한국생물공학회지 제18권 제1호
Korean J. Biotechnol. Bioeng.
Vol. 18, No. 1, 39-44(2003)

Gene Content Tree를 이용한 Archaebacteria와 Bacteria 분류

[1]이 동 근 · [2]강 호 영 · 김 수 호 · 이 상 현 · [2]김 철 민 · [3]김 상 진 · † 이 재 화
신라대학교 공과대학 생명공학과, [1]신라대학교 마린바이오산업화지원센터,
[2]부산대학교 의과대학부설 부산지놈센터, [3]한국해양연구원 미생물연구실
(접수 : 2002. 12. 2., 게재승인 : 2003. 1. 30.)

Classification of Archaebacteria and Bacteria using a Gene Content Tree Approach

Dong-Geun Lee[1], Ho-Young Kang, Cheol-Min Kim[2,3], Sang-Jin Kim[3], and Jae-Hwa Lee†
Department of Bioscience and Biotechnology, College of Enginneering, Silla University, Pusan 617-736, Korea
[1]Marine Biotechnology Center for Bio-Functional Material Industries, Silla University, Pusan 617-736, Korea
[2]Busan Genome Center, College of Medicine, Pusan National University, Pusan 617-736, Korea
[3]Microbiology Laboratory, Korea Ocean Research and Dvelopment Institute, PO Box 29 Ansan 425-600, Korea
(Received : 2002. 12. 2., Accepted : 2003. 1. 30.)

A Gene content phylogenetic tree and a 16S rRNA based phylogenetic tree were compared for 33 whole-genome sequenced procaryotes, neighbor joining and bootstrap methods (n=1,000). 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*). This meant that the ratio was diverse among analyzed procaryotes and indicated the possibility of searching for useful genes. Over 20% of orthologs were independent among the same species. The gene content tree and the 16S rDNA tree showed coincidence and discordance in Archaeabacteria, *Proteobacteria* and *Firmicutes*. This might have resulted from non-conservative genes in the gene content phylogenetic tree and horizontal gene transfer. The COG based gene content tree could be regarded as a midway phylogeny based on biochemical tests and nucleotide sequences.

Key Words : Archaebacteria, *proteobacteria*, *Firmicutes*, 16S rRNA, gene content tree

서 론

미생물에서 유용한 유전자나 효소는 이미 밝혀진 종(species)과 유연관계가 높은 종에서 발견될 가능성이 높을 것이다. 하지만 동·식물과 달리 미생물은 형태적으로 매우 단순하여 구분이 힘들기 때문에 분류는 주로 생리적, 생화학적 특성에 의존하여 왔다. 그러나 분자생물학적 방법의 발달에 따라 이를 이용한 분류법들이 많이 시도되어 왔으며 현재까지 16S rRNA 염기서열을 이용한 분류법이 널리 사용되어 왔다. 16S rRNA는 constant region과 variable region으로 구성되어 있는데 variable region은 종 (species)과 속 (genus)간의 분화에 따른 다양성이 큰 부분을 함유하고 있어 특정 분류군에서만 나타나는 염기서열을 함유하고 있다고 알려져왔다. 또한 16S rRNA의 특정 부분은 진화 속도가 매우 느리기 때문에 많은 생물체가 공통적으로 갖는 보존된 염기서열과 이차구조를 나타내어 다양한 분류군의 상호 비교를 가능하게 하여 가장 널리 이용되어 왔다 (1).

하지만 아직도 미생물분류에 있어서는 아직 해결할 수 없는 문제로 적합한 분류기준의 설정과 공통조상에서 어떤 경로로 분화되었고 각 분류그룹들이 어떻게 연관되는 지를 명확히 해야하는 것이다 (2). 16S rRNA 유전자와 다르면서 적합한 분류기준이 되는 유전자의 설정을 위하여 rpo 유전자, hsp 유전자, ITS 염기서열 등이 일부 이용되기도 하였다. 한편 DNA-DNA hybridization, GC contents, RFLP (restriction fragment length polymorphisms), RAPD (random-amplified polymorphic DNAs), 탐침유전자 (gene probe) 등을 이용하여 각 미생물의 유연관계를 밝히려는 시도가 있었다. 이들은 모두 염기서열에 기초한 방법이라고 할 수 있을 것이다.

