ISSN (Online) 2287-3406
Journal of Life Science **2022** Vol. 32. No. 3. 249~255 DOI : https://doi.org/10.5352/JLS.2022.32.3.249

Metabolic Pathways of 1309 Prokaryotic Species in Relation to COGs

Dong-Geun Lee[1], Ju-Hui Kim[2] and Sang-Hyeon Lee[1,2]*

[1]*Department of Pharmaceutical Engineering, Silla University, Busan 46958, Korea*
[2]*Department of Green-Chemistry Convergence Engineering, Graduate School, Silla University, Busan 46958, Korea*

Received January 20, 2022 / Revised February 4, 2022 / Accepted February 6, 2022

Metabolism is essential for survival and reproduction, and there is a metabolic pathways entry in the clusters of orthologous groups of proteins (COGs) database, updated in 2020. In this study, the metabolic pathways of 1309 prokaryotes were analyzed using COGs. There were 822 COGs associated with 63 metabolic pathways, and the mean for each taxon was between 200.50 (mollicutes) and 527.07 (cyanobacteria) COGs. The metabolic pathway composition ratio (MPCR) was defined as the number of COGs present in one genome in relation to the total number of COGs constituting each metabolic pathway, and the number of pathways with 100% MPCR ranged from 0 to 26 in each prokaryote. Among 1309 species, the 100% MPCR pathways included murein biosynthesis associated with cell wall synthesis (922 species); glycine cleavage (918); and ribosomal 30S subunit synthesis (903). The metabolic pathways with 0% MPCR were those involving photosystem I (1263 species); archaea/vacuolar-type ATP synthase (1028); and Na^+-translocation NADH dehydrogenase (976). Depending on the prokaryote, three to 49 metabolic pathways could not be performed at all. The sequence of most highly conserved metabolic pathways was ribosome 30S subunit synthesis (96.1% of 1309 species); murein biosynthesis (86.8%); arginine biosynthesis (80.4%); serine biosynthesis (80.3%); and aminoacyl-tRNA synthesis (82.2%). Protein and cell wall synthesis have been shown to be important metabolic pathways in prokaryotes, and the results of this study of COGs related to such pathways can be utilized in, for example, the development of antibiotics and artificial cells.

Key words : COG (cluster of orthologous proteins), conservative metabolic pathway, MPCR (metabolic pathway composition ratio), prokaryotes

서 론

원핵생물은 지구에 최초로 출현한 생물로 간주되며 서식지와 대사과정 등이 아주 다양하고 기초과학과 의료, 경제적인 측면에서도 중요하다. 원핵생물을 포함한 생명체는 영양성분이나 독성성분 등의 환경 변화에 따라 물질대사 등 생명현상을 조절하며 이를 통하여 현재 각 서식지에서 자기의 역할을 수행할 수 있었을 것이다.

동일한 속(genus)이나 과(family)의 원핵생물들은, 속이나 과에 공통된 대사경로를 갖고 종(species)이나 속에 특이한 대사경로도 가진다. 대개 하나의 물질대사는 여러 효소들의 연속된 작용으로 가능하며, 각 효소 합성에 필요한 유전자는 게놈에 존재한다. 공통된 대사경로를 보이면 유전자 염기 서열의 차이는 있더라도 공통된 유전자를 갖고 있을 것이다. 공통 조상의 유전자가 종분화(speciation)로 여러 종들에 분포할 때 동일한 기능을 수행하며 서열도 유사한데, 이런 유전자들의 집합을 ortholog라고 하며 동일 ortholog에서 유래한 단백질의 집합을 COG (Cluster of Orthologous Groups of proteins)라고 한다[10].

Lee와 Lee [10]는 2020년에 업그레이드된 1,309종의 원핵생물이 갖는 COG 자료를 이용하여 4,877개의 COGs를 구성하는 3,455,853개의 단백질들에 대한 분석으로 각 원핵생물과 문(phylum) 수준에서 보유한 COG의 수 등을 보고하였다. 한편 Lee [9]는 COG 방법이 개별적인 보존적 유전자만 파악 가능한 한계가 있다며, 대사경로 데이터베이스가 구축된 MetaCyc 대사경로 데이터베이스를 이용하여, 단독배양이 가능한 원핵생물 중에 게놈크기가 가장 작은 *Mycoplasma genitalium*보다 유전자의 수가 작은 14개의 원핵생물의 대사경로를 검토하였다. MetaCyc 대사경로 데이터베이스는 2016년에 업그레이드되었지만[1], 2020년에 업그레이드된 COG 데이터베이스도 pathways라는 항목을 따로 독립시켰다[7]. 이 논문에서는 COG pathways [7]를 구성하는 63개의 대사경로에 대한 분석결과를 보고한다.

