한국생물공학회지 제17권 제6호
Korean J. Biotechnol. Bioeng.
Vol. 17, No. 6, 560-565(2002)

COG 알고리즘을 통한 *Proteobacteria*의 보존적 유전자 파악

[1]이 동 근 · 강 호 영 · [1,2]이 재 화 · †김 철 민
부산대학교 유전공학연구소부설 생물정보학센터, 부산대학교 의과대학부설 부산지놈센터
[1]신라대학교 마린바이오산업화지원센터, [2]신라대학교 공과대학 생명공학과
(접수 : 2002. 11. 27., 게재승인 : 2002. 12. 24.)

Detection of Conserved Genes in *Proteobacteria* by using a COG Algorithm

Dong-Geun Lee[1], Ho-Young Kang, Jae-Hwa Lee[1,2], and Cheol-Min Kim†
Pusan Bioinformatics and Biocomplexity Research Center, Pusan National University
Busan Genome Center, College of Medicine, Pusan National University, Pusan 609-735, Korea
[1]Marine Biotechnology Center for Bio-Functional Material Industries, Silla University, Pusan 617-736, Korea
[2]Department of Bioscience and Biotechnology, Silla University, Pusan 617-736, Korea
(Received : 2002. 11. 27., Accepted : 2002. 12. 24.)

A COG(clusters of orthologous groups of proteins) algorithm was used to detect conserved genes within *Proteobacteria* and to figure out their relationships. Restricting comparison to the sequences of 42 procaryotes, 33 eubacteria and 16 *Proteobacteria*, the number of conserved genes was increased. All analyzed procaryotes shared 75 COGs. COG0195, COG0358 and COG0528 were only represented by the 42 procaryotes. Sixtyfour COGs were added as conserved genes in 33 eubacteria. Each *Proteobacteria* group has a unique repertoire of COGs. Metabolic COGs were more diverse in the beta *Proteobacteria* group than in the other groups. These results could be used to determine the origins and the evolutionary relationships of *Proteobacteria*. The possibilities of detecting new biological molecules is high in phylogenetically related organisms, hence the identification of useful proteins by using this algorithm is possible.

Key Words : Clusters of orthologous groups of proteins (COG), conserved gene, *Proteobacteria*

서 론

현재 지구상에 존재하는 생물들은 크게 원핵생물(procaryote)과 진핵생물(eucaryote)로 나눌 수 있고 원핵생물은 다시 고세균(archae-bacteria)과 진정세균(eubacteria)으로 나눌 수 있다. 계통발생학적 분류에 의하면 생물들을 bacteria, archaea, eukarya의 3 domain으로 분류하기도 한다. 단백세균(*Proteobacteria*)은 purple non-sulfur bacteria라고도 하며 다양한 그룹으로 진정세균 혹은 bacteria에 속하는 그룹으로 총 1534 species로 이루어져 알려진 종의 32.3%를 구성하는 것으로 알려져 왔다(1).

생명체는 화학적, 생물학적 진화를 통하여 형성된 것으로 생각되고 있다. 이에 따르면 생명기원의 시작인 원핵생물에서 시작된 공통조상의 유전자(ancestral gene)는 생물학적 진화를 거치면서 종분화(speciation)와 복사(duplication)에 의하여 각 생물의 유전체(genome)에 분포하게 되었고, 이러한 유전자 분포에 의하여 생물체의 공통 생명현상과 각 분류단위에 특이한 생명현상을 띄게 되었을 것으로 생각되어왔다. 따라서 계통발생학적 계보(phylogenetic lineage)에 있어 서로 다른 생명체들이 나타내는 각 분류단위의 독특한 생명현상과 모든 생명체가 공통적으로 나타내는 필수기능(housekeeping function)에 대한 이해는 생명 자체를 이해하는 핵심이라 할 것이며(2) 원핵생물에 대한 이해와 계통분류의 파악은 태초의 생명체와 진화를 밝히기 위해서 매우 중요한 것이라고 할 수 있을 것이다. 유용한 생체물질은 진화적으로 유연관계가 높은 생물 사이에 공통적으로 분포할 가능성이 높으므로 생물들의 계통학적 분류는 중요하다고 할 것이다.

