ISSN (Print) 1225-9918
ISSN (Online) 2287-3406
Journal of Life Science **2016** Vol. 26. No. 6. 711~720 DOI : http://dx.doi.org/10.5352/JLS.2016.26.6.711

- Note -

Phylogenetic Analysis of 680 Prokaryotes by Gene Content

Dong-Geun Lee and Sang-Hyeon Lee*

Major in Pharmaceutical Engineering, Division of Bioindustry, College of Medical and Life Sciences, Silla University, Kwaebop-dong 1-1, Busan 617-736, Korea

Received February 12, 2016 /Revised April 27, 2016 /Accepted May 10, 2016

To determine the degree of common genes and the phylogenetic relationships among genome-sequenced 680 prokaryotes, the similarities among 4,631 clusters of orthologous groups of protein (COGs)' presence/absence and gene content trees were analyzed. The number of COGs was in the range of 103–2,199 (mean 1377.1) among 680 prokaryotes. *Candidatus* Nasuia deltocephalinicola str. NAS-ALF, an obligate symbiont with insects, showed the minimum COG, while *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen, represented the maximum COG. The similarities between two prokaryotes were 49.30–99.78 % (mean 72.65%). *Methanocaldococcus jannaschii* DSM 2661 (hyperthermophilic and autotrophic, Euryarchaeota phylum) and *Mesorhizobium loti* MAFF303099 (mesophilic and symbiotic, alpha-Proteobacteria class) had the minimum amount of similarities. As gene content may represent the potential for an organism to adapt to each habitat, this may represent the history of prokaryotic evolution or the range of prokaryotic habitats at present on earth. COG content trees represented the following. First, two members of Chloroflexi phylum (*Dehalogenimonas lykanthroporepellens* BL-DC-9 and *Dehalococcoides mccartyi* 195) showed a greater relationship with Archaea than other Eubacteria. Second, members of the same phylum or class in the 16S rRNA gene were separated in the COG content tree. Finally, delta- and epsilon-Proteobacteria were in different lineages with other Proteobacteria classes in neighbor-joining (NJ) and maximum likelihood (ML) trees. The results of this study would be valuable to identifying the origins of organisms, functional relationships, and useful genes.

Key words : COG (Clusters of Orthologous Groups of protein), gene content tree, maximum likelihood, neighbor-joining

서 론

생물의 분류는 전통적으로 표현형에 기초하여 이루어졌으나, 세균 등 미생물의 경우 형태적으로는 단순하고 구분이 매우 어려워 표현형을 이용한 분류에 한계가 있어 생리학 혹은 생화학적 특성을 이용한 동정과 분류가 이루어져왔다. 하지만 분자생물학적 방법의 발달에 따라 염기서열분석을 통한 동정의 장점들이 부각되어 현재는 세균의 경우 16S rRNA 유전자 염기서열을 이용한 동정과 분류법이 널리 사용되고 있다. 16S rRNA 유전자는 모든 세균에 존재하며 보존성이 높고 또한 방대한 데이터베이스가 구축된 장점이 있다[26].

하지만 rRNA유전자 만으로 분류하는 것에 한계가 제기되어 elongation factor Tu/1α, proton translocating ATPase의 subunit, recA, heat shock protein인 hsp 60 등 여러 housekeeping gene 들에 대한 분석이 제안되고 있다[24]. 한편 염기서열분석법의 발달로 미생물 유전체(genome) 전체의 염기서열에 대한 보고가 증가하는 상황에서 비교유전체학 등 생물정보학적 기법으로 유전체에서 유전자의 위치와 기능의 파악이 시도되고 있다[10]. Orthologs는 생물종들의 공통조상에 존재하던 유전자가 종분화(speciation)되어 서로 다른 종들에 분포하게 된 유전자들의 집합으로 정의하며, 같은 ortholog에 속하는 유전자들은 유사한 서열과 동일한 기능을 갖는다[10]. COG (Clusters of Orthologous Groups of protein)는 동일 ortholog에서 유래된 단백질의 집합으로, 각 COG는 최소 3가지 이상의 생물종에 분포해야 한다. COG 기법으로 게놈서열에서 실험 없이 ortholog 등을 탐색하며 동시에 생물학적 기능을 파악할 수 있다[10].

16S rRNA 유전자를 이용한 생물의 분류와 계통분석이 널리 사용되어 왔는데, 한계가 있어 유전자 순서 등의 다른 대상과 방법으로 계통분석을 시도하고 있으며, 최근 게놈에 산재하는 여러 유전자들을 이용한 계통분석이 보고되고 있다[4, 7, 18, 26, 28]. 한편 COG를 이용한 생물의 분류도 가능한데 동일한 기능을 갖는 단백질의 보유여부를 하나의 형질(character)로 파악하고, 하나의 게놈이 보유하는 COG들의 종류를 서로 비교하면 각 미생물의 유연관계를 파악할 수 있을 것이다[19, 23]. 따라서 COG를 이용한 생물의 분류는 미생물의 기능적 분류로 간주할 수 있으며 실제 실험을 하는 생화학적

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

분류법과 게놈의 염기서열에 기초한 분자생물학적 분류법의 중간자라 할 수 있을 것이다[5, 19]. Lee 등[19]은 각 COG 보유 여부에 따른 gene content tree를 시도하였지만 43종의 세균 중에서 Archaebacteria 9종, Proteobacteria 15종, Firmicutes 9종을 각각 분석하여 전체 생물종을 대상으로 파악하지 않았으며 생물종의 개수가 비교적 적었다.

2016년 1월 현재 고세균 83종과 진정세균 628종 등 총 711종의 세균이 보유한 COG가 보고되어 있다[9]. Lee와 Lee [21]는 711개의 세균에 포함된 4,631개의 COG 분석을 통해 보존적 유전자들을 밝히고 유전자들의 평균과 분산을 이용하여 변이가 큰 고세균과 변이가 적은 진정세균의 유전체들로 나누어진다고 하였지만 각 문(phylum)이나 강(class) 분류수준에서의 관계는 밝히지 못했다.

이 연구에서는 Lee 등[19]의 43개보다 훨씬 많은 680여개의 세균을 문(phylum)이나 강(class) 단위로 나누어 분석하지 않고 전체 생물종을 함께 분석하여 계통수를 작성하고 유전자보 유계통수(gene content tree)의 양상을 분석하고자 하였다.

재료 및 방법

재료

세균 유전체의 유전자 유사성에 관한 자료는 COGs에서 정리된 자료를 이용하였다[9]. 각 세균이 함유하고 있는 COG 자료를 확보하였다. 이들은 2016년 1월 현재 711종의 세균 유전체에 포함된 총 1,962,317개의 유전자들을 4,631개의 COG 그룹으로 분류해 놓았다[10]. 분석에서는 분류가 확실하지 않은 other bacteria에 포함된 진정세균의 31개 유전체는 제외하였다. Table 1은 실제로 분석한 자료인 680종의 세균들의 문이나 강 등의 분류학적 위치와 구성하는 생물종의 개수를 나타내고 있다. Proteobacteria와 Firmicutes 문은 강으로 나누어 분석하였다.

