

## 유전자보유 계통수를 이용한 *Archaea*와 *Proteobacteria* 분류

이동근<sup>1</sup>, 이진옥, 이재화

<sup>1</sup>신라대학교 마린바이오산업화지원센터, 신라대학교 생명공학과 발효공학연구소  
전화 (051) 999-5748, FAX (051) 999-5636

### Abstract

A Gene content phylogenetic tree and a 16S rRNA based phylogenetic tree were compared for 9 *Archaea* and 15 *Proteobacteria*, whole-genome sequenced, by neighbor joining and bootstrap methods (n=1000). Ratio of conserved COG (clusters of orthologous groups of proteins) to ortholog revealed that they were within the range of 4.60% (*Mezorhizobium loti*) or 56.57% (*Mycoplasma genitalium*). The diversity of ratio meant the possibility of searching for useful genes, as they possess peculiar genes. The gene content tree and the 16S rDNA tree showed coincidence and discordance in *Archaea* and *Proteobacteria*.

### 서론

미생물에서 유용한 효소는 이미 밝혀진 종 (species)과 유연관계가 높은 종에서 발견될 가능성이 높을 것이다. 미생물 분류는 분자생물학의 발달에 따라 이를 이용한 분류법들이 많이 시도되어 왔으며 현재까지 16S rRNA 염기서열을 이용한 분류법이 널리 사용되어 왔다. 각 COG는 하나의 공통조상유전자 (ancient conserved domain)에 해당하는 것으로 간주할 수 있으며 그룹의 구성원들은 유사한 기능을 보인다<sup>1)</sup>. 따라서 동일 혹은 유사한 기능을 수행하는 단백질의 유무를 하나의 character로 파악하고 전체 계통에서 이러한 작업을 수행하여 각 미생물의 유연관계를 파악할 수 있을 것이다. 본 연구에서는 *Archaea*와 *Proteobacteria*에 대해 적용하여 현재 널리 통용되는 16S rRNA에 기초한 분류법과 COG의 보유정도에 따른 분류법을 비교·분석하였다.

### 재료 및 방법

미생물 유전자의 유사성에 관한 자료는 COGs에서 정리된 자료를 이용하였고<sup>2)</sup> 분석대상은 고세균 (*Archaea*) 9종, 진정세균 (*Bacteria*)중 15종의 단백질세균 (*Proteobacteria*)이었다. 보존적 COG를 구하고, 각 생물종의 3,307개 각 COG를 보유

유무를 행렬로 작성하였다. 게놈 염기서열에서 16S rRNA를 추출하였고, 불가능한 경우에는 RDP-II<sup>3)</sup>에서 1,400 bp 이상의 염기서열만을 추출하고 National Center for Biotechnology Information (NCBI)I number로 표시하였다 (Table 1). COG 보유 행렬과 다중염기배열 (multiple alignment)을 수행한 16S rRNA를 ClustalX (ver. 1.64b)로 neighbor joining method와 통계적 의미를 갖는 bootstrap method (n=1000)를 이용하여 분석하였다<sup>4)</sup>.

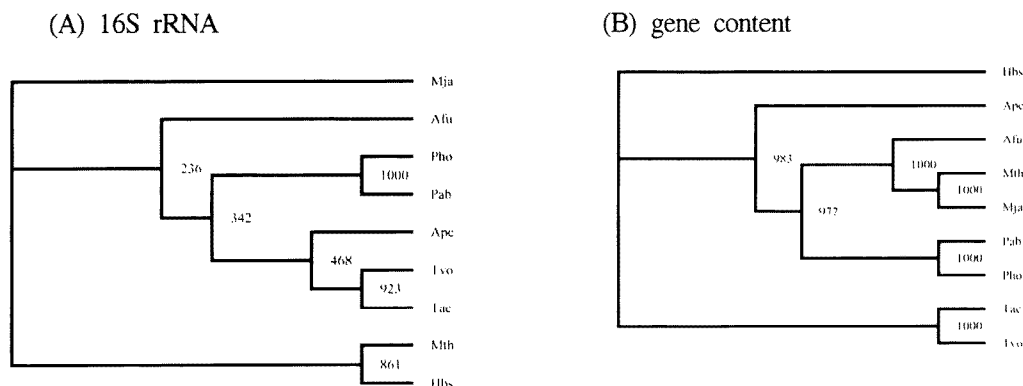
### 결과 및 토의

**게놈 비교 (보존된 COG 비율)** Table 1은 분석대상 각 종의 보존적 유전자 탐색 결과를 나타내고 있다. Archaea는 17.36 ~ 28.29%, Proteobacteria는 4.60 ~ 42.54% 범위의 보존적 COG 비율을 보였다.