컴퓨터와 정보공학의 발전으로 생물정보학이 발달하게 되었으며 이를 이용한 비교유전체학(comparative genomics) 으로 많은 생물체를 유전자 수준에서 관찰할 수 있게 되었다(4,5). 현재 유전자의 유사성 분석에 있어 개별 유전자의 염기서열을 배열(alignment)하는 단순서열비교가 주로 사용되고 있다. 기타 단백질의 3차원 구조를 이용하여 유사성을 판단하는 방법도 있으나, 현재까지는 다소 실용적이지 못하다고 할 수 있다.

† Corresponding Author : Busan Genome Center, College of Medicine, Pusan National University, Busan, Korea
Tel : +82-51-240-7736, Fax : +82-51-248-1118
E-mail : kimcm@pusan.ac.kr

Table 1. Studied genomes derived from COGs database, number and percentage of conserved genes for 16 *Proteobacteria*

Phylogenetic group	organism	number of ortholog	number of conserved gene	Percentage of conserved gene
alpha	*Caulobacter crescentus*	2,880	561	19.48
	Mesorhizobium loti	5,390		10.41
	Rickettsia prowazekii	723		77.59
beta	*Neisseria meningitidis MC58*	1,555	1163	74.79
	Neisseria meningitidis Z2491	1,540		75.52
gamma	*Escherichia coli O157*	3,900	390	10.00
	Escherichia coli K12	3,618		10.78
	Haemophilus influenzae	1,595		24.45
	Pasteurella multocida	1,838		21.22
	Pseudomonas aeruginosa	4,698		8.30
	Vibrio cholerae	2,998		13.01
	Xylella fastidiosa	1,687		23.12
	Buchnera sp. APS	583		66.90
epsilon	*Campylobacter jejuni*	1,344	786	58.48
	Helicobacter pylori 26695	1,135		69.25
	Helicobacter pylori J99	1,114		70.56

*Haemophilus influenzae*와 *Mycoplasma genitalium*의 두 genome에서 공통적으로 존재하는 유전자를 밝히는 연구가 있었는데(5), 이는 비교유전체학을 응용하여 보존적 유전자(conserved gene)를 밝히는 방식으로 진행되었으며 이를 바탕으로 최소기능유전자(minimal function gene)에 대한 많은 연구가 이어져 왔다(6-8).

Orthologs는 공통의 조상으로부터 종분화되어 서로 다른 종에 있는 유전자들의 집합으로 정의하며, paralogs는 한 유전체내에서 복사(duplication)로 생성된 유전자들을 총칭하는 용어이다(2). Clusters of Orthologous Groups of protein(COG)는 ortholog들에서 유래된 단백질의 집합을 이르는 말로 대개 유사한 구조와 기능을 갖는 것으로 알려져 있다(9-11). 각 COG는 적어도 3가지 이상의 계보에서 유래된 paralog 그룹 혹은 개별의 단백질들로 구성되어 있어 하나의 공통조상유전자(ancient conserved domain)에 해당하는 것으로 간주할 수 있다(2). 한편 단백질의 유사도를 이용한 COG 접근법은 미지의 단백질에 대한 기능 추측 등이 가능하므로 생물공학적인 측면에서도 이용가능성이 높다고 할 것이다.

본 연구에서는 생물정보학적인 접근방법으로 염기서열이 알려진 미생물 중 단백세균(*Proteobacteria*)이 가진 유전자집합(gene pool) 중 필수적으로 유지되고 있는 유전자의 존재를 확인하고자 하였다. 보존적 유전자들의 종류와 기능을 파악하고자 COG 알고리즘을 이용한 접근법을 시도하였고 42종의 원핵생물과 33종의 진정세균에 대하여도 동일한 접근법을 시도하여 각 분류단계별로 추가되는 보존적 유전자도 조사하였다.