Table 1. Numbers of studied organisms and their 16S rRNA gene based phylogenetic groups, and range of possessing COG number at each phylum or class level

Phylogenetic group		# of COG			Number of organisms
(Superkingdom) Phylum	Class	Min.	Max.	Avg.	
(Archaea)					
Creanarchaeota		801	1,133	954.3	21
Euryarchaeota		898	1,538	1,204.1	56
Thaumarchaeota		346	931	638.5	4
other Archaea		853	1,057	925.0	2
(Bacteria)					
Acidobacteria		598	1,964	1,484.3	6
Actinobacteria		1,399	1,950	1,648.5	74
Aquificae		1,071	1,253	1,139.3	8
Bacteroidetes		223	1,804	1,325.1	55
Chlorobi		1,202	1,409	1,305.6	5
Chlamydiae		610	1,126	892.0	6
Chloroflexi		888	1,764	1,424.8	9
Cyanobacteria		1,205	1,859	1,631.5	31
Deinococcus-Thermus		1,314	1,566	1,433.7	6
Fusobacteria		841	1,520	1,244.8	5
Planktomycetes		1,391	1,841	1,605.3	6
Spirochaetes		579	1,614	1,211.3	7
Synergistetes		842	1,334	1,167.0	5
Thermotogae		1,156	1,222	1,185.1	7
Tenericutes	Mollicutes	316	1,575	1,019.6	10
Firmicutes	Bacilli	841	1,988	1,455.1	33
	Clostridia	900	1,803	1,382.4	49
	other Firmicutes	949	1,575	1,238.3	6
Proteobacteria	Alpha-Proteobacteria	125	2,242	1,589.0	75
	Beta-Proteobacteria	103	2,230	1,644.6	52
	Delta-Proteobacteria	966	2,053	1,616.1	28
	Epsilon-Proteobacteria	945	1,532	1,245.9	11
	Gamma-Proteobacteria	155	2,284	1,633.8	103

게놈 비교 및 유전자보유 계통수(gene content tree)

분석대상 680 종의 고세균과 진정세균 각 구성원이 보유하는 COG는 보존적 유전자 탐색에서 구하였으며[21], 각 COG의 보유유무에 따라 분석하였다. 즉 각 생물종이 4,631개의 각 COG를 보유하고 있는 지를 행렬로 작성하고 이를 Mega 프로그램(ver 5.1)의 phylogeny analysis를 이용하여 NJ (neighbor joining), ML (maximum likelihood), UPGMA, ME (minimum evolution), MP (maximum parsimony) tree를 작성하면서 bootstrap method (n=1,000)로 분석하였다[19, 20].

결과 및 고찰

보유 COG 수

Table 1에 분석대상 세균 680종을 문과 강의 분류단위로 구분한 후 각 분류단위를 구성하는 세균들이 보유한 COG 수의 최소, 최대, 평균을 나타내었다. COG database에서 다운받은 파일들 사이에 오류가 있어 검증 후 최소로 보유한 것을 기준으로 정리하였다. 전체적으로 각 생물이 보유한 COG 개수의 평균은 1,377.1개 였다. COG 보유개수를 보면 곤충과 절대공생성이며[3] beta-Proteobacteria 강에 속하는 *Candidatus* Nasuia deltocephalinicola str. NAS-ALF가 가장 적은 103개, 대사적으로 다양하며 기회성 병원균인[16] gamma-Proteobacteria 강의 *Pseudomonas aeruginosa* PAO1가 가장 많은 2,199개였다. 하나의 COG는 최소한 3종 이상의 세균의 게놈에 존재하는 것이므로 보유 COG 수가 많다는 것은 다른 세균들과 공통되는 생명현상이 많다고 할 수 있다[10]. 따라서 절대공생성인 *Candidatus* Nasuia deltocephalinicola str. NAS-ALF 균주는 곤충과 공생하면서 다른 세균들과의 공통되는 생명현상이 적어지는 방향으로 진화했다고 할 수 있을 것이다.

COG 보유개수를 500개 이하, 501~1,000개, 1,001~1,500개, 1,501~2,000개, 2,001개 이상으로 가진 세균의 수는 각각 21, 80, 259, 276, 44개였다. 분석대상 680개 세균의 78.67%는 보유한 COG의 개수가 1,001~2,000개 사이였다. 500개 이하인 세균 중 고세균은 *Nanoarchaeum equitans* Kin4-M 하나뿐이고 나머지는 모두 진정세균으로 Proteobacteria, Bacteroidetes, Tenericutes 문에 속하는 세균이 각각 9, 3, 8개였다. Tenericutes 문의 분석대상 세균은 모두 Mollicutes 강에 속하는데 이들은 세포벽이 없으며 다양한 동식물에 기생하며, 숙주 세포의 외부 또는 내부에 산다고 알려져 있다[24]. 2,000개 이상의 COG를 보유하는 세균 44개는 모두 Proteobacteria 문에 속하였고 alpha-, beta-, delta-, gamma-Proteobacteria 강에 각각 11, 8, 4, 21개의 세균들이 분포하였다. 보유 COG가 많다는 것은 다른 세균과 공통되는 유전자가 많다는 것이며, 다양한 환경에서 생존할 수 있는 잠재력이 높다고 혹은 생명현상의 범위가 넓다고 할 수 있을 것이다.

COG 보유 유사도

분석대상 680개의 세균에서 선택한 2개의 세균들 사이에 나타내는 전체 4,631개 COG 보유 유무의 유사도를 파악하니 최소 49.30%, 최대 99.78%, 평균 72.65%로 나타났다. Euryarchaeota 문의 *Methanocaldococcus jannaschii* DSM 2661와 alpha-Proteobacteria 강의 *Mesorhizobium loti* MAFF303099 사이가 최저였고, *M. jannaschii* DSM 2661와 gamma-Proteobacteria 강의 *Pseudomonas aeruginosa* PAO1 사이가 49.38%로 2번째의 최저유사도를 나타내었다. Tenericutes 문의 Strawberry lethal yellows phytoplasma (CPA) str. NZSb11와 *Candidatus* Phytoplasma australiense 사이가 최대였다. 그리고 2개의 세균 사이에서 99% 이상의 보유 COG 유사도를 보이는 것은 총 7개였다(자료미제시). 이들은 모두 Tenericutes 문 소속으로 보유 COG 수가 400개 이하였다. 전체 4,631개의 COG 중에서 보유하지 않은 COG들에 의해 유사도가 높게 나타났을 가능성이 있다.

하나의 세균이 나머지 분석대상 세균 679개와의 보유한 COG 종류의 유사도 평균을 구하였다. 전체 평균은 72.56%였고 범위는 65.43~76.62% 였다. Euryarchaeota 문의 *M. jannaschii* DSM 2661가 최소였고 Actinobacteria 문의 *Mycobacterium leprae* TN이 최대였다. *M. jannaschii* DSM 2661와 *My. leprae* TN은 각각 1142개와 917개의 COG를 보유하고 있었다. COG 수가 더 많은 *M. jannaschii* DSM 2661가 다른 세균들과의 COG 보유 유사도 평균이 낮다는 것은 독특한 생명활동을 하는 것을 나타낸다고 판단되었다.