재료 및 방법

재료

원핵생물의 각 유전체가 가지는 COG자료는 2022년 현재까

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

지 변화가 없는 2020년에 정리된 데이터베이스를 이용하였다 [10]. 각 원핵생물이 가지는 COG 데이터베이스를 확보한 후, COG pathways [7]에서 정한 각 대사경로를 구성하는 COG의 종류들을 파악하였고, 1,309개의 원핵생물 모두가 이들을 보유하는 지를 파악하였다. Table 1은 실제로 분석한 1,309종의 원핵생물을 문(phylum) 수준에서 각 문을 구성하는 종(species)의 개수를 나타내고 있다. Firmicutes와 Proteobacteria 문은 강(class)의 수준으로 정리하였다.

대사경로별 단백질 구성원들의 수

Perl과 엑셀 프로그램을 이용하여 전체 3,455,853개의 COG에 속하는 단백질들에서, COG pathways [7]에서 정한 각 대사경로를 구성하는 822개의 각 COG에 속하는 단백질들로 구분하여 1,309종의 원핵생물별로 나열하고 분석하였다.

대사경로 orthologs를 이용한 유전체 분석

대사경로 관련 보존 유전자(ortholog) 보유 계통수는 행렬로 작성된 분석대상 원핵생물들의 COG 자료들을 ClustalX (ver. 2.1) 프로그램으로 다중서열비교를 수행한 후, neighbor joining 방법으로 Presence-Absence phylogenetic tree를 작성할 때에 bootstrap method (n=100)를 적용하여 distance value를 담고 있는 phb 파일을 생성하였다. Dendroscpe 프로그램

Table 1. Phylogenetic groups, numbers of studied species, and average among 822 COGs in COG pathways

Phylogenetic Group Phylum Class	Abbre-viations	# of species	Average (range) among 822 COGs of COG pathways
Kingdom Archaebacteria			
Crenarchaeota	aC	25	364.92 (273~428)
Euryarchaeota	aE	79	400.27 (202~450)
Thaumarchaeota	aT	12	365.08 (342~386)
Other Archaea	aZ	6	240.67 (126~433)
Kingdom Eubacteria			
Acidobacteria	Ad	7	481.43 (465~503)
Actinobacteria	At	155	443.23 (283~525)
Aquificae	Aq	9	437.56 (413~456)
Bacteroidetes	Bc	107	439.62 (141~519)
Chlamydiae	Cm	6	329.17 (270~377)
Chlorobi	Cb	5	466.40 (455~481)
Chloroflexi	Cf	14	425.21 (313~502)
Cyanobacteria	Cy	41	527.07 (320~567)
Deferribacteres	De	5	486.00 (481~490)
Deinococcus-Thermus	DT	6	475.83 (447~490)
Firmicutes			
Bacilli	fB	73	432.26 (255~522)
Clostridia	fC	79	428.58 (239~505)
Negativicutes	fN	10	443.20 (349~511)
Tissierellia	fT	9	349.00 (300~429)
Other Firmicutes	fO	4	343.25 (274~468)
Fusobacteria	Fu	6	378.67 (265~485)
Mollicutes	Mo	14	200.50 (141~324)
Planctomycetes	Pm	14	470.00 (404~514)
Proteobacteria			
α-Proteobacteria	aP	158	468.77 (93~533)
β-Proteobacteria	bP	102	471.51 (71~543)
δ-Proteobacteria	dP	39	483.28 (337~535)
ε-Proteobacteria	eP	12	449.25 (372~498)
γ-Proteobacteria	gP	224	482.79 (124~579)
Other Proteobacteria	oP	6	439.00 (397~499)
Spirochaetes	Sp	11	370.55 (197~471)
Synergistetes	Sy	5	416.60 (386~445)
Thermotogae	Tt	9	377.56 (360~410)
Verrucomicrobia	Vr	9	448.11 (325~486)
Other eubacteria	oB	48	314.54 (86~495)

(ver 0.8)을 이용하여 distance value를 구하였고[8], distance value를 포함한 자료의 분석에는 엑셀 프로그램을 이용하였다.