재료 및 방법

재 료

분석에 이용된 42종의 원핵생물(procaryote) 유전체(microbial genome)는 National Center for Biotechnology Information (NCBI)의 공개 서버로부터 추출하였다(12). 미생물 유전자의 유사성에 관한 자료는 COGs에서 정리된 자료를 이용하였는데(13) 이들은 2002년 12월 현재 43종의 미생물 유전체를 ortholog 그룹으로 분류하여, 총 74,059개의 유전자들을 3,307개의 단백질 그룹으로 분류해 놓았다. 분석대상 원핵생물은 archaea가 9종, bacteria 중 *Firmicutes*가 9종, *Proteobacteria*가 16종, 기타 8종이었다(14). Table 1은 본 연구에서 분석한 16 *Proteobacteria*와 COG 자료를 나타내고 있다.

보존적 유전자 탐색

분석 방법은 강 등(14)의 방법을 이용하였다. 여러 분석 단계 중에서 CLUSTALW를 이용한 다중서열비교를 통해 distance value를 계산하였고(7,15), 자료의 분석과 정리에는 perl language(Practical Extraction and Report Language)를 사용하였다.

Figure 1은 본 연구에서 사용된 미생물 유전체의 분석 작업 순서도이다. NCBI의 공개 데이터베이스로부터 전체 유전체가 공개된 75종의 미생물 자료를 수집하여(12,16,17), 이들 유전체로만 구성된 로컬 데이터베이스를 제작하였다(14). 그리고 COGs 데이터베이스의 공개파일전송(ftp) 서버로부터 원핵생물에 대한 ortholog를 확보하였다(12). 이후 각 그룹별로 보존적 유전자의 항목을 배열화시키는 분석을 수행하였다. COGs 데이터베이스의 자료를 종(species)과 ortholog에 따라 2차원으로 재정렬하여 개별 ortholog를 종을 기준으로 분류하였다. 모든 원핵생물에서 공통적으로 관찰되는 COG들을 ancestral gene에서 유래된 보존적 유전자로 간주하였다(14).

위의 보존적 유전자 탐색 과정을 33종의 진정세균(eubacteria)과 16종의 단백세균(*Proteobacteria*)에 대하여 적용하여 각 분류단계별로 추가되는 보존적 유전자를 탐색하였다. *Proteobacteria*를 구성하는 alpha 그룹 3종, beta 그룹 2종, gamma 그룹 8종, 그리고 epsilon 그룹 3종에도 적용하여 각 그룹 사이의 보존적 유전자의 차이점과 공통점을 비교·분석