유전자는 소실(gene loss)이나 획득(gene gain)의 결과이고 현재 보유하고 있는 유전자의 종류가 서식지와 연관이 있다[3, 34]. 그리고 본 연구의 분석대상 680종의 세균들이 각자가 검출된 서식환경을 대표한다는 전제하에서 아래와 같은 유추가 가능하다.

첫째, 보유 COG의 유사도가 나타내는 범위가 넓은 것을 서식환경으로 파악하여 서식환경이 완전히 다른 곳에 분포하는 3종 이상의 세균이 존재하면 서로간의 COG 유사도의 범위(최대와 최소의 차이)가 넓어질 것이다. 이러한 경우 중 하나는 진화의 초기에 발생하여 초기 지구의 환경이 현재도 유사하게 유지되는 곳에서 계속 서식하는 세균과 초기 지구와 많이 달라진 환경에 적응하면서 생존한 세균까지 포함하여 3종 이상이라면 서로간의 COG 유사도의 범위가 넓어질 것이다. COG 유사도의 범위가 큰 상위 10개 세균의 분류학적 위치는 Fig. 1과 Fig. 2에서 ⇧ 표시로 나타내었다. 이들은 43% 이상의 보유 COG 유사도 범위를 보였다. ⇧ 표시에 중복을 표시하지 않았지만 고세균(Fig. 1A, Fig. 2A 분계)이 상위 10개 중 7개를 차지하였으며, 유사도가 나타내는 범위가 가장 큰 세균은 심해의 열수구에 서식하는 초고온성이며 자가영양을 나타내는 Euryarchaeota 문의 *M. jannaschii* DSM 2661 [13]이었다. 이 세균은 초고온성이며 화학자가영양을 하는 Euryarchaeota 문

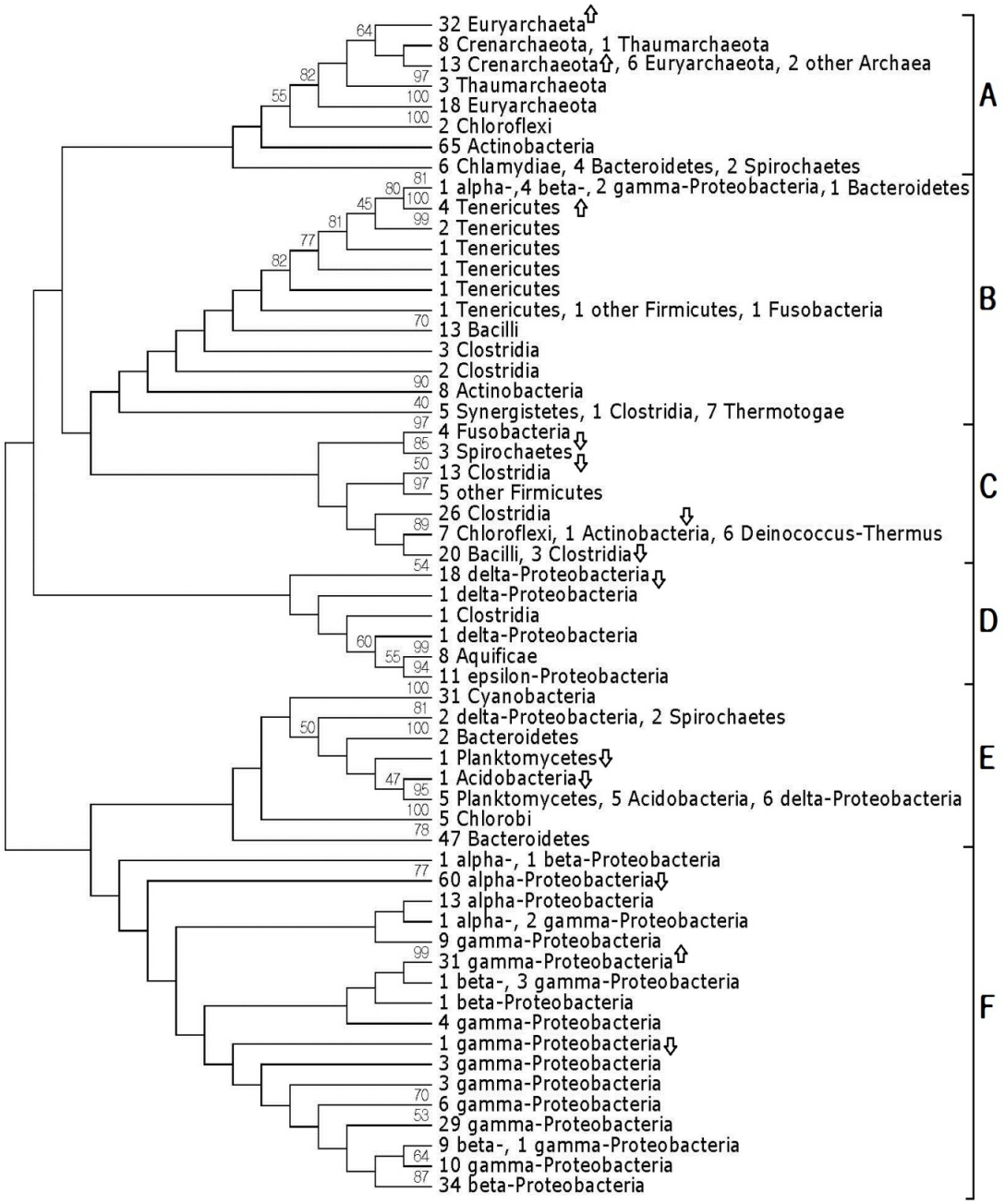

Fig. 1. ML (maximum likelihood) phylogenetic tree of 680 prokaryotes in the point of presence or absence of 4,631 COG. Bootstrap values at each node are expressed as a percentage of 1,000 trials and values lower than 40% were not expressed. Terminal branches have been extended for clarity and their length is therefore not meaningful.

의 *Methanotorris igneus* Kol 5와 최대인 95.57%, 중온성이며 뿌리에 공생하며 질소고정에 관여하는 alpha-Proteobacteria 강의 *Mesorhizobium loti* MAFF303099와 최소인 49.30%의 유사도를 보여 46.27%의 COG 유사도 범위를 보였다. Fig. 1의 B 분계와 Fig. 2의 C 분계 최상단에 위치하는 alpha-Proteobacteria인 *Candidatus* Hodgkinia cicadicola Dsem과 beta-Proteobacteria인 *Candidatus* Nasuia deltocephalinicola str. NASALF도 유사도의 범위가 컸다. 이들은 보유 COG수가 적은데, 비교대상 4,631개의 COG 중에서 보유하는 COG의 개수가 360개 이하여서 보유하지 않은 COG에 의해 유사성이 높아지는 한계가 있어 서로간의 COG 유사성이 높다고 판단할 수 없어 추후 연구가 필요할 것으로 사료되었다.