Orthologs는 공통의 조상으로부터 종분화되어 서로 다른 종에 있는 유전자들의 집합으로 정의하며, COG (clusters of orthologous groups of protein)는 ortholog들에서 유래된 단백질의 집합을 이르는 말로 대개 유사한 구조와 기능을 갖는

† Corresponding Author : Department of Bioscience and Biotechnology, Silla University, Kwaebop-dong 1-1, Pusan 617-736, Korea
Tel : 051 - 999 - 5748 Fax : 051 - 999 - 5636
E-mail : jhalee@silla.ac.kr

것으로 알려져 있다 (3-5). 각 COG는 적어도 3가지 이상의 계보 (lineage)에서 유래된 paralog 그룹 혹은 개별의 단백질들로 구성되어 있어 하나의 공통조상유전자(ancient conserved domain)에 해당하는 것으로 간주할 수 있다 (6). 유전체 염기서열 분석과 gene annotation을 통하여 2002년 4월 현재 3800여 개의 COG 그룹으로 분류해 놓은 것을 인터넷을 통하여 접근할 수 있다(8). 따라서 동일하거나 비슷한 기능을 수행하는 단백질의 유무를 하나의 character로 파악하고 전체 게놈에서 이러한 작업을 수행하여 각 미생물의 유연관계를 파악할 수 있을 것이다. 이러한 측면에서 COG를 이용한 분류법은 일종의 기능적 측면에서 미생물을 분류하는 것으로 생화학적 분류법과 전체게놈의 염기서열에 기초한 분류법의 중간자적 위치에 있는 분류법이라 할 수 있을 것이며 아울러 유용한 단백질을 탐색하는 작업 등에 유용하게 응용될 수도 있을 것이다.

본 연구에서는 Archaeabacteria와 Bacteria의 *Firmicutes*와 *Proteobacteria*에 대해 적용하여 현재 널리 통용되는 16S rRNA에 기초한 분류법과 COG의 보유정도에 따른 분류법을 비교·분석하였다.

재료 및 방법

재료

총 43종의 원핵생물 (procaryote) 유전체 (microbial genome)를 NCBI의 공개 서버로부터 추출하였다(7). 미생물 유전자의 유사성에 관한 자료는 COGs에서 정리된 자료를 이용하였는데 (8) 이들은 2002년 4월 현재 43종의 미생물 유전체를 ortholog 그룹으로 분류하여, 총 77,069개의 유전자들을 3,852개의 단백질 그룹으로 분류해 놓았다. 43종의 원핵생물은 Archaea가 9종, Bacteria중 *Firmicutes*가 9종, *Proteobacteria*가 16종, 기타 8종이었다 (9). 이 중에서 Archaea 9종, Bacteria 중 *Firmicutes* 9종, *Proteobacteria* 15종을 연구대상으로 하여 분석하였다. epsilon-Prpteobacteria에 속하는 *Helicobacter pylori J99*는 긴 염기서열의 16S rRNA 염기서열을 확보할 수 없어 비교에서 제외하였다. Table 1은 본 연구에서 분석한 33종의 원핵생물이 보유하고 있는 COG 자료를 나타내고 있다.

Table 1. Studied 34 genomes derived from COGs database and 16S rRNA from NCBI
Sequences of 16S rDNA was extracted from whole genome sequences. In the case of absence of 16S rDNA sequence in whole genome sequences, RDP-II was used and accession number at NCBI was recorded.

Phylogenetic Group		organism	Abbreviation	number of ortholog	16S rDNA accession # at NCBI
Archaea	Crenarchaeota	*Aeropyrum pernix*	Ape	1,202	
	Euryarchaeota	*Archaeoglobus fulgidus*	Afu	1,958	
		Halobacterium sp. NRC-1	Hbs	1,818	
		Methanobacterium thermoautotrophicum	Mth	1,464	
		Methanococcus jannaschii	Mja	1,407	
		Pyrococcus abyssi	Pab	1,516	L19921
		Pyrococcus horikoshii	Pho	1,442	
		Thermoplasma acidophilum	Tac	1,258	
		Thermoplasma volcanium	Tvo	1,268	
Bacteria	Firmicutes	*Bacillus halodurans*	Bha	3,032	AB027713
		Bacillus subtilis	Bsu	3,030	
		Lactococcus lactis	Lla	1,694	AF515225
		Mycobacterium leprae	Mle	1,213	
		Mycobacterium tuberculosis	Mtu	2,760	
		Mycoplasma genitalium	Mge	396	X77334
		Mycoplasma pneumoniae	Mpn	441	
		Streptococcus pyogenes	Spy	1,287	
		Ureaplasma urealyticum	Uur	414	
	Proteobacteria	*Buchnera sp. APS*	Buc	583	
		Campylobacter jejuni	Cje	1,344	AF393204
		Caulobacter crescentus	Ccr	2,880	
		Escherichia coli K12	Eco	3,618	
		Escherichia coli O157	EcZ	3,900	
		Haemophilus influenzae	Hin	1,595	
		Helicobacter pylori 26695	Hpy	1,135	
		Mesorhizobium loti	Mlo	5,390	X67230
		Neisseria meningitidis MC58	Nme	1,555	
		Neisseria meningitidis Z2491	NmA	1,540	AL162757
		Pasteurella multocida	Pmu	1,838	AF294412
		Pseudomonas aeruginosa	Pae	4,698	
		Rickettsia prowazekii	Rpr	723	
		Vibrio cholerae	Vch	2,998	
		Xylella fastidiosa	Xfa	1,687	