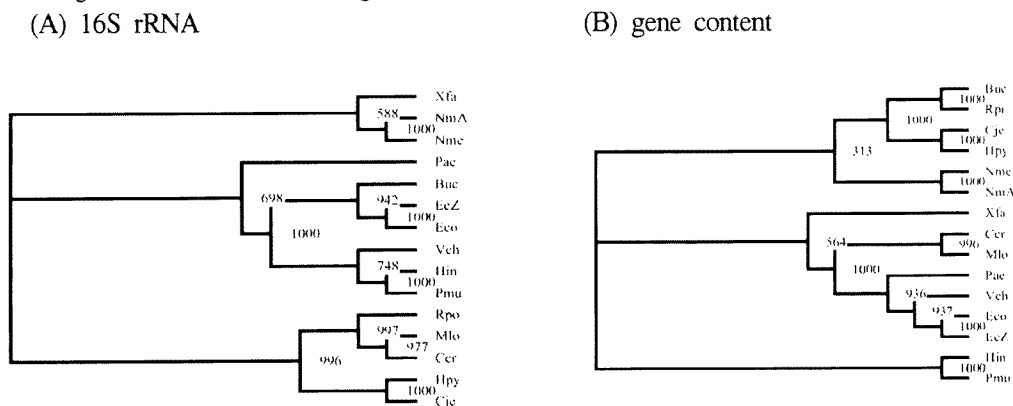
**Table 1.** Studied 24 genomes derived from COGs database and 16S rRNA from NCBI. Percentage (%) of conserved COG over ortholog are also presented

Phylogenetic Group	Organism	Abbreviation	number of ortholog	16S rDNA accession # at NCBI	number of conserved COG	Percentage of conserved COG (%)
Archaea	<i>Aeropyrum pernix</i>	Ape	1,202		340	28.29
	<i>Archaeoglobus fulgidus</i>	Afu	1,958			17.36
	<i>Halobacterium sp. NRC-1</i>	Hbs	1,818			18.70
	<i>Methanobacterium thermoautotrophicum</i>	Mth	1,464			23.22
	<i>Methanococcus jannaschii</i>	Mja	1,407			24.16
	<i>Pyrococcus abyssi</i>	Pab	1,516	L19921		22.43
	<i>Pyrococcus horikoshii</i>	Pho	1,442			23.58
	<i>Thermoplasma acidophilum</i>	Tac	1,258			27.03
	<i>Thermoplasma volcanium</i>	Tvo	1,268			26.81
Proteobacteria	<i>Buchnera sp. APS</i>	Buc	583		248	42.54
	<i>Campylobacter jejuni</i>	Cje	1,344	AF393204		18.45
	<i>Caulobacter crescentus</i>	Ccr	2,880			8.61
	<i>Escherichia coli K12</i>	Eco	3,618			6.85
	<i>Escherichia coli O157</i>	EcZ	3,900			6.36
	<i>Haemophilus influenzae</i>	Hin	1,595			15.55
	<i>Helicobacter pylori 26695</i>	Hpy	1,135			21.85
	<i>Mesorhizobium loti</i>	Mlo	5,390	X67230		4.60
	<i>Neisseria meningitidis MC58</i>	Nme	1,555			15.95
	<i>Neisseria meningitidis Z2491</i>	NmA	1,540	AL162757		16.10
	<i>Pasteurella multocida</i>	Pmu	1,838	AF294412		13.49
	<i>Pseudomonas aeruginosa</i>	Pae	4,698			5.28
	<i>Rickettsia prowazekii</i>	Rpr	723			34.30
	<i>Vibrio cholerae</i>	Vch	2,998			8.27
	<i>Xylella fastidiosa</i>	Xfa	1,687			14.70

토양에서 발견되는 *Mezorhizobium loti*의 경우는 보존된 COG 비율이 4.60%로 가장 낮았다. 이는 다른 미생물들과 공통되는 기능이 낮다는 것으로, 특징적 유전자를 많이 함유하고 있을 가능성이 높을 것으로 사료되었다. 따라서 새로운 유용유전자 등을 발견할 확률이 상대적으로 높다고 할 수 있을 것이다.