둘째, 반대로 보유 COG 유사도의 범위가 좁은 세균은 분석대상 세균들이 검출된 여러 서식지들의 평균적인 곳에 서식하는 세균일 것으로 판단할 수 있다. 한가지 예로 원시부터 현재까지 서식환경 변화의 평균적인 곳에 서식하는 세균은 원시지

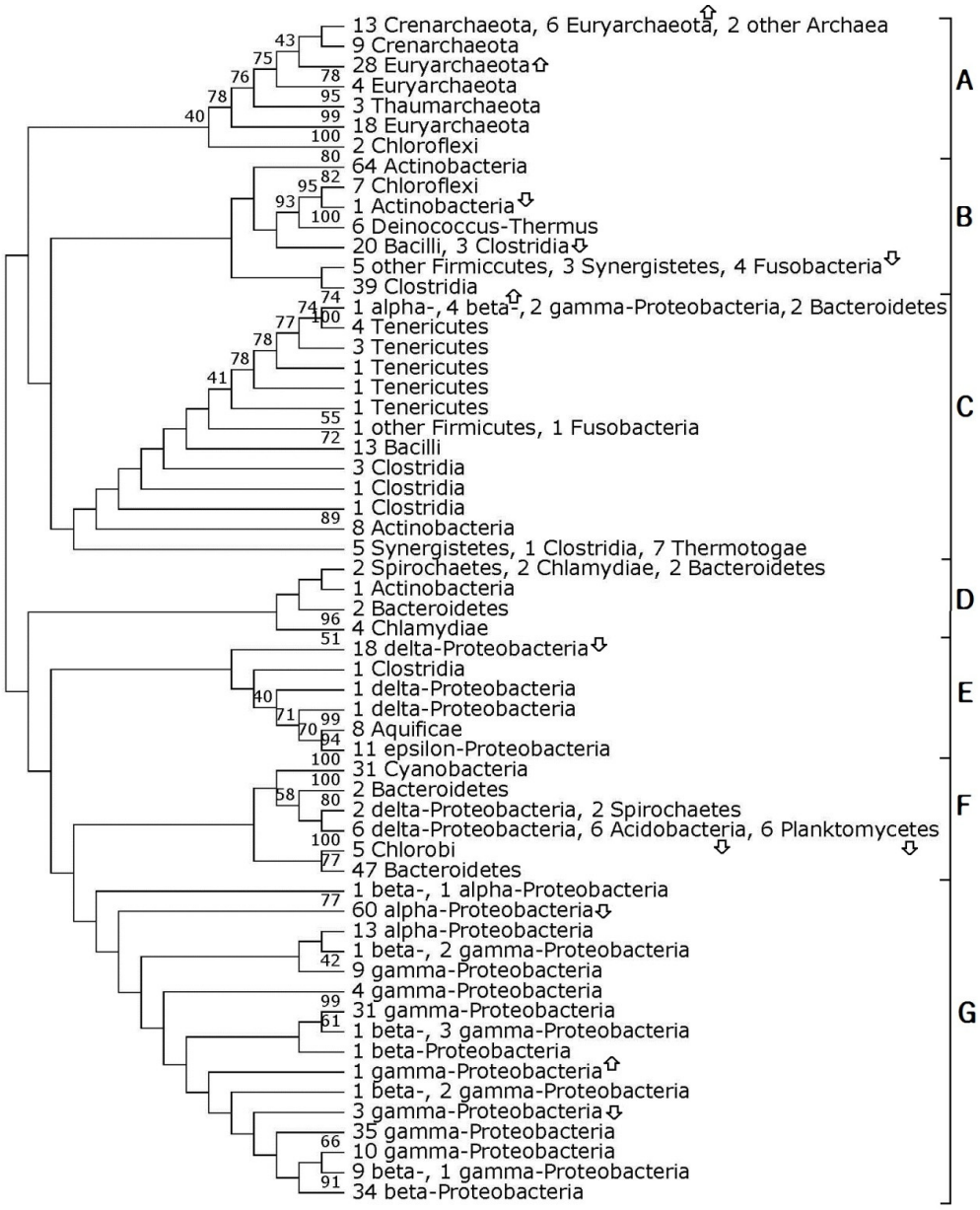

Fig. 2. NJ (neighbor joining) phylogenetic tree of 680 prokaryotes in the point of presence or absence of 4,631 COG. Bootstrap values at each node are expressed as a percentage of 1,000 trials and values lower than 40% were not expressed. Terminal branches have been extended for clarity and their length is therefore not meaningful.

구의 형태를 유지하는 서식지와 가장 많이 변화한 서식지에서 검출된 두 가지 세균들과의 COG 유사도를 구하면 그 차이가 크지 않을 것이다. COG 유사도의 범위가 작은 상위 10개 세균의 분류학적 위치는 Fig. 1과 Fig. 2에서 ⇩ 표시로 나타내었다. 이들은 20% 이하의 보유 COG 유사도 범위를 보였다. 범위가 넓은 세균은 고세균에 많이 분포하였는데 범위가 좁은 세균은 Fig. 1의 C 분계에 4개가 있었고 D, E, F 분계에 각각 2개씩 고루 분포하였다. 범위가 가장 작은 세균은 Cyanobacteria mat에서 분리된 50℃가 생육최적인 고온성 및 미호기성이며 광종속영양생물인 Acidobacteria 문의 *Chloracidobacterium thermophilum* B [30]로 Chlorobi 문의 *Chloroherpeton thalassium* ATCC 35110 와 최대인 81.82%, Euryarchaeota 문의 *M. jannaschii* DSM 2661와 최소인 64.95%의 유사도를 보여 16.87%의 범위를 보였다.

동일한 문 혹은 강에 속하는 구성원들 사이의 pairwise distance 들의 평균으로 구한 진화거리(evolutionary distance)를 최소부터 정렬하면 Tenericutes, Aquificae, Chlorobi, Chlamydiae, Crenarchaeota, Thermotagae, Thaumarchaeota, Cyanobacteria, epsilon-Proteobacteria의 순서였고 최대부터 정렬하면 gamma-Proteobacteria, delta-Proteobacteria, Spirochaetes,

beta-Proteobacteria, Chloroflexi, Clostridia, Bacilli, Actinobacteria, Bacteroidetes, other Firmicutes, Planktomycetes, Fusobacteria, Euryarchaeota의 순서로 Table 1에 표시된 구성원들의 개수와 연관성은 낮았다. 구성원들의 진화거리가 작다는 것은 서로간의 변이 즉 다른 COG를 함유할 가능성이 낮다는 것으로 구성원 서로 간의 생명현상이 유사할 것이라고 판단할 수 있을 것이다. 하지만 COG는 3종류 이상의 생물종에 존재하는 유전자에 기반하므로 2종류 이하의 생물종에 존재하는 유전자의 수를 많이 가진 종이 독특한 생명현상과 서식지를 나타낼 가능성이 높을 수도 있을 것이다.