게놈 비교 및 유전자보유 계통수(gene content tree)

Bacteria 그룹의 *Firmicutes*와 *Proteobacteria* 그리고 Archae bacteria에 공통적인 COG를 보존적 유전자 탐색에서 구한 다음, 각 COG의 보유유무에 따라 분석하였다. 즉 각 생물종이 3,852개의 각 COG를 보유하고 있는 지를 행렬로 작성하고 이를 ClustalX(ver. 1.64b)를 이용하여 neighbor joining method와 bootstrap method (n=1,000)로 상관관계를 분석하였다(10).

16S rDNA 염기서열 추출 및 분석

전체 유전자 염기서열에서 rRNA 그룹을 조사하여 16S rRNA를 추출하였고, 이것이 불가능한 경우에는 RDP-II(http://rdp.cme. msu.edu)의 hierachy browser를 이용하여 16S rRNA의 염기서열이 1400bp 이상의 것만을 추출하였다. RDP-II에서 유래된 16SrRNA 유전자의 acession number는 NCBI number로 Table 1에 표시되어 있다. 그 후 ClustalX (ver. 1.64b)를 이용하여 다중염기배열 (multiple alignment)을 수행한 후 게놈 비교와 동일하게 neighbor joining method와 bootstrap method (n=1,000)로 분석하여 게놈 비교에서 유래된 결과와 대조하였다.

결과 및 토의

게놈 비교 (보존된 COG 비율)

Table 2는 유전자보유 계총수를 작성하기 위한 보존적 유전자 탐색의 결과를 나타내고 있다. 각 분류그룹에서 보존된 COG의 비율은 Archaea에서는 17.36~28.29% 정도로 나타났고 *Firmicutes*에서는 7.39~56.57% 였으며 *Proteobacteria*에서는 4.60~42.54% 범위였다. 보존된 COG의 비율이 높은 종은 보유하고 있는 ortholog의 수가 많지 않은 종들로, 이들은 *Mycoplasma* 종들., *Ureaplasma urealyticum*, *Buchnera* sp. *APS* 등으로 병원성을 보이는 종류들로 생명현상에 필요한 많은 대사산물을 자신들이 대사하지 않아도 되는 것이 이러한 결과를 보이는 하나의 원인이 될 수 있을 것이다. 그러나 *Mycobacterim* 종들과 같이 세포내에 기생하는 병원성세균이나 *Vibrio cholerae*에서 보존된 COG의 비율이 낮은 것으로 보아 더 많은 연구가 필요한 것으로 사료되었다. 한편 토양

Table 2. Percentage (%) of conserved COG over ortholog for each organism and number of conserved COG in Archaebacteria, *Firmicutes* and *Proteobacteria*.