결과 및 고찰

대사경로와 구성 COG 수

COG pathways [7]에서 지정하는 대사경로는 총 63개였고 관련 COG는 822개였다. Table 1에 문과 강 등의 분류단위별로 COG pathways와 관련된 822개의 COG 보유 평균과 범위를 표시하였다. 최소 71개의 COG를 *Candidatus* Nasuia deltocephalinicola NAS-ALF가 보유하였고 최대 579개의 COG를 *Hafnia alvei* FB1가 보유하였으며 평균은 442.06개였다. 분류단위별 COG 수 평균은 최저 200.50개(Mollicutes 문)이고 최대 527.07개(Cyanobacteria 문)이었다. Mollicutes 문과 other Archaea가 250개 이하였고 나머지는 모두 300개 이상이었다. 분류단위별로 최소와 최대 차이의 범위는 Deferribacteres 문이 최소인 9였고 β-Proteobacteria 강이 472로 최대였다. 범위를 구성원 수로 나누면 최소가 1.56(Actinobacteria), 최대가 51.17(other Archaea), 평균이 11.68이었다. 통계에서 중요한 중심극한정리를 적용할 수 있는 30개 이상의 구성원을 가진 분류단위 중에 other eubacteria를 제외하면 10개였다. 이들은 모두 400개 이상의 COG를 가졌고, 범위를 구성원의 수로 나누면 최소가 1.56(Actinobacteria), 최대가 6.02(Cyanobacteria), 평균이 3.58이었다. 향후 더 많은 원핵생물에 대한 COG 자료가 확보되면 더 정확한 비교가 가능할 것이지만, 현재로는 대사관련 COG 수의 분포가 Cyanobacteria 문에는 고르지 않고 Actinobacteria 문에는 고르다고 할 수 있었다.

각 대사경로의 종류와 구성하는 COG의 개수는 인터넷을 통하여 알 수 있어[7] 본고에서는 생략하였다. 각 대사경로를 구성하는 COG의 개수는 최소 1개(Entner-Doudoroff pathway)에서 최대 59개(tRNA modification)로 평균 13.05개였다. 각 대사경로를 구성하는 COG 개수가 많은 순서대로 나열하면 tRNA modification에 59개, CRISPR-Cas system에 46개, archaeal ribosomal proteins와 ribosomal 50S subunit에 각각 33개, aminoacyl-tRNA synthetases와 photosystem II에 각각 26개, cobalamin/B12 biosynthesis에 24개, aromatic amino acid biosynthesis에 23개, ribosome 30S subunits에 21개, translation factors에 21개, purine biosynthesis에 20개 등이다. 리보솜 구성 단백질과 번역인자들을 대사경로에 포함시킨 것이 특이하였다.

59개의 COG가 있는 tRNA modification 대사관련 COG는 *Candidatus* Hodgkinia cicadicola Dsem에서 2개만 존재하여 최소였고 *Myxococcus xanthus* DK 1622에 39개가 존재하여 최대였다. COG pathways [7]를 보면, 16S rRNA modification 대사관련에 15개, 23S rRNA modification 대사관련에 12개, tRNA modification에 59개 등 총 86개의 COG가 RNA modification에 관여하였다. RNA modification으로 RNA의 구조, 유전자 발현과 번역의 조절을 할 수 있고, DNA의 유전정보는 RNA modification에 의해 원래 서열과 다른 형태로도 발현될 수 있다. 미생물은 tRNA modification을 통하여 tRNA가 amino-acyl tRNA synthetase, 리보솜, mRNA 등과 반응하는 것에 영향을 주어 다양한 환경 및 영양 요인에 관여하면서 대사의 항상성이 유지되도록 한다[3].

46개의 COG가 있는 CRISPR-Cas system은 원핵생물의 박테리오파지나 플라스미드에 대한 면역시스템[6]이면서 세균의 유전자 전사 조절에도 관여[11]하는 것으로 알려져 있다. 이들은 고세균과 진정세균에 널리 분포하였지만 484개의 원핵생물에는 하나도 없었다. Chloroflexi 문의 *Herpetosiphon aurantiacus* DSM 785와 β-Proteobacteria 강의 *Vitreoscilla filiformis* ATCC 15551은 각각 22개의 COG를 가져 최대분포를 나타내었다.