Table 2는 동일한 속(genus)에 포함된 세균들이 나타내는 보유 COG의 유사도(%)이다. 구성원의 수는 *Bacillus* 속만 4개의 종(species)이 있었고 나머지는 모두 2개의 종이었다. 그리고 *Escherichia* 속은 종까지 동일한 *E. coli* 균주 2개를 비교하였다. 유사도의 범위는 88.1~99.1%의 범위였으며 *Bacillis* 속에서 88.1~96.8%로 동일 속에서도 종에 따른 차이를 보였다. Lee 등[19]은 종까지 동일한 균주들을 비교하여 *Helicobacter pylori*의 26695와 J99, *Neisseria meningitidis*의 MC58과 Z2491 균주 사이에서 각각 전체 COG 중 20% 이상과 25% 정도는 상이하다고 보고하였다. COG 최신판[10]에서는 *H. pylori* 26695 그리고 *N. meningitides* Z2491 균주가 삭제되어 Lee 등[19]의 결과와 비교는 불가능하였다. COG 최신판에 유일하게 *E. coli*만 O157과 K-12 두 균주가 있었다. 이들은 COG 보유유무에서 96.2%가 동일하였다(Table 2). *E. coli* 두 균주보다 높은 유사도를 보인 속은 *Rickettsia*, *Wolbachia*, *Sulfolobus*, *Pyrococcus*, *Mycoplasma* 등의 5개 속으로 이들은 동일한 종이 아님에도 불구하고 동일한 종에 속하는 *E. coli*의 두 균주 사이보다 유사도가 높았다.

강이 다르지만 동일한 속에 속하는 세균들보다 보유 COG

Table 2. Similarity (%) of COG content in same genus level

Phylum / Class	Genus	Similarity (%)
Crenarchaeota	*Sulfolobus*	96.3
Euryarchaeota	*Pyrococcus*	96.7
	Thermoplasma	89.0
Firmicutes/Bacilli	*Bacillus*	88.1~96.8
	Streptococcus	93.1
Alpha-Proteobacteria	*Bradyrhizobium*	92.7
	Rickettsia	97.8
	Wolbachia	98.8
Firmicutes/Clostridia	*Clostridium*	88.4
Gamma-Proteobacteria	*Escherichia*	96.2
	Spiribacter	96.0
Acidobacteria	*Granulicella*	92.1
Actinobacteria	*Mycobacterium*	89.4
Mollicutes	*Mycoplasma*	99.1
Cyanobacteria	*Nostoc*	93.5

유사도가 높은 세균들 중 최고의 유사도를 보인 것은 alpha-Proteobacteria인 *Candidatus* Hodgkinia cicadicola Dsem와 beta-Proteobacteria인 *Candidatus* Nasuia deltocephalinicola str. NAS-ALF로 98.19%였다. 문이 다르지만 유사도가 가장 높은 세균들은 Bacteroidetes인 *Candidatus* Uzinura diaspidicola str. ASNER가 beta-Proteobacteria인 *Candidatus* Tremblaya phenacola PAVE와 97.69%, gamma-Proteobacteria인 *Candidatus* Carsonella ruddii DC와 97.13%, gamma-Proteobacteria인 *Candidatus* Portiera aleyrodidarum BT-QVLC와 97.04%였다. 하지만 이들은 보유 COG 수가 500개 이하로 전체 4,361개의 COG 중에 보유하지 않은 COG의 수가 3,800개 이상이었다. 즉 보유하지 않은 COG에 의해 유사성이 높아질 수 있으므로 서로간의 COG 유사성이 높다고 판단할 수는 없었다.

유전자보유 계통수(gene content tree)

분석대상 680개 세균의 COG 보유 유무를 bootstrap (n=1,000)을 적용하여 작성한 maximum-likelihood (ML) 계통수는 Fig. 1에, neighbor-joining (NJ) 계통수는 Fig. 2에 나타내었다. Protebacteria 문과 Firmicutes 문은 강으로 표시하였고 other Archae와 other Firmicutes는 그대로 나타내었다(Table 1). 각 문이나 강 앞의 숫자는 해당 계통수에 포함된 구성원의 개수를 나타내는 것이다. 계통수 작성을 위한 자료는 형질기반자료와 거리기반자료가 있고 계통수를 제작하는 방법은 optimality와 clustering의 두 가지가 있는데 ML과 parsimony 계통수는 형질기반자료를 optimality 방법으로 제작하고 NJ와 UPGMA (unweighted pair-group method with arithmetic averages) 계통수는 거리기반자료를 clustering 방법으로 제작한다[5]. ML 계통수는 계산시간이 길지만 통계적 기법을 이용하여 가장 신뢰성이 높은 것으로 알려져 있다[17].

전반적으로 보면 첫째 고세균과 진정세균들이 확연히 분리되지 않았고 진정세균의 Chloroflexi 문에 속하는 *Dehalogenimonas lykanthroporepellens* BL-DC-9와 *Dehalococcoides mccartyi* 195 두 균주가 고세균과 가장 높은 COG 보유의 유사도를 보였다. 이러한 양상은 UPGMA 계통수를 제외한 ML (Fig. 1), NJ (Fig. 2), ME (minimum evolution), MP (maximum parsimony) 계통수에서도 거의 동일하였다(자료미제시). 고세균과 COG 보유 유연관계가 높은 Chloroflexi 문의 두 진정세균은 모두 혐기성이며 산소대신 유기할라이드(organohalide)를 이용하여 호흡하는 세균들로 각각 염소함유 물질로 오염된 지하수와 오염환경처리 소화조의 슬러지에서 분리되어 환경정화 등에 사용될 수 있다[23, 25]. ML 계통수에서는 Actinobacteria, Chlamydiae 문의 일부가 다른 진정세균보다 고세균과 COG 보유의 유연관계가 높은 것으로 나타났다(Fig. 1A). Lee와 Lee [21]는 본 연구의 분석대상들을 포함하는 711개의 유전체들을 대상으로 보존적 유전자들의 평균과 분산으로 각 유전체를

분석하여 변이가 큰 고세균과 상대적으로 변이가 작은 진정세균으로 확연히 나누어진다고 하였는데 본 연구의 COG 보유 계통수는 조금 달랐다. Lee와 Lee [21]는 일부 COG의 변이로, 본 연구는 모든 COG의 보유유무로 분석하였다.

ML, NJ, ME 계통수에서는 고세균 모두가 하나의 분계에 속하였고, MP와 UPGMA 계통수는 Table 1에서 other Archaeota로 분류되었지만 Nanoarchaeota 문에 속하는 *Nanoarchaeum equitans* Kin4-M이 다른 고세균보다 Tenericutes 문과 더 가까운 것으로 보였다. *N. equitans*는 초고온성 고세균인 *Ignicoccus hospitalis*의 공생체 혹은 기생체로 보유한 유전자의 수가 적다[15]. 분석에 사용된 Tenericutes 문의 Mollicutes 강은 다양한 동식물에 기생한다[25]. 이들은 기생하며 보유 유전자의 수가 적다는 공통점이 있다.