Phylogenetic Group		organism	Abbreviation	number of ortholog	number of conserved COG	Percentage of conserved COG (%)
Archaea	Crenarchaeota	*Aeropyrum pernix*	Ape	1,202	340	28.29
	Euryarchaeota	*Archaeoglobus fulgidus*	Afu	1,958		17.36
		Halobacterium sp. NRC-1	Hbs	1,818		18.70
		Methanobacterium thermoautotrophicum	Mth	1,464		23.22
		Methanococcus jannaschii	Mja	1,407		24.16
		Pyrococcus abyssi	Pab	1,516		22.43
		Pyrococcus horikoshii	Pho	1,442		23.58
		Thermoplasma acidophilum	Tac	1,258		27.03
		Thermoplasma volcanium	Tvo	1,268		26.81
Bacteria	Firmicutes	*Bacillus halodurans*	Bha	3,032	224	7.39
		Bacillus subtilis	Bsu	3,030		7.39
		Lactococcus lactis	Lla	1,694		13.22
		Mycobacterium leprae	Mle	1,213		18.47
		Mycobacterium tuberculosis	Mtu	2,760		8.12
		Mycoplasma genitalium	Mge	396		56.57
		Mycoplasma pneumoniae	Mpn	441		50.79
		Streptococcus pyogenes	Spy	1,287		17.40
		Ureaplasma urealyticum	Uur	414		54.11
	Proteobacteria	*Buchnera sp. APS*	Buc	583	248	42.54
		Campylobacter jejuni	Cje	1,344		18.45
		Caulobacter crescentus	Ccr	2,880		8.61
		Escherichia coli K12	Eco	3,618		6.85
		Escherichia coli O157	EcZ	3,900		6.36
		Haemophilus influenzae	Hin	1,595		15.55
		Helicobacter pylori 26695	Hpy	1,135		21.85
		Mesorhizobium loti	Mlo	5,390		4.60
		Neisseria meningitidis MC58	Nme	1,555		15.95
		Neisseria meningitidis Z2491	NmA	1,540		16.10
		Pasteurella multocida	Pmu	1,838		13.49
		Pseudomonas aeruginosa	Pae	4,698		5.28
		Rickettsia prowazekii	Rpr	723		34.30
		Vibrio cholerae	Vch	2,998		8.27
		Xylella fastidiosa	Xfa	1,687		14.70

에서 발견되는 *Mezorhizobium loti*의 경우는 보존된 COG 비율이 4.60%로 가장 낮았다. 이는 다른 미생물들과 공통되는 기능이 낮다는 것으로, 다른 미생물들과 구별되는 유전자를 많이 함유하고 있을 가능성이 높다는 것을 내포하고 있다. 따라서 새로운 유용유전자 등을 발견할 확률이 상대적으로 높다고 할 수 있을 것이다.

특이한 것은 같은 종 내에서도 strain에 따라서 전체 ortholog에 대해 보존적 유전자의 비율이 아주 높지는 않다는 것이었다. 16S rRNA와의 비교·분석대상에는 포함되지 않지만 *Helicobacter pylori* 26695(1135 ortholog)와 *Helicobacter pylori J99*(1114 ortholog) 사이의 보존적 유전자가 889개에 불과하여 산술적으로 보면 전체 ortholog 중 20% 이상이 상이한 것으로 판명되었다. 그리고 *Neisseria meningitidis* MC58와 *Neisseria meningitidis* Z2491 사이에도 각 균주가 보유하는 ortholog의 25% 가량은 일치하지 않았다. 이는 전체 ortholog중 25%는 다른 종들과 COG를 형성하는 것으로 horizontal gene transfer에 의한 영향 등으로 사료되었다(11). 전체 gene 수에서 ortholog가 아닌 유전자의 개수까지 고려된다면 같은 종 내라도 보유유전자의 공통성은 줄어들 가능성이 큰 것으로 사료되었다.

유전자보유 계통수(gene content tree)

같은 종이라도 보유하는 COG에서 20% 이상의 차이를 보이는 것은 전체 게놈의 입장에서 보면 16S rRNA에 의존한 분류법에 변화가 있어야 한다는 것을 내포하고 있는 것으로 사료되었다. 이러한 관점에서 COG의 보유유무에 따른 neighbor joining analysis를 수행하여 유전자보유 계통수를 작성하였고 현재 널리 통용되는 16S rRNA 염기서열 유사성에 의한 neighbor joining analysis 결과와 비교를 수행하였다.

Archaeaebacteria

Fig. 1은 Archaeabacteria를 16S rRNA와 gene content를 이용하여 분류한 결과이다. 그림의 각 분기점에서 보이는 bootstrap number는 무작위로 표본을 1,000번 추출하였을 때 서열들이 같이 위치하는 개수로, 계통수에 있어서 분석대상의 서열이 다른 것과 얼마만큼 독립적인가를 나타내어 주는 지표로 이용할 수 있다. *Pyrococcus* 종들(Pho, Pab)과 *Thermoplasma* 종들(Tvo, Tac)은 16S rRNA와 gene content tree 두 가지 모두 아주 높은 bootstrap number로 같은 그룹에 위치하는 것을 알 수 있었다. *Methanobacterium thermoautotrophicum* (Mth)과 *Halo-bacterium sp. NRC-1* (Hab)의 경우는 16S rRNA에서는 높은 bootstap number로 같이 위치하였지만 유전자보유 계통수에서는 독립적인 것으로 나타났다. 이는 유전자보유 계통수와 16S rRNA 계통수가 일치하는 부분과 일치하지 않는 부분으로 나뉘어 진다는 것을 의미하는 것으로 horizontal gene transfer에 의한 영향 등이 원인이라 할 수 있을 것이다(11). 그리고 유전자보유 계통수에서 halophile이 독립적으로 위치한 것은 다른 archaebacteria와 구분되는 유전자 조성을 가지고 있는 것을 나타내는 결과가 될 수 있는 것이었다. 즉 호열성인 다른 archaeabacteria와 달리 고염도라는 서식지에 적합한 유전자 조성을 갖춘 점이 이러한 결과를 초래될 가능성도 있는 것으로 사료되었다.