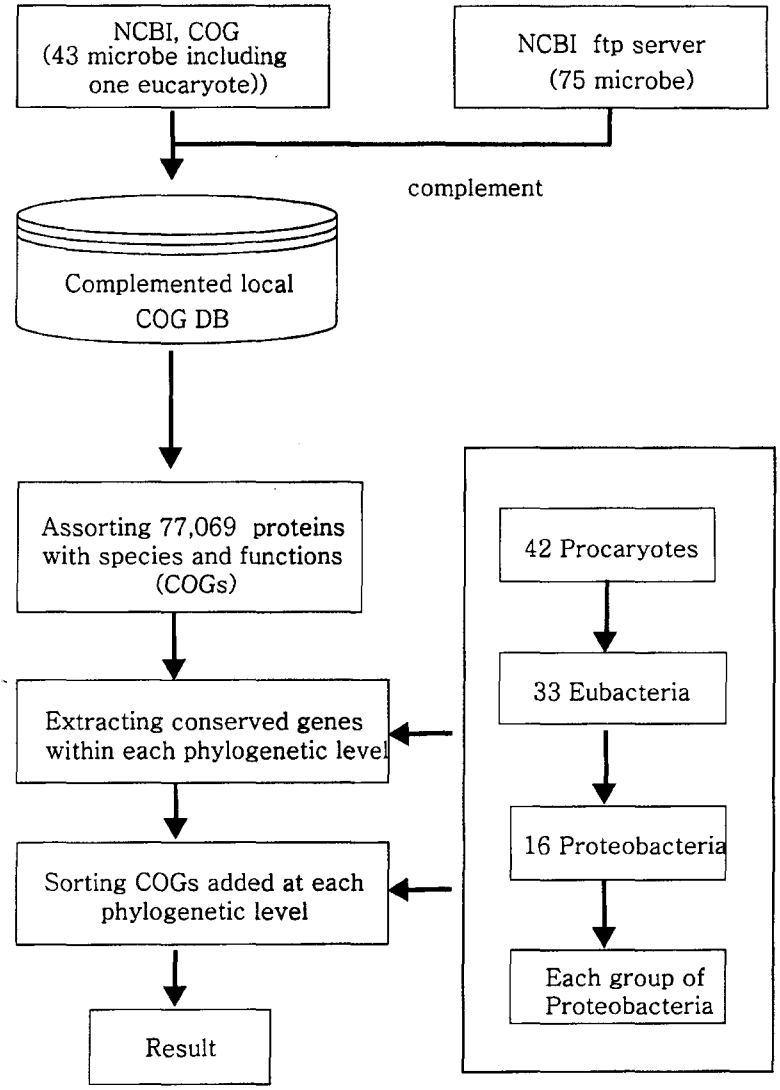


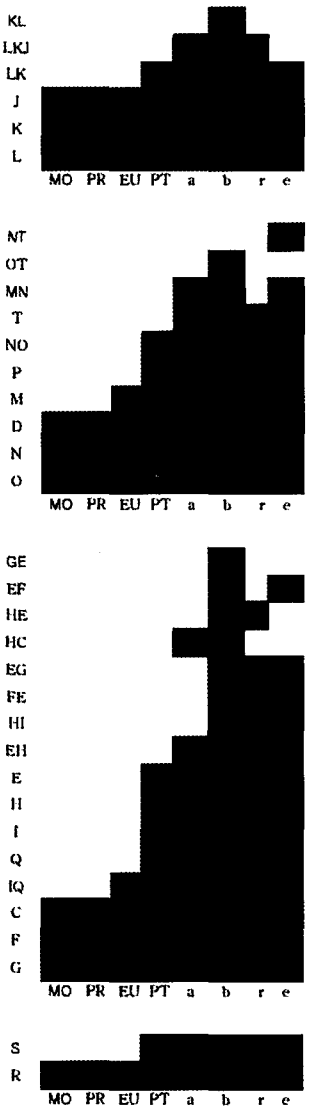
Figure 1. Flow chart to determine conserved genes within *Proteobacteria* using microbial genomes. Forty-three microbes cover 1 *Saccharomyces cerevisiae*, 9 archaebacteria, 16 proteobacteria, and 17 other eubacteria(see reference 14).

하였다.

결과 및 고찰

비교대상으로 하는 genome을 미생물, 원핵생물, 진정세균 및 단백세균 등으로 제한하면 보존적 유전자의 수가 늘어나고 각 분류단계별로 특이적인 COG들이 보존적이라는 것을 확인할 수 있었다. 이는 비교대상의 genome 수가 줄어들면서 공통적으로 분포하는 유사기능을 수행하는 유전자의 수가 증가될 가능성이 높다는 점과 각 분류단계별로 특징적인 COG들이 추가된 결과라고 할 수 있을 것이다.

이전의 연구(14)에서 원핵생물 42종과 효모 1종(*Saccharomyces cerevisiae*)이 공통적으로 소유하는 보존적 유전자를 검색하여 총 72종의 COG들은 검출하였고, 그들 대부분이 단백질의 합성에 관여된 것으로 전체 보존적 유전자 중 단일분류로 52개(72.2%)에 이른다는 것을 확인하여 생명현상에서 단백질 합성의 중요성을 유추할 수 있었다. 본 연구에서는 진핵생물인 효모를 제외하고 42종의 원핵생물이 공통적으로 보유하는 COG, 33종의 진정세균(eubacteria)에 공통적인 COG, 그리고 16종의 *Proteobacteria*에 공통적인 COG를 모두 조사하여 진화적인 관점에서 원핵생물들이 각 환경에 적응하여 추가되는 COG를 조사하고자 하였다. 원핵생물 42종은 효모균과 72종의 COG를 공유한 것 외에 3가지 COG를 더 공유하는 것으로 판명되었다. 즉 transcription elongation factor인 COG0195, bacterial type의 DNA primase인 COG0358 그리고 uridylate kinase 인 COG0528이 추가적인 것이었다. COG의 분류에 따르면 COG0195와 COG0358을 'information storage and processing' 기능분류에 속하는 것이었고 COG0528은 'metabolism' 기능분류에 속하는 것이었다. 이는 진핵생물과는 다른 transcription과 DNA replication 관련 단백질의 존재를 확인하는 것으로 유전정보를 처리하는 진화학적인 분기점이 추가되는 것을 추측할 수 있었다. 또한 이 유전자들은 *rpo*나 *hsp*과 같이 분류의 parameter로 이용할 가능성이 높은 것으로 사료되었다(in preparation). 보존적 유전자의 탐색을 33종의 진정