사람의 위장관 등에서 Euryarchaeota가 검출되고 있으며 메탄생성 고세균(methanogenic archaea)의 경우는 건강과의 상관성이 보고되고 있고, 위장관에 1,000종류 이상의 세균이 존재하는 것으로 보고되어 있는데[12], 위장관 내에서의 유전자의 수평적 전달의 가능성이 있으므로[33, 34] 이들에 대한 자료가 추가되어도 고세균이 본 연구처럼 하나의 분계를 구성할 지는 의문이다.

둘째 동일한 문이나 강의 구성원들 모두가 계통수에서 단계통군(monophyletic taxon)처럼 묶이는 것은 모든 계통수에서 문 수준에서 Aquificae (구성원 수: 8개), Chlorobi (5개), Cyanobacteria (31개), Deinococcus-Thermus (6개), Thermotogae (7개), 강 수준에서 epsilon-Proteobacteria (11개) 등의 6개였다(Fig. 3). 단계통군(monophyletic taxon)은 공통 조상 및 그 조상으로부터 진화한 모든 생물을 포함하는 분류군인데[2] 본 연구에서는 16S rRNA 유전자 기반 분류에서 동일한 문이나 강에 속하는 분류군들 만으로 계통수에서 하나의 분류군을 형성하였을 때 단계통군으로 간주하였다. ME 계통수에서 Synergistetes (5개)가, ML 계통수에서 Synergistetes (5개)와 Chlamydiae (6개)가, NJ와 UPGMA 계통수에서 Planktomycetes (6개)가 추가로 단계통군으로 나타났다. MP 계통수는 가장 많은 9개의 문이나 강이 단계통군으로 나타났다. Fig. 3에서 UPGMA와 NJ 계통수에서만 단계통군으로 나타나는 Planktomycetes 문은 ME와 ML 계통수에서 인접하였고 Synergistetes는 NJ와 UPGMA 계통수에서 인접하였다(Fig. 1, Fig. 2, 자료미제시). 이것은 단계통군으로 나타난 문이나 강에서 구성원들의 COG 보유 정도가 진화적으로 매우 가깝거나 다른 문이나 강과는 확연히 구분되는 생화학적 반응 혹은 단백질의 소유를 나타내는 것으로 판단되었다. 이들의 구성원 수를 보면 31개의 Cyanobacteria와 11개의 epsilon-Proteobacteria를 제외하면 모두 8개 이하였다(Table 1). 하지만 8개 이하의 구성원이 있는 Acidobacteria는 5개와 1개로(Fig. 1E), Fusobacteria는 4개와 1개로(Fig. 1B, Fig. 1C), Chloroflexi는 7개와 2개로(Fig. 1A, Fig. C), Spirochaetes는 3개, 2개, 2개로(Fig. 1A, Fig. 1C, Fig. 1E), Thaumarchaeota는 3개와 1개로(Fig. 1A) 나뉘어지는 등 5개의 문을 구성하는 세균들은 동일 문보다 다른 문과 COG 보유의 유연관계가 높게 나타났다(Fig. 1, Fig. 2). 따라서 구성원의 수가 적다고 해서 단계통군처럼 위치하는 것이 아니라는 사실을 알 수 있었다.

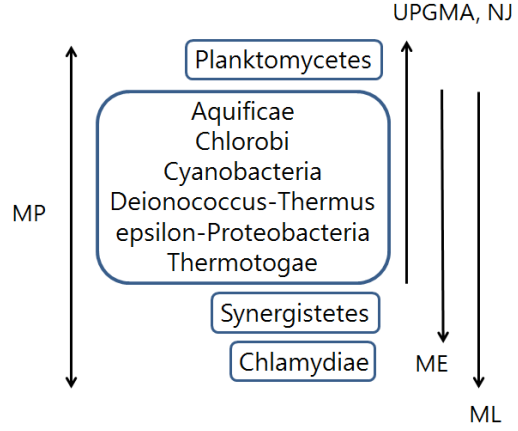

Fig. 3. Monophyletic phylum or class at NJ, ML, UPGMA, ME and MP trees.

셋째 16S rRNA 유전자 기반의 계통수와 본 연구의 COG 기반의 계통수에서 차이를 보였다. 위 첫째의 고세균과 진정세균의 사례 이외에도 Planktomycetes, Verrucomicrobia, Chlamydiae 문은 PVC group에 속하여[32] 16S rRNA 유전자를 이용한 계통수에서는 인접하지만 COG 보유 계통수에서는 Planktomycetes와 Chlamydiae 문이 Fig. 1의 A와 E 분계에, 그리고 Fig. 2의 E와 F 분계에 서로 떨어져 있다. 반면에 16S rRNA 유전자에서 Bacteroidetes/Chlorobi group에 속하는 구성원들은[14] COG 보유 계통수에서 함께 분포하는 것으로 나타났다(Fig. 1E, Fig. 2F). 한편 Proteobacteria 문의 delta-와 epsilon-Proteobacteria 강은 계통수에서 다른 Proteobacteria 강들과 계통수의 다른 위치에 분포하였다. 대부분의 alpha-, beta-, gamma-Proteobacteria 강 구성원들이 ML (Fig. 1F)과 NJ (Fig. 2G)에서 하나의 분계에 존재하는데 epsilon-Proteobacteria는 Aquificae 문과 유사한 것으로 나타났고(Fig. 1D, Fig. 2E) delta-Proteobacteri는 두 분계에 존재하였다(Fig. 1의 D와 E, Fig. 2의 E와 F). 이러한 양상은 ME, UPGMA, MP 계통수에서도 유사하였다(자료미제시). Lienau 등[22]은 166개의 게놈을 이용한 유전자보유계통수에서도 본 연구와 유사한 결과를 얻었다.

유전자보유계통수(gene content tree)는 다른 taxonomy와 불일치현상을 보이는데, 원인으로는 유전자의 수평전달(horizontal transfer) 같은 비수직적 현상(nonvertical gene events), 계통수 작성 때의 부적합한 기준, 다른 크기의 게놈을 비교할 때의 오류 등이 있다[6, 27, 29, 34]. 부정확한 계통수를 방지하는 첫걸음은 정확한 유전자 사이의 상동성(homology) 확립이 중요한데[22] 본 연구에서는 확립된 상동성 개념인

COG의 보유유무를 기반으로 NJ, ML 등으로 계통수를 작성하였다. Chaffron 등[6]은 유전자의 수평전달과 세포내 기생/공생하면서 게놈크기가 작은 세균 등의 두 경우를 제외하면 16S rRNA 유전자의 계통수와 유전자보유계통수가 대개 일치한다고 보고하였다. Dutilh 등[8]은 유전자의 수평전달이나 소실이 유전자보유계통수에 체계적인 오류를 생성하지는 않는다고 보고하였다. 한편 Langille 등[18]은 marker gene을 이용하면 16S rRNA 유전자 기반의 phylogeny와 function이 충분히 연계되어 있지만 환경유래 균주들은 유전자의 다양성이 커서 연계가 힘들 수 있다고 보고하였다.