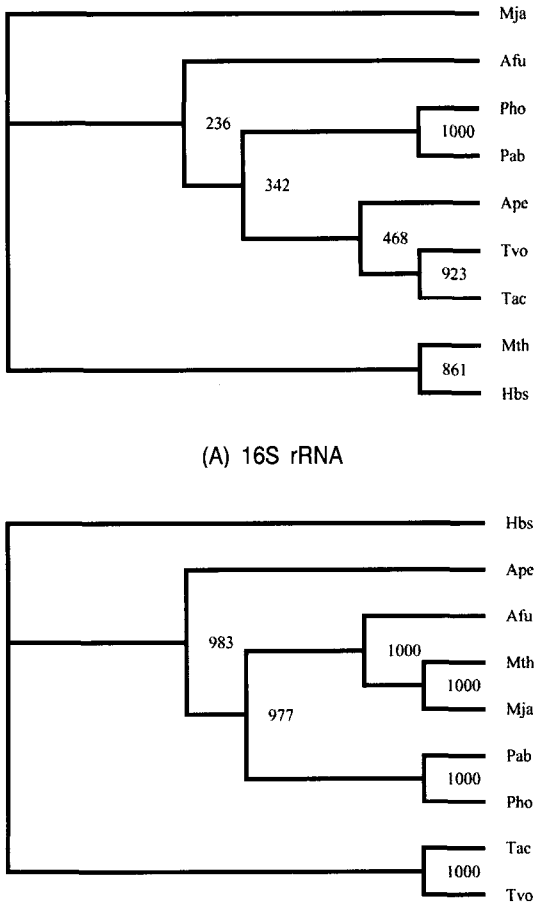

Figure 1. Comparison of the phylogenetic trees of Archaeabacteria obtained from neighbor-joining analysis of either 16S rRNA gene sequence (A) and gene content (B). Bootstrap values at each node are expressed as a number over 1,000 trials. Terminal branches have been extended for clarity and their length is therefore not meaningful.

Crenarchaeota인 *Aeropyrum pernix*가 Euryarchaeota와 함께 위치하는 것을 알 수 있었지만 bootstap number의 비율이 높지 않아 확신할 수는 없었다. *Thermoplasma acidophilum* (Tac)의 게놈염기서열에서 16S rRNA의 길이가 430bp로 나타났고 (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/rnatab?gi=168&db=Genome) alignment 수행결과 염기서열도 다른 archaeabacteria와 많이 다른 것으로 나타났는데 이로 인한 간섭의 결과로도 사료할 수 있었다.

Proteobacteria

Fig. 2는 *Proteobacteria*를 16S rRNA와 gene content를 이용하여 분류한 결과이다. 16S rRNA gene을 이용한 계통수에서는 alpha-와 epsilon-*Proteobacteria*사이의 유연관계가 아주 높은 것을 확인할 수 있었다. gamma-*Proteobacteria*는 *Xylella fasdidious*를 제외하고는 하나의 group을 형성하는 것을 알 수 있었고 bootstrap number가 70% 이하인 경우도 있어 내부적 응집성은 매우 높지는 않은 것을 알 수 있었다. *Xylella fasdidious*는 beta-*Proteobacteria*인 *Neisseria meningitidis*

MC58과 유연관계가 높은 것으로 드러났다. 전체 게놈에서 특정 유전자의 보유 유무에 따라 작성된 tree를 보면 16S rRNA와 같이 alpha- *Proteobacteria*인 *Caulobacter crescentus*와 *Mezorhizobium loti*, epsilon group인 *Helicobacter pyroli* 26695와 *Camphylobacter jejuni*, gamma-*Proteobacteria*인 *E. coli* 두 strain이 함께 분류되는 것을 알 수 있었다(Fig. 2). 한편 16S rRNA와 달리 alpha 그룹인 *Caulobacter crescentus*와 *Mezorhizobium loti*는 bootstap 비율은 낮지만 gamma 그룹과 함께 위치하는 것을 알 수 있었다. 이러한 현상은 parsimony method 등을 이용한 결과에서도 유사함을 확인하였다. alpha-와 epsilon-*Proteobacteria*의 경우 16S rRNA에서는 유연관계가 높은 것으로 나타났지만 유전자보유 측면에서는 유연관계가 16S rRNA보다 낮은 것을 알 수 있었다.