33개의 COG로 구성된 archaeal ribosomal proteins은 고세균에서 최소 19개(Euryarcheota 문의 *Nanohaloarchaea archaeon* SG9)~최대 33개(Crenarchaeota 문의 6개 고세균)로 파악되었고 57.6~100%의 사이였다. 진정세균에서도 최대 3개의 archaeal ribosomal proteins가 관찰되었는데, 이들은 모두 Firmicutes 문에 속하는 3개의 종(species)이었다. 전체 진정세균 1,187종 중에서 archaeal ribosomal proteins의 보유 현황은 921종에서 0개였고, 189종이 1개, 74종이 2개였다. 진정세균의 50S ribosomal subunits도 33개의 COG로 구성되어 있는데 고세균은 최소 13개(*Candidatus* Nanopusillus acidilobi)에서 최대 19개(Salinarchaeum Harcht Bsk1)를 가져, 39.4~57.6%의 비율이었다. 진정세균은 최소 16개(*Candidatus* Nasuia deltocephalinicola NAS-ALF)에서 최대 33개(851종의 진정세균) 범위로, 54.5~100%의 비율이었다. 전체 진정세균의 71.7%에 해당하는 851종의 진정세균은 33개의 COG를 모두 가졌고, 1,162종의 진정세균이 30개 이상의 COG를 가지고 있어 연구대상 진정세균의 97.9%였다. 122종 고세균의 20.5%에 해당하는 25종의 고세균만이 archaeal ribosome proteins의 30개 이상의 COG를 가져 서로 차이를 보였다. 21개의 COG로 구성된 ribosomal 30S subunit는 고세균에서 13~16개(61.9~76.2%)가 관찰되었고, 진정세균에서 13~21개(61.9~100%)가 관찰되었다. Ribosome 30S subunits는 고세균과 진정세균에서 보존된 비율이 높고, 진정세균의 ribosome 50S subunits는 고세균에서도 상대적으로 많이 보존되어 있지만 고세균의 ribosome 50S subunits는 진정세균에는 많이 보존되어 있지 않은 것을 알 수 있었다. 따라서, ribosome 30S subunits 유전자는 고세균과 진정세균에 공통적으로 널리 분포하지만, 50S subunits 유전자는 서로 다른 진화경로를 밟았다고 할 수 있었다. 이는 고세균의 생활환경이 극한환경으로 진정세균의 일반적 환경과는 차이가 나고 분석대상인 1,309개의 원핵생물 중에서 고세균은 단지 122종뿐인 점이 그 원인으로 유추되었다.

완전한 대사경로

COG pathways에서 지정한 각 대사경로를 구성하는 COG들이 하나의 원핵생물에 모두 존재한다면, 그 대사경로는 완전한 것이라고 생각할 수 있다. 대사경로구성율(MPCR, metabolic pathway composition ratio)을 각 대사경로를 구성하는 전체 COG 개수에 대비하여 하나의 게놈에 존재하는 COG 개수의 비율로 정의하면, 구성 COG가 하나도 없는 0%부터 모든 COG가 존재하는 100%까지로 계량화가 가능하였다. 63개의 COG pathways 중에서 100% 대사경로구성율을 가진 원핵생물 수를 파악하고 가장 많은 대사경로부터 (완전한 대사경로 보유 원핵생물 수, 대사경로구성 COG 수, 대사경로)의 형식으로 600개 이상의 원핵생물에 존재하는 대사경로를 내림차순으로 나열하면 (922, 11, murein biosynthesis), (918, 5, glycine cleavage), (903, 21, ribosome 30S subunit), (851, 33, ribosome 50S subunit), (752, 2, proline degradation), (702, 1, Entner-Doudoroff pathway), (674, 4, serine biosynthesis), (623, 9, lipid A biosynthesis) 등의 순이었다. Mureine biosynthesis는 세포벽 성분인 peptidoglycan의 합성에 필요하고, lipid A는 세균의 세포벽을 구성하는 LPS의 내부에 위치한다. Glycine과 proline 아미노산 분해대사가 serine과 lipid A 생합성 경로보다 더 많은 원핵생물에 완전히 보존되어 있다.