본 연구결과는 Fig. 3처럼 16S rRNA 유전자 기반의 계통수와 일치하는 목이나 강도 있지만 불일치하는 경우도 있었다. Zheng 등[34]은 균주들 사이에 "translation, ribosomal structure and biogenesis" 같은 핵심유전자(core gene)들로 구성된 core genome은 양성선택(positive selection)을 하고 나머지 유전자들은 유전자 획득(gain)과 소실(loss)의 결과이며, 이것이 진화를 이끈다고 하였다. 본 연구의 COG는 공통조상 유래의 유전자 유래라는 개념을 가지고 있지만[10] 16S rRNA 유전자의 계통수와 달리 여러 목과 강의 균주들이 COG 보유 계통수에서 단계통군이 아니므로 이들 COG들의 획득, 소실, 수평전달이 비교적 광범위했던 것으로 판단할 수 있었다[7]. 이외에도 16S rRNA 유전자 계통수와 차이를 보이는 원인은 유전자들의 진화비율의 차이, 파악 못한 paralog, 유전자의 수렴(convergence) 등이 있다[17]. 16S rRNA 유전자같이 단 하나의 유전자로 계통수를 작성하는 것의 오류 가능성과 다른 대안들이 제시되었으며[26] 염기서열보다 단백질서열이 계통분석에 우수하다는 보고[1, 24]가 있지만 16S rRNA 유전자의 universal primer로 접근할 수 있는 편의성과 데이터베이스의 방대한 양으로 아직 표준으로 사용되고 있다.

넷째, ML 계통수(Fig. 1B)와 NJ 계통수(Fig. 2C)의 분계를 보면 alpha-, beta-, gamma-Proteobacteria와 Bacteroidetes 그리고 Tenericutes 문의 4개의 구성원들이 높은 유연관계를 나타냈다. 이들이 보유한 COG 개수는 모두 360개 이하로 보유한 COG 개수가 적어 계통수에서 유연관계가 높게 나타난 것으로 판단될 수 있었다. 즉 비교대상 4,631개의 COG 중에서 보유하는 COG의 개수가 360개 이하이니 4,000개 이상의 COG들을 보유하지 않는 공통점이 높은 유사성을 보인 결과로 판단될 수 있었다. 하지만 355개의 COG를 보유한 beta-Proteobacteria의 *Candidatus* Profftella armatura 가 316~327개의 COG를 보유한 4개의 Tenericutes 목보다 235개 이하의 COG를 보유한 alpha-, beta-, gamma-Proteobacteria와 Bacteroidetes 와 더 높은 유연관계를 보였으며, 346개의 COG를 보유하고 other Archaea에 속하는 *Nanoarchaeum equitans* Kin4-M이 NJ, ML 계통수 모두에서 900개 이상의 COG를 보유한 Euryarchaeota의 *Aciduliprofundum boonei* T469와 가장 가깝고 other Archaea로 분류된 Korarchaeota 문의 *Candidatus* Korarchaeum cryptofilum OPF8와 그 다음으로 가까웠다. 그리고 다른 고세균들과 함께 계통수의 같은 영역에 존재하여(Fig. 1A, Fig. 2A) COG 개수의 정량성 요인 외에 정성적 요인도 계통수에서 이들의 위치에 영향을 미쳤다는 것을 알 수 있었다.

다섯째, 계통수에서 bootstrap의 비율이 모든 가지에서 높지 않았다. Zheng 등[34]은 *Lactobacillus* 속의 세균 16종의 계통수를 비교하여 1,240여개의 유전자로 구성된 core genome 계통수에서는 높은 bootstrap 비율을 보였지만 게놈의 전체 유전자보유 계통수에서는 낮은 bootstrap 비율을 보였다. 본 연구에서는 680개의 세균이 서로 다른 문(phylum)을 형성하며 분석대상 세균이 보유하지 않을 수도 있는 4,631개의 COG로 계통수를 형성하여 bootstrap 비율이 40% 이하인 것이 많은 것은(Fig. 1, Fig. 2) Zheng 등[34]의 결과와 어느 정도 일치한다고 할 수 있었다.

유전자보유 계통수의 적용과 유용성

유전자보유 자료를 이용하여 생물의 기원을 파악하고, 기능적 연관성과 진화경로, 대사경로 등을 파악할 수 있다[1].

유용미생물을 검출하는 방법은 유전자의 염기서열과 기능에 기반할 수 있다. 유전자의 염기서열은 유용유전자 같은 marker gene이나 16S rRNA 유전자를 이용할 수 있다[18]. 하지만 Langille 등[18]도 사람의 장과 같이 국한된 환경에서는 16S rRNA 유전자가 유용하지만 환경에서 검출된 세균은 16S rRNA 유전자가 아주 가까운 종이라도 기능의 다양성이 높다고 하였다. 실제로 속이나 종이 같아도 보유 유전자의 종류는 차이가 많을 수 있었다(Table 2). 따라서 환경에서는 marker gene의 서열법이나 기능에 기반하는 것이 좋다고 할 것이다. 항생제 생성 균주의 탐색에 phylogeny를 이용한 보고와 genome mining으로 polyphenol들을 검출한 보고가 있었다[11, 31]. 실험의 난이도나 비용측면에서 전체 게놈정보 파악을 통한 COG 접근법이 분리한 세균에서 원하는 유전자 탐색에 유용할 경우도 있을 것이다.

References

1. Baldauf, S. L., Roger, A. J., Wenk-Siefert, I. and Doolittle, W. F. 2000. A kingdom level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972-977.
2. Baum, D. 2008. Reading a phylogenetic tree: The meaning of monophyletic groups. *Nat. Edu.* **1**, 190.
3. Bennett, G. M. and Moran, N. A. 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome Biol. Evol.* **5**, 1675-1688.
4. Boeckmann, B., Marcet-Houben, M., Rees, J. A., Forslund, K., Huerta-Cepas, J., Muffato, M., Yilmaz, P., Xenarios, I., Bork, P., Lewis, S. E. and Gabaldón, T. 2015. Quest for orthologs entails quest for tree of life: In search of the gene stream. *Genome Biol. Evol.* **7**, 1988-1999.