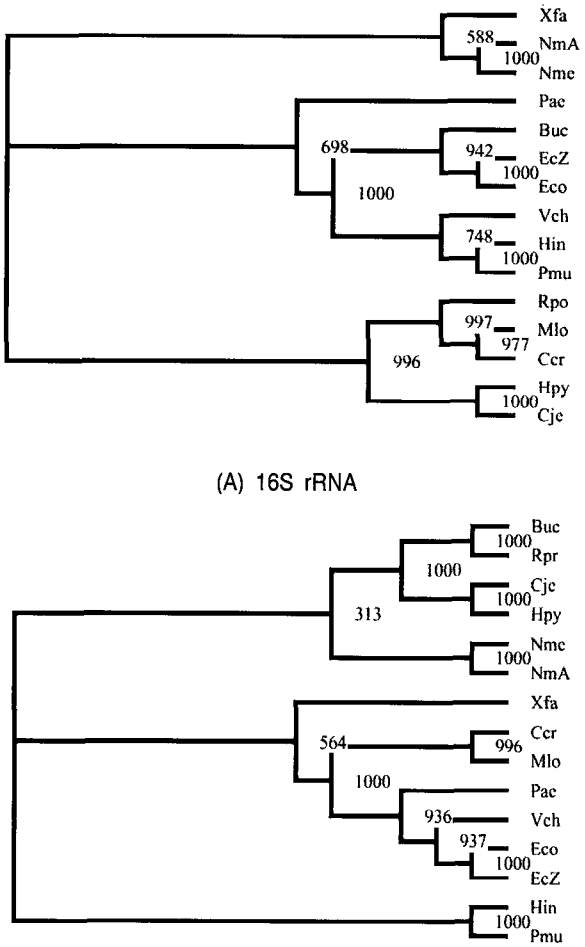

Figure 2. Comparison of the phylogenetic trees of *Proteobacteria* obtained from neighbor-joining analysis of either 16S rRNA gene sequence (A) and gene content (B). Bootstrap values at each node are expressed as a number over 1,000 trials. Terminal Branches have been extended for clarity and their length is therefore not meaningful.

Firmicutes

Fig. 3은 *Firmicutes*를 16S rRNA와 gene content를 이용하여 분류한 결과이다. 결과를 보면 16S rRNA 계통수와 유전자보유 계통수 모두 아주 높은 bootstrap number를 갖는 것으로 나타났다. *Bacillus* 종들(Bsa, Bsu)과 *Mycoplasma* 종들(Mge, Mpn) 그리고 *Mycobacterium* 종들(Mle, Mtu)은 두 계통수 모두에서 같은 그룹에 위치하는 것을 볼 수 있었다. 유전자보유 계통수에서 보이는 각 결절점(node)은, 그들의 상위 조상 결절점과 비교하여 추가되거나 혹은 사라진 유전자의 결과를 포함하는 게놈을 갖는, 공통 조상을 나타내는 것이라 할 것이다. 이 측면에서 보면 본 연구에서 분석한 *Firmicutes*에 속하는 종들은 공통조상으로부터 진화하면서 유전자조성이 서로 명확히 다르다는 것을 암시하는 것으로 해석할 수 있었다.

