidatus* 구성원이었다. 각 COG에 속하는 단백질의 수는 2개에서 22,048개 사이였으며, 상위 11위 COG들은 12,000개 이상의 단백질을 포함하였다. 상위 11개 중 5개는 DNA에 결합하고 유전자 발현에 관여하는 COG로, 원핵생물에서 유전자 발현 조절의 중요성을 알 수 있었다. COG 데이터베이스는 게놈에 포함된 유전자를 식별하고 균주 개선을 위한 유전자를 선택하는 데 사용할 수 있어 많은 활용이 기대된다.