1,309종의 원핵생물은 0~26개의 완전한 대사경로(100% MPCR)를 가졌다. 완전한 대사경로의 수가 0개인 원핵생물이 총 53개였고, γ-Proteobacteria 강에 속하는 *Serratia marcescens* sub marcescens Db11이 26개를 가졌다. 고세균과 진정세균의 Mollicutes 문이 평균 3개 이하의 100% MPCR 대사경로를 나타내었다. 각 분류단위가 갖는 완벽한 대사경로 수의 평균은 Thaumarchaeota 문이 0.67개로 최소였고, γ-Proteobacteria 강이 13.6개로 최대였다. 원핵생물이 보유한 완전한 대사경로가 0개라는 것이 대사를 전혀 할 수 없다는 것이라면 물질대사로 생존하는 생물에 성립될 수 없다. 이는 첫째로 대사경로를 구성하는 COG가 완전하지 않아도 대사가 가능하거나, 둘째로 COG pathways의 대사경로 수가 63개로 많은 대사경로를 포함하지 못한다는 것 등을 유추할 수 있다. Lee [9]는 단독배양이 가능한 원핵생물 중에서 보존적 유전자의 수가 가장 작은 *Mycoplasma genitalium*보다 보존적 유전자의 수가 작은 14종의 원핵생물이 1~19개의 완전한 대사경로를 가진다고 보고하였는데, Lee [9]의 보고는 2,526개의 대사경로에서 조사한 것이 본 연구의 63개의 대사경로에서 연구한 것과의 차이이고 대사경로가 완전하지 않아도 대사가 가능한 경우들을 토의하였다.

대사경로를 구성하는 COG의 완전 결핍

원핵생물에 각 대사경로를 구성하는 COG가 전혀 없다면, 그 대사경로는 없는 것으로 간주할 수 있을 것이고 그때 0% MPCR이 될 것이다. 63개의 각 대사에서 보면 0% MPCR을 나타내는 원핵생물은 대사경로에 따라 0~1,263종의 사이였다. Aminoacyl-tRNA synthetases, heme biosynthesis, RNA polymerase, ribosome 30S subunits와 ribosome 50S subunits, translation factors, tRNA modification 등 7개의 대사에 관련된 COG (보존적 유전자)는 모든 원핵생물에 1개 이상은 존재하였다. 63개의 대사경로 중에서 56개 대사경로가 0% MPCR인 원핵생물이 존재하였다. 1,309종의 원핵생물 중에서 600종 이상의 원핵생물에서 0% MPCR인 COG pathway를 (대사경로 완전결핍 원핵생물 수, 대사경로구성 COG 수, 대사경로)의 형식으로 나열하면, (1,263, 17, photosystem I), (1,229, 26, photosystem II), (1,028, 8, A/V (Archaeal/vacuolar)-type ATP synthase), (976, 7, Na^+-translocating NADH dehydrogenase), (921, 33, archaeal ribosomal proteins), (789, 2, TCA cycle-glyoxylate bypass), (680, 6, Na^+-translocating Fd:NADH oxidoreductase), (607, 1, Entner-Doudoroff pathway)의 순이었다. Photosystem은 광합성과 연계되어 있고 A/V-type ATP synthase와 archaeal ribosomal proteins는 고세균과 연관되어 있다. 각 대사의 MPCR이 0%인 원핵생물이 많은 것은 본 연구의 1,309종의 원핵생물 구성에 진정세균이 1,187개인 것 등 분석대상의 구성과 연관이 있는 것으로 추정된다. 전자전달계에서 전자가 이동하며 H^+를 이동시키지 않고 Na^+를 이동시키는 전자전달계를 구성하는 Na^+-translocating NADH dehydrogenase와 Na^+-translocating Fd:NADH oxidoreductase도 연구대상 원핵생물에 널리 분포된 것은 아니었다. Entner-Doudoroff pathway는 해당경로의 glyceraldehyde-3-phosphate를 피루브산으로 바꾸는데, 구성하는 COG는 1개라서 큰 의미를 부여하기 어려웠다.

원핵생물별 대사경로 구성 COG의 부재

1,309종의 원핵생물은 최소 3개에서 최대 49개의 대사경로가 0% MPCR이었다. 즉, 대사경로를 구성하는 COG를 전혀 가지고 있지 않았다. 3개의 대사경로가 0% MPCR인 원핵생물은 8종으로 γ-Proteobacteria 강에서 7종, Cyanobacteria 문에서 1종이었다. Other eubacteria에 속하는 bacterium AB1에 49개의 대사경로를 구성하는 COG가 하나도 없었다. 30개 이상의 대사경로가 전혀 없는 원핵생물은 31종이었다. 이중에 단독배양이 가능한 원핵생물은 30개의 대사경로가 전혀 없는 *Mycoplasma genitalium* G37과 34개가 전혀 없는 *Ureaplasma parvum* serovar 3 ATCC 700970 [5] 등 2개뿐이었다. 31~49개의 대사경로가 전혀 없는 원핵생물 중에서 초고온성이며 세포외공생 혹은 기생을 하는 *Nanoarchaeum equitans* Kin4-M [2]와 공생체의 메타게놈에서 유래한 bacterium AB1 [12]를 제외하면, 모두 명명이 *Candidate* 혹은 *Candidatus*로 시작되는 24종 등의 원핵생물로 모두 단독배양이 불가능하고[4] 공생이나 기생을 하였다.