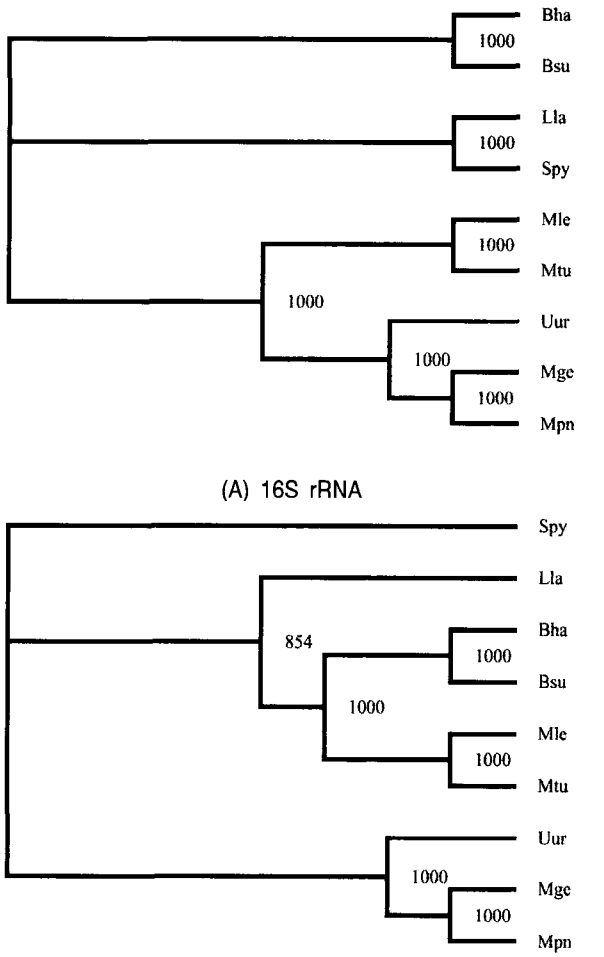

Figure 3. Comparison of the phylogenetic trees of *Firmicutes* obtained from neighbor-joining analysis of either 16S rRNA gene sequence (A) and gene content (B). Bootstrap values at each node are expressed as a number over 1,000 trials. Terminal Branches have been extended for clarity and their length is therefore not meaningful.

유전자보유 계통수의 적용과 유용성

16S rRNA 유전자는 세균들을 종 수준으로 구분할 수 있는 정보를 담고 있는 영역이며, 염기서열의 변화는 미생물간의 유연관계를 파악하는 데에 유용하다. 이는 16S rRNA의 특정 부분은 진화 속도가 매우 느리기 때문에 많은 생물체가 공통적으로 갖는 보존된 염기서열과 이차구조를 나타내는 사실에 기초하는 것으로 다양한 분류군의 상호 비교를 가능하

게 한다(10). 그러나 이는 게놈 전체에서 일어나는 염기서열의 변화가 아닌 하나의 유전자에 대한 염기서열의 유사성(similarity) 비교라는 한계를 갖는다(12). Eisen (5)은 서열의 유사성 자체보다는 어떤 경로를 통하여 유사하게 되었는지 파악하는 것이 중요하다고 하면서 진화학적 분석을 통한 미지의 유전자 예측을 시도하였다.

본 연구에서 수행한 게놈의 특정 COG 보유 유무에 따른 계통수 (phylogenetic tree)는 전체 게놈을 고려한다는 점과 16S rDNA처럼 보존적이지 않은 유전자까지 고려한다는 점이 특이한 점이라 할 수 있을 것이다. 하지만 각 genome에 대하여 binary sequence가 주어지므로 maximum parsimony와 bootstrap 같은 다양한 계통서열 분석법 (phylogenetic sequence analysis tool)이 이용 가능하다(12). 이러한 방법은 Fitz-Gibbon와 House(13) 그리고 Montague와 Hutchison(12) 등도 시도한 방법으로 Fitz-Gibson 등은 gene family를 구하는데 이용하였고 Montague 등은 herpesvirus 간의 phylogenetic tree를 구하는데 이용하였는데 genome rearrangement와 biological properties 등에 의한 분류와 일치하는 결과를 구하였다. Snel 등(14)과 Tekaia 등(15)도 genome에서 전체 유전자를 고려하여 계통수를 그렸지만 그들은 BeT (best hit) 알고리즘을 이용하였고 본 연구에서는 COG 알고리즘을 이용한 것이 차이라 할 것이다. Montague와 Hutchison(12)은 COG 알고리즘이 통계적 버림(statistical cutoff)를 이용한 BeT 알고리즘에 비해 민감도(sensitivity)와 설득력(stringency)이 높다고 하면서 우수성을 주장하였다. 이 방법은 수평적 유전자 전달(lateral gene transfer)에 의하여 진화적 계통수와 다른 계통수가 그려질 가능성이 있지만 이러한 가능성과 함께 전체 게놈에서 차지하는 비율은 높지 않을 것이라는 것을 추정할 수 있으므로 계통수를 작성하였고 neighborjoining method를 이용하였다(12).