Information storage and processing

J : Translation, ribosomal structure and biogenesis
K : Transcription
L : DNA replication, recombination and repair

Cellular processes

D : Cell division and chromosome partitioning
O : Posttranslational modification, protein turnover, chaperones
M : Cell envelope biogenesis, outer membrane
N : Cell motility and secretion
P : Inorganic ion transport and metabolism
T : Signal transduction mechanisms

Metabolism

C : Energy production and conversion
G : Carbohydrate transport and metabolism
E : Amino acid transport and metabolism
F : Nucleotide transport and metabolism
H : Coenzyme metabolism
I : Lipid metabolism
Q : Secondary metabolites biosynthesis, transport and catabolism

Poorly characterid

R : General function prediction onlyze
S : Function unknown

Figure 2. Conserved functions according to phylogenetic group. Definition of functions were followed as COGs. Abbreviations in horizontal form represent phylogenetic group (MO; 42 procaryote and 1 eucaryote, PR; 42 procaryotes, EU; 33 eubacteria, PT; 16 *Proteobacteria*, a; 3 alpha-*Proteobacteria*, b: 2 beta-*Proteobacteria*, r; 8 gamma-*Proteobacteria*, e; 3 epsilon-*Proteobacteria*).

세균(eubacteria)에 대하여 시도한 결과 64개의 COG가 추가되어 총 139개의 보존적 유전자를 보였다. 진정세균(eubacteria) 그룹에서 ancient conserved domain 즉 보존적인 COG는 'celllular process' 기능분류 중 'cell envelope biogenesis'에 관여하는 methyltransferase(COG0275)와 diacylglyceryltransferase (COG0682)가 보존적인 것으로 나타났고 추가적으로 물질대사(metabolism) 관련 기능분류 중 acyl carrier protein(COG0236) 보존적인 것으로 판명되었다. Figure 2는 각 분류단계별로 보존적 유전자로 판명된 것을 COG web site의 기능분류(functional category) 별로 정리한 후 정성적으로 나타낸 그래프이고 Figure 3은 각 분류단계별로 추가된 보존적 유전자의 수를 각 기능분류별로 그리고 정량적으로 나타낸 것이다. 이를 통해 보면, 효모를 포함한 미생물 43종이나 42종의 원핵생물이나 기능분류 면에서 차이를 확연히 보이는 COG가 없다는 것을 알 수 있고(Figure 2) 33종의 진정세균(eubacteria) 그룹에서 M 계열(cell envelope biogenesis, outer membrane)과 IQ 계열(I-lipid metabolism, Q-secondary metabolite biosynthesis, transport and catabolism)의 COG들이 전술한 바와 같이 추가로 보존적인 것을 알 수 있었다. 추가로 확인된 COG를 보면 J계열이 30개로 ribosome의 large subunit를 구성하는 단백질이 12개, small subunit를 구성하는 단백질이 3개, tRNA charging에 필요한 단백질이 3개 그리고 기타 ribosome binding factor와 amino acid에 methyl 잔기와 formyl 잔기를 전달하는데 필요한 단백질이 보존적인 것으로 드러났다. 한편 K계열이 3개 그리고 L계열이 11개로 이들은 sigma factor(COG0568), transcription termination factor(COG0781), bacterial nucleoid DNA-binding protein(COG0776) 및 gyrase (COG0187, COG0188)였다. 이는 진화과정에서 archaea 또는 eukarya와 차별화되는 진화경로를 밟은 것으로 알려진 진정세균류의 공통조상에 내포되었던 것으로 사료되었다. *Proteobacteria*에서는 기능분류상(functional category) NO 계열인 protease (COG0616, COG0740)와 P 계열인 superoxide dismutase