원핵생물과 대사경로 구성 COG의 비율

COG pathways에서 지정하는 각 대사경로에 필요한 대사경로구성율(MPCR)을 각 원핵생물과 각 대사경로별로 평균을 구하였다. 원핵생물별로 대사경로구성율의 평균은 5.3% (*Candidatus* Nasuia deltocephalinicola NAS-ALF) 에서 80.2% (*Hafnia alvei* FB1) 범위로 평균 56.7% 이었다. 분류단위별로 대사경로구성율 평균은 20.4% (Mollicutes 문)~64.6% (γ-Proteobacteria 강)의 범위이고 평균은 51.4%였다.

Ribosome의 30S 소단위체를 구성하는 21개의 COG를 원핵생물별로 최소 61.9%에서 최대 100%로 보유하였고, 원핵생물 1,309종의 평균 대사경로구성율은 96.1%로 전체 대사경로 중에서 최대였다. 이렇게 대사경로별로 원핵생물에 따른 대사경로구성율은 최소가 3% (photosystem II)였고 평균이 56.7% 이었다. 각 대사경로구성율 70% 이상을 갖는 원핵생물의 비율은 리보솜의 30S 소단위체가(21개 COG) 96.1%, 리보솜의 50S 소단위체가(33개 COG) 80.3% 였다. 세포벽 성분인 peptidoglycan 합성에 필요한 mureine biosynthesis가(11개 COG) 86.8%, 아미노산을 합성하는 arginine biosynthesis가(12개) 80.4%, serine biosynthesis가(4개) 80.3%, isoleucine-leucine-valine biosynthesis가(11개) 79.5%, histidine biosynthesis가(11개) 78.5%, glutamine biosynthesis가(6개) 77.0%, lysine biosynthesis가(12개) 72.1% 이었다. tRNA에 아미노산을 결합시키는 aminoacyl-tRNA synthetases (26개)가 82.2%이므로, 원핵생물에서 단백질과 세포벽 합성관련 대사들의 보존성이 높고 중요한 것을 알 수 있었다. 핵산의 염기를 합성하는 pyrimidine biosynthesis가(10개) 85.7%와 purine biosynthesis가(20개) 73.5%였고, fatty acids biosynthesis도(15개) 74.9%였다. 분해관련 대사로는 glycine cleavage가(5개) 77.2%, proline degradation이(2개) 73.6%, pyrimidine degradation이(7개) 71.3% 등으로 3개의 대사경로가 70% 이상을 차지하였다.

분류그룹별 COG pathways의 분포

Fig. 1은 COG pathways를 구성하는 전체 822개 COG의 보유 유무를 1,309종의 원핵생물에서 구한 후, presence/absence 계통수를 neighbor-joining과 bootstrap (n=100)을 적용하여 작성하였고, 계통수에서 구한 각 원핵생물의 distance value를 Table 1의 분류단위별로 나타낸 것이다. Distance value의 평균의 관점에서 보면 고세균(aC, aE, aT, aZ)과 진정세균으로 나눌 수 있고, Mollicutes 문은 고세균과 나머지 진정세균들 사이에 있다고 할 수 있다. Distance value의 표준편차를 보면, β-Proteobacteria, Spirochaetes, other eubacteria가 0.08 이상이었고, 아직 분류가 불명확한 other eubacteria를 제외하면 11개의 구성원을 가진 Spirochaetes 문에서 가장 높았다. 한 분류단위의 distance value가 높은 평균이라는 것은 구성원들이 다른 분류단위와 유전자의 보유 측면에서 차이가 많다는 것이다. 각 COG를 획득 혹은 상실하는데 모든 원핵생물에게 동일한 시간이 소요되었다면 고세균이 진정세균보다 오래전에 지구상에 출현한 것으로도 해석할 수 있다. Distance value의 표준편차가 크다는 것은 각 분류단위의 구성원들이 보유하는 유전자의 종류가 평균에서 차이를 보이는 구성원들이 존재한다는 것으로 해석할 수 있다. 분류가 명확하지 않은 구성원들로 모인 고세균의 aZ와 진정세균의 oB가 상대적으로 표준편차가 컸다.