COG를 이용한 분류법은 염기서열을 이용하여 일종의 기능적 측면에서 미생물을 분류하는 것으로 직접 실험을 수행해야하는 생화학적 분류법과 전체 게놈의 염기서열에 기초한 분류법의 중간자적 위치에 있는 분류법이라 할 수 있을 것이며, 유용한 단백질을 탐색하는 작업 등에 그 유용성이 높다고 할 수 있을 것이다.

사 사

본 연구는 과기부 21세기 프론티어 미생물유전체활용기술개발사업(과제번호 MG02-0101-001-1-0-0)에 의해 지원받았습니다.

요 약

유전자보유 유무에 따른 계통수와 16S rRNA에 의한 계통수를 염기서열 분석이 완료된 33종의 미생물에 대하여 neighbor joining method와 bootstrap method(n=1,000)를 이용하여 상관관계를 분석하였다. 각 분류그룹에서 공통적으로 보존된 COG와 각 미생물이 보유하고 있는 ortholog 수에 대한 비율을 조사한 결과, *Mezorhizobium loti*의 4.60% *Mycoplasma genitalium*의 56.57% 사이에 분포하는 것으로 파악되었다. 이는 미생물 종류에 따라서 공통 유전자의 보유정도가 차이를 보이는 것으로 독특한 유전자를 탐색할 수 있는 가능성을 제시하는 결과로 사료되었다. 그리고 같은 종 내에서도 20% 이상의 ortholog가 서로 독립적인 것을 알 수 있었다. Archaeabacteria와 *Proteobacteria* 그리고 *Firmicutes*모두 유전자보유 계통수와 16S rRNA 계통수가 일치하는 부분과 일치하지 않는 부분으로 나뉘어진다는 것을 알 수 있었다. 이러한 결과는 16S rDNA처럼 보존적이지 않은 유전자까지 고려한 결과이거나 horizontal gene transfer에 의한 영향 등으로 사료되었다. COG에 기초한 유전자보유 계통수는 생화학적 실험과 염기서열에 기초한 분류의 중간자적 입장에서 유용유전자 탐색에 이용될 수 있을 것이다.

REFERENCES

1. Woese, C. R. (1987), Bacterial evolution, *Microbiol. Rev.* **51**, 221-271.
2. Gupta, R. S. and E. Griffiths (2002), Critical issues in bacterial phylogeny, *Theor. Popul. Biol.* **61**, 423-434.
3. Henikoff, S., E. A. Greene, B. S. Pietrokovski, T. K. Attwood, and L. Hood (1997), Gene Families: The taxonomy of protein paralogs and chimeras, *Science* **278**, 609-614
4. Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin (2000), The COG database a tool for genome-scale analysis of protein functions and evolution, *Nucleic Acids Res.* **28**, 33-36
5. Eisen, J. A. (1998), Phylogenomics: improvoing functional predictions for uncharacterized genes by evolutionary analysis, *Genome Research* **8**, 163-187.
6. Tatusov, R. L., E. V. Koonin, and D. L. Lipman (1997), A genomic perspective on protein families, *Science* **278**, 631-637
7. ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria
8. http://www.ncbi.nlm.nih.gov/COG/
9. Kang, H. Y., C. J. Shin, B. C. Kang, J. H. Park, D. H. Shin, J. H. Choi, H. G. Cho, J. H. Cha, D. G. Lee, J. H. Lee, H. K. Park, and C. M. Kim (2002), Investigation of conserved gene in microbial genomes using *in silico* analysis, *Kor. J. Life Sci.* **5**, 610-621.
10. Amann, R., W. Ludwig, and K. H. Schleifer (1994), Identification of uncultured bacteria: a challenging task for molecular taxonomists. *ASM News* **60**, 360-365.
11. Jain, R, M. Rivera, and J. A. Lake (1999) Horizontal gene transfer among genomes: The complexity hypothesis, *Proc. Natl. Acad. Sci. USA.* **96**, 3801 -3806.
12. Montague, M. G. and C. A. Hutchison III (2000), Gene content phylogeny of herpesviruses, *Proc. Natl. Acad. Sci. USA.* **97**, 5334-5339.
13. Fitz-Gibbon, S. T. and C. H. House (1999), Whole genome-based phylogenetic analysis of free-living microorganisms, *Nucleic Acids Research* **27**, 4218-4222.
14. Snel, B. P. Bork, and M. A. Huynen (1999), Genome phylogeny based on gene content, *Nature Genetics* **21**, 108-110.
15. Tekaia, F., A. Lazcano, and B. Dujon (1999), The genomic trees as revealed from whole protein comparison, *Genome Res.* **9**, 550-557.