5. Bos, D. H. and Posada, D. 2005. Using models of nucleotide evolution to build phylogenetic trees. *Dev. Comp. Immunol*. **29**, 211-227.
6. Chaffron, S., Rehrauer, H., Pernthaler, J. and von Mering, C. 2010. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947-959.
7. Chung, Y. and Ané, C. 2011. Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage sorting and horizontal gene transfer. *Syst. Biol.* **60**, 261-275.
8. Dutilh, B. E., Huynen, M. A., Bruno, W. J. and Snel, B. 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol*. **58**, 527-539.
9. ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data
10. 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.
11. Guo, J., Ran, H., Zeng, J., Liu, D. and Xin, Z. 2016. Tafuketide, a phylogeny-guided discovery of a new polyketide from *Talaromyces funiculosus* Salicorn 58. *Appl. Microbiol. Biotechnol.* in press.
12. Horz, H. P. and Conrads, G. 2010. The discussion goes on: What is the role of euryarchaeota in humans? *Archaea* **2010**, 967271
13. http://microbes.ucsc.edu/cgi-bin/hgGateway?db=methJann1
14. http://www.ncbi.nlm.nih.gov/taxonomy/?term=Bacteroidetes/ Chlorobi%20group
15. Jahn, U., Huber, H., Eisenreich, W., Hugler, M. and Fuchs, G. 2007. Insights into the autotrophic CO_2 fixation pathway of the archaeon *Ignicoccus hospitalis*: comprehensive analysis of the central carbon metabolism. *J. Bacteriol.* **189**, 4108-4119.
16. Klockgether, J., Munder, A., Neugebauer, J., Davenport, C. F., Stanke, F., Larbig, K. D., Heeb, S., Schöck, U., Pohl, T. M., Wiehlmann, L. and Tümmler, B. 2010. Genome diversity of *Pseudomonas aeruginosa* PAO1 laboratory strains. *J. Bacteriol*. **192**, 1113-1121.
17. Lang, J. M., Darling, A. E. and Eisen, J. A. 2013. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS One* **8**, e62510.
18. Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G. and Huttenhower, C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol*. **31**, 814-821.
19. Lee, D. G., Kang, H. Y., Kim, S. H., Lee, S. H., Kim, C. M., Kim, S. J. and Lee, J. H. 2003. Classification of archaebacteria and bacteria using a gene content tree approach. *KSBB J.* **18**, 39-44
20. 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.
21. Lee, D. G. and Lee, S. H. 2015. Investigation of conservative genes in 711 prokaryotes. *J. Life Sci*. **25**, 1007-1013.
22. Lienau, E. K., DeSalle, R., Rosenfeld, J. A. and Planet, P. J. 2006. Reciprocal illumination in the gene content tree of life. *Syst. Biol*. **55**, 441-453.
23. Löffler, F. E., Yan, J., Ritalahti, K. M., Adrian, L., Edwards, E. A., Konstantinidis, K. T., Müller, J. A., Fullerton, H., Zinder, S. H. and Spormann, A. M. 2013. *Dehalococcoides mccartyi* gen. nov., sp. nov., obligately organohalide-respiring anaerobic bacteria relevant to halogen cycling and bioremediation, belong to a novel bacterial class, *Dehalococcoidia classis* nov., order Dehalococcoidales ord. nov. and family Dehalococcoidaceae fam. nov., within the phylum Chloroflexi. *Int. J. Syst. Evol. Microbiol*. **63**, 625-635.
24. Ludwig, W. and Klenk, H. P. 2000. Overview: A phylogenetic backbone and taxonomic framework for procaryotic systematics. pp. 49-65. *In* Boone, D. R., Castenholz, R. W. and Garrity, G. M. (eds.) Bergey's Manual of Systematic Bacteriology Volume 1. 2nd edition. Springer-Verlag , NY.
25. Mukherjee, K., Bowman, K. S., Rainey, F. A., Siddaramappa, S., Challacombe, J. F. and Moe, W. M. 2014. *Dehalogenimonas lykanthroporepellens* BL-DC-9T simultaneously transcribes many rdhA genes during organohalide respiration with 1,2-DCA, 1,2-DCP, and 1,2,3-TCP as electron acceptors. *FEMS Microbiol. Lett*. **354**, 111-118.
26. Rajendhran, J. and Gunasekaran, P. 2011. Microbial phylogeny and diversity: Small subunit ribosomal RNA sequence analysis and beyond. *Microbiol. Res*. **166**, 99-110.
27. Salichos, L. and Rokas, A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327-331.
28. Shi T. 2016. Impact of gene family evolutionary histories on phylogenetic species tree inference by gene tree parsimony. *Mol. Phylogenet. Evol*. **96**, 9-16.
29. Szöllősi, G. J., Tannier, E., Daubin, V. and Boussau, B. 2015. The inference of gene trees with species trees. *Syst. Biol.* **64**, e42-e62.
30. Tank, M. and Bryant, D. A. 2015. *Chloracidobacterium thermophilum* gen. nov., sp. nov.: an anoxygenic microaerophilic chlorophotoheterotrophic acidobacterium. *Int. J. Syst. Evol. Microbiol*. **65**, 1426-1430.
31. Tian, J., Chen, H., Guo, Z., Liu, N., Li, J., Huang, Y., Xiang, W. and Chen, Y. 2016. Discovery of pentangular polyphenols hexaricins A-C from marine *Streptosporangium* sp. CGMCC 4.7309 by genome mining. *Appl. Microbiol. Biotechnol*. in press.
32. Wagner, M. and Horn, M. 2006. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* **17**, 241-249.
33. Walter, J. and Ley, R. 2011. The human gut microbiome: Ecology and recent evolutionary changes. *Annu. Rev. Microbiol*. **65**, 411-429.
34. Zheng, J., Zhao, X., Lin, X. B. and Gänzle, M. 2015. Comparative genomics *Lactobacillus reuteri* from sourdough reveals adaptation of an intestinal symbiont to food fermentations. *Sci. Rep*. **5**, 18234.

초록 : 유전자 보유 계통수를 이용한 원핵생물 680종의 분석

이동근 · 이상현*
(신라대학교 의생명과학대학 바이오산업학부 제약공학전공)

게놈분석이 완료된 680개의 세균의 공통 유전자 보유 정도와 유연관계를 파악하기 위해 4,631개의 COG (Clusters of Orthologous Groups of protein) 보유 유사도와 COG 보유 계통수를 작성하여 다음과 같은 결과를 얻었다. 균주별 COG 보유개수는 103~2,199개 사이였고 평균 1377.1개 였다. 곤충과 절대공생성인 *Candidatus* Nasuia deltocephalinicola str. NAS-ALF가 최저였고 기회성병원균인 *Pseudomonas aeruginosa* PAO1가 최대였다. 2개의 세균들 사이에 나타내는 COG 보유 유무의 유사도는 49.30~99.78% 사이였고 평균 72.65%였다. 초고온성이며 자가영양생활을 하는 *Methanocaldococcus jannaschii* DSM 2661과 중온성이며 공생생활을 하는 *Mesorhizobium loti* MAFF303099 사이가 최소였다. 유전자 보유 정도가 생물이 각 서식지에 적응하는 정도를 나타내므로 이 결과는 원핵생물 진화의 역사 혹은 현재 지구의 원핵생물 서식지 범위를 나타내는 것일 수도 있다. COG 보유계통수를 통하여 첫째 진정세균인 Chloroflexi문의 일부는 진정세균보다 고세균과 유연관계가 높았고, 둘째 16S rRNA유전자에서 동일한 문(phylum)이나 강(class)으로 분류되지만 COG 보유 계통수에서는 일치하지 않는 경우가 많았으며, 셋째 delta-와 epsilon-Proteobacteria는 다른 Proteobacteria와 다른 분계(lineage)를 이루었다. 본 연구결과는 생물의 기원 파악과 기능적 연관성 파악 그리고 유용유전자 탐색 등에 이용할 수 있을 것이다.