Fig. 2는 전체 4,788개의 COG의 보유 유무를 1,309종의 원핵생물에서 구한 후, neighbor-joining과 bootstrap (n=100) 기법으로 presence/absence 계통수를 작성하고, 계통수에서 각 원핵생물의 distance value를 Table 1의 분류단위별로 나타낸 것이다. Fig. 1과의 공통점은 고세균이 진정세균보다 distance value의 평균이 크고 진정세균과 구분되는 것이었다. Distance value의 표준편차는 Proteobacteria 문에 속하는 gP, bP, oP와 Spirochaetes 문에서 0.02 이상이었다. β-Proteobacteria 강과 Spirochaetes 문은 Fig. 1에서도 상대적으로 높은 표준편

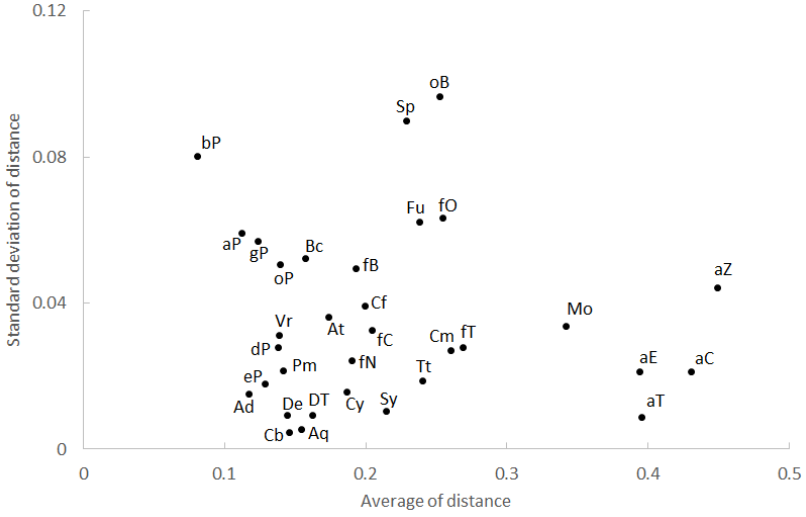

Fig. 1. Distribution patterns of prokaryotic genomes by distance value of phylogenetic tree based on presence/ absence of 822 COGs consisting 63 COG pathways.

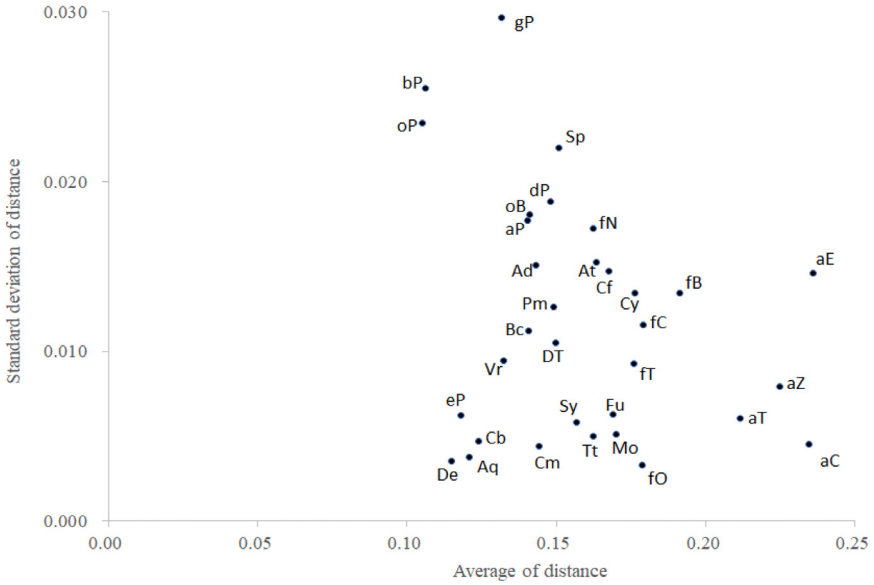

Fig. 2. Distribution patterns of prokaryotic genomes by distance value of phylogenetic tree based on presence/absence of all 4,788 COG.