(COG0605), rhodanese-related sulfurtransferases(COG0607), membrane protein TerC(COG0861), 그리고 LK 계열의 transcription-repair coupling factor - superfamily II helicase(COG1197) 등이 진정세균들의 보존적 유전자에서 볼 수 없었던 *Proteobacteria* 그룹에서만 발견되는 새로운 보존적 유전자들임을 확인하였다. 한편 metabolism에 관여하는 4 계열(E, H, I, Q)의 15 COG 그룹이 *Proteobacteria*에서 보존적인 것으로 나타났다. 이는 원시진정세균류에서 *Proteobacteria* 그룹으로 진화하면서 다양한 대사적 기구들이 독특한 ecological niche에 적응하는 과정에서 추가되었던 것으로 추측할 수 있었다.

각 *Proteobacteria* 그룹에서 추가로 보존적인 것으로 밝혀진 COG를 보면 기능계열 LK인 COG1200(recG-like helicases)이 gamma-*Proteobacteria*를 제외한 *Proteobacteria* 그룹들이 공통적으로 보유하였고, 기능계열 LKJ인 COG0513 (superfamily II DNA and RNA helicases)는 epsilon-*Proteobacteria*를 제외한 나머지 그룹들에 공통적이었다. 특히 beta-*Proteobacteria*는 기능계열 KL인 COG2901(factor for inversion stimulation Fis, transcriptional activator)가 추가로 보존적인 것으로 판명되었다. *Proteobacteria* 각 그룹 사이에서 가장 많은 차이가 나는 것은 metabolism에 해당하는 기능분류(functional category)로(Figure 2) beta 그룹이 351종, gamma 그룹은 98종, alpha 그룹이 116종, epsilon 그룹이 212종의 COG가 보존적인 것으로 나타났다(Figure 3). 피상적으로는 beta-*Proteobacteria*의 경우가 진화과정에서 아주 다양한 대사 경로가 추가되었다는 것을 알 수 있었다. 하지만 비교대상으로 하는 생물종의 수가 늘어날수록 공통적인 COG의 수는 감소할 수 밖에 없으므로 이러한 작용을 감안하여야 할 것이다(Table 1). 각 *Proteobacteria* 그룹에서 종의 수와 추가된 보존적 유전자 수를 보면 alpha 그룹은 116 COG/3 species, beta 그룹은 351 COG/2 species, gamma 그룹은 98 COG/8 species, epsilon 그룹이 212 COG/3 species로 나타났다. Alpha-*Proteobacteria*에 비해 gamma-*Proteobacteria*는 3배에 가까운 종 수에도 불구하고 보존적 유전자를 1.2 배 작은 것으로 드러났다. 이는 조사대상 gamma 그룹 서로간의 COG 보유유무의 차이가 높지 않을 것이라는 것과 각 구성원들의 유연관계가 높을 것이라는 것을 추가적으로 추측할 수 있었다. 앞으로 gamma 그룹 외에 다른 그룹의 *Proteobacteria*에서 더 많은 게놈 서열이 밝혀지면 서로 비교하면서 각 그룹 내에서 유연관계의 정도에 대한 파악이 가능할 것이다. Cellular process 기능분류에서 signal transduction mechanisms을 나타내는 T 기능계열은 *Proteobacteria*에서는 공통적인 COG가 관찰되지 않았지만 각 그룹에서는 T 기능계열에 속하는 서로 다른 COG들이 검출되었다(Figure 1). 이러한 양상은 metabolism 기능분류에서의 EH 기능계열에서도 관찰되었다. 이는 horizontal gene transfer 등 서로 다른 생물그룹 사이에서 보이는 형질의 독자적 획득과 상실에 의한 homoplasy에 의한 영향일 것으로 추측되었다(18). COG의 응용 장점은 기존의 자료와 유전자 서열만으로 미지의 유전자에 대한 기능추측이 가능하다는 것이므로 본 연구에서 적용된 것과 같은 방법으로 극한 미생물 등에서 유용한 효소를 탐색하는 일이 충분히 가능할 것으로 사료된다(in preparation).