**Fig. 1.** Comparison of the phylogenetic trees of *Archaea* obtained from neighbor-joining analysis of either 16S rRNA gene sequence(A) and gene content(B). Bootstrap values at each node are expressed as a number over 1000 trials. Terminal Branches have been extended for clarity and their length is therefore not meaningful.



**Fig. 2.** Comparison of the phylogenetic trees of *Proteobacteria* obtained from neighbor-joining analysis of either 16S rRNA gene sequence(A) and gene content(B). Bootstrap values at each node are expressed as a number over 1000 trials. Terminal Branches have been extended for clarity and their length is therefore not meaningful.

**유전자보유 계통수 (gene content tree)** Fig. 1은 *Archaea*를 16S rRNA와 gene content를 이용하여 분류한 결과이다. 두 계통수가 일치하는 부분 (Pho와 Pab, Tvo와 Tac)과 일치하지 않는 부분 (Mth와 Hab)이 있었고 이는 horizontal gene transfer에 의한 영향 등

이 원인으로 사료되었다<sup>5)</sup>. 호염성인 Hbs가 유전자보유 계통수에서 독립된 것은 호염성인 다른 고세균과 유전자 조성이 다르다는 것을 나타내는 결과로 판단되어졌다.

Fig. 2는 *Proteobacteria*를 16S rRNA와 유전자보유 유무를 이용하여 분류한 결과이다. alpha-와 epsilon-*Proteobacteria*의 유연관계를 보면 16S rRNA에서는 높고 유전자보유 측면에서는 상대적으로 낮게 나타났다. Ccr과 Mlo는 16S rRNA와 달리 유전자보유 계통수에서는 gamma 그룹과 함께 위치하였고, parsimony method 등을 이용한 결과에서도 유사함을 확인하였다.

본 연구에서 수행한 계놈의 특정 COG 보유 유무에 따른 계통수는 전체 계놈을 고려한다는 점과 16S rDNA처럼 보존적이지 않은 유전자까지 고려한다는 점이 특이한 점이라 할 수 있을 것이다. COG를 이용한 분류법은 염기서열을 이용한 기능적 측면에서 미생물을 분류하는 것으로 직접 실험을 수행하는 생화학적 분류법과 전체계놈 염기서열에 기초한 분류법의 중간자적 위치에 있다고 할 수 있을 것이며, 유용단백질 탐색 작업 등에 그 유용성이 높다고 할 수 있을 것이다.

## 요 약

염기서열 분석이 완료된 9종의 고세균 (Archaea)과 15종의 단백세균 (*Proteobacteria*)에 대하여 유전자보유 유무와 16S rRNA에 의한 계통수를 neighbor joining method와 통계적 의미를 갖는 bootstrap method (n=1000)를 이용하여 분석하였다. 보존적 COG와 각 미생물 보유 ortholog수에 대한 비율은 4.60% (*Mezorhizobium loti*)와 56.57% (*Mycoplasma genitalium*) 사이로 종에 따라서 공통 유전자의 보유정도가 차이를 보이는 것으로 독특한 유전자를 탐색할 수 있는 가능성을 제시하는 결과로 사료되었다. Archaeobacteria와 *Proteobacteria* 그리고 *Firmicutes*모두 유전자보유 계통수와 16S rRNA 계통수가 일치하는 부분과 일치하지 않는 부분으로 나뉘어진다는 것을 알 수 있었다.

## References

1. Henikoff, S., E. A. Greene, B. S. Pietrokovski, T. K. Attwood, and L. Hood, Gene Families: The Taxonomy of Protein Paralogs and Chimeras (1997), *Science* **278**, 609-614
2. <http://www.ncbi.nlm.nih.gov/COG/>
3. <http://rdp.cme.msu.edu>
4. Amann, R., W. Ludwig, and K. H. Schleifer, Identification of uncultured bacteria: a challenging task for molecular taxonomists (1994). *ASM News* **60**, 360-365
5. Jain, R., M. Rivera, and J. A. Lake, Horizontal gene transfer among genomes: The complexity hypothesis (1999), *PNAS*, **96**, 3801 -3806