차를 보였다. COG pathways를 구성하는 822개의 COG가 전체 4,788개의 COG의 보유 유무를 모두 반영하지는 못했지만 고세균과 진정세균의 분리 등이 나타나 어느 정도는 대표성을 띄는 것을 알 수 있었다. Fig. 1과 Fig. 2는 maximum-likelihood와 bootstrap (n=100)을 적용하여도 유사하였다(자료미제시). COG pathways는 아직 대사경로의 수가 MetaCyc [1] 보다 많이 부족하지만 특정 대사경로를 구성하는 COG와 원핵생물별로 부족한 COG의 파악, 목적 COG를 대상으로 하는 항생제 개발, 인공세포의 개발 등에 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. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S. and Karp, P. D. 2016. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nuc. Acids Res.* **44**, D471-D481.
2. Das, S., Paul, S., Bag, S. and Dutta, C. 2006. Analysis of *Nanoarchaeum equitans* genome and proteome composition: indications for hyperthermophilic and parasitic adaptation. *BMC Genomics* **7**, 186.
3. Edwards, A. M., Addo, M. A. and Dos Santos, P. C. 2020. Extracurricular functions of tRNA modifications in microorganisms. *Genes (Basel)* **11**, 907.
4. Firrao, G., Gibb, K. and Streten, C. 2005. Short taxonomic guide to the genus 'Candidatus Phytoplasma'. *J. Plant Pathol.* **87**, 249-263.
5. Glass, J. I., Lefkowitz, E. J., Glass, J. S., Heiner, C. R., Chen, E. Y. and Cassell, G. H. 2000. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* **407**, 757-762.
6. Horvath, P., Romero, D. A., Coûté-Monvoisin, A. C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. and Barrangou, R. 2008. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J. Bacteriol.* **190**, 1401-1412.
7. https://www.ncbi.nlm.nih.gov/research/cog/pathways/
8. Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M. and Rupp, R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**, 460.
9. Lee, D. G. 2018. Comparison of metabolic pathways of less orthologous prokaryotes than *Mycoplasma genitalium*. *J. Life Sci.* **28**, 369-375.
10. 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.* **31**, 834-839.
11. Li, M., Gong, L., Cheng, F., Yu, H., Zhao, D., Wang, R., Wang, T., Zhang, S., Zhou, J., Shmakov, S. A., Koonin, E. V. and Xiang, H. 2021. Toxin-antitoxin RNA pairs safeguard CRISPR-Cas systems. *Science* **372**, eabe5601.
12. Miller, I. J., Weyna, T. R., Fong, S. S., Lim-Fong, G. E. and Kwan, J. C. 2016. Single sample resolution of rare microbial dark matter in a marine invertebrate metagenome. *Sci. Rep.* **6**, 34362.

초록 : COG pathways에서 원핵생물 1,309종의 대사경로

이동근[1] · 김주희[2] · 이상현[1,2]*
([1]신라대학교 제약공학과, [2]신라대학교 일반대학원 그린화학융합공학과)

대사는 생존과 번식에 필수적이다. 2020년에 업그레이드된 COG (cluster of orthologous proteins) 데이터베이스에는 "pathways" 항목이 있다. 본 연구에서는 COG pathways를 이용하여 1,309개의 원핵생물의 대사 경로를 분석하였다. 63개의 대사경로와 관련된 822개의 COG가 있었고, 각 분류단위의 대사관련 COG의 평균은 200.50개(phylum Mollicutes)에서 527.07개(phylum Cyanobacteria)의 사이였다. MPCR을 대사경로구성율(하나의 게놈에 존재하는 COG 수 / 각 대사 경로를 구성하는 COG의 총 수)로 정의하였다. MPCR이 100%인 대사경로의 수는 원핵생물에 따라 0에서 26의 범위였다. 다수의 원핵생물에서 100% MPCR인 대사경로는 세포벽 합성과 관련된 murein biosynthesis (922종), glycine cleavage (918종), ribosome 30S subunits (903종) 등이었다. MPCR이 0%인 대사경로(종의 수)는 photosystem I (1,263종), A/V (archaea/vacuolar)-type ATP synthase (1,028종) 및 Na^+-translocation NADH dehydrogenase (976종) 등이었다. 원핵생물에 따라 3~49개의 대사경로를 전혀 수행할 수 없었다. MPCR의 보존성이 높은 대사경로의 순서는 ribosome 30S subunit (1,309종의 96.1%), murein biosynthesis (86.8%), arginine biosynthesis (80.4%), serine biosynthesis (80.3%) 및 aminoacyl-tRNA synthetases (82.2%) 등이었다. 단백질과 세포벽 합성이 원핵생물에서 중요한 대사경로인 것을 알 수 있었다. 본 연구의 결과와 원핵생물 사이의 대사경로와 관련된 COG는 항생제 및 인공세포의 개발 등에 활용될 수 있을 것이다.