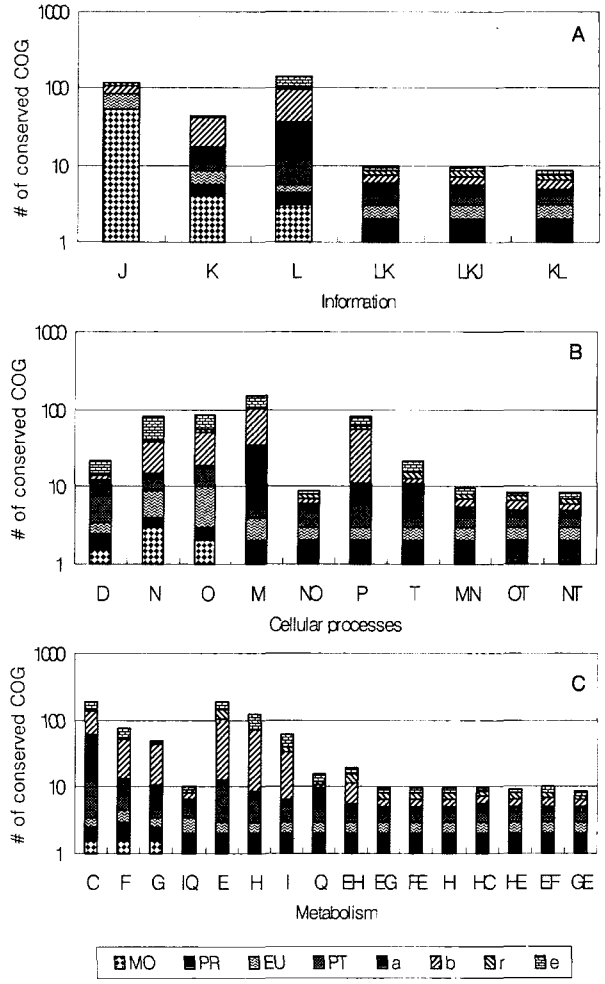


Figure 3. Number of newly detected conserved genes within each phylogenetic group in the functional grouping of information storage and processing(A), cellular processes(B), and metabolism(C). Characters in X-axis represent the functional category of COG(see Fig. 2) and characters in box represent phylogenetic group(MO: 42 procaryotes and 1 *S. cerevisiae*, PR: 42 procaryotes, EU: 33 eubacteria, PT: 16 *Proteobacteria*, a: 3 alpha-*Proteobacteria*, b: 2 beta-*Proteobacteria*, r: 8 gamma-*Proteobacteria*, e: 3 epsilon-*Proteobacteria*).

요 약

생태계에서 중요한 역할을 담당하는 단백세균(*Proteobacteria*)의 보존적유전자(conserved gene)를 파악하고 서로간의 유연관계를 밝히고자 clusters of orthologous groups of proteins (COG) 알고리즘을 이용한 접근법을 시도하였다. 42종의 원핵생물과 33종의 진정세균, 16종의 단백세균으로 가면서 보존적유전자가 증가하는 것을 확인하였다. 분석대상 원핵생물 모두에서 75종의 COG 즉 보존적 유전자가 관찰되었다. COG0195, COG0358 그리고 COG0528은 원핵생물에서만 관찰되어 새로운 분류의 parameter로 이용될 가능성이 있는 것으로 추측되었다. 64종류의 보존적 유전자가 33종의 진정세균(eubacteria)에서 관찰되었다. 이는 각 분류단계를 특징짓는 새로운 COG의 추가에 의한 결과로 사료되었다. 각 단백세균 그룹은 독자적인 COG 레퍼토리를 소유하였으며 물질대사에

관련된 보존적 유전자는 beta 그룹이 다른 그룹에 비해 다양한 것을 확인하였다. 본 연구는 단백세균의 기원과 진화적 유연관계를 파악하는데 도움을 줄뿐만 아니라 향후 세균분류학과 생명공학에 필수적인 유용유전자 탐색 등에서도 충분한 연구가치가 있는 것으로 사료되었다.

REFERENCES

1. Garrity, G. (2001), Bergey's manual of Systematic Bacteriology, 2nd ed., Springer-Verlag, New York.
2. Tatusov, R. L., E. V. Koonin, and D. L. Lipman (1997), A genomic perspective on protein families, *Science* **278**, 631-637.
3. Woese, C. R. (1987), Bacterial evolution, *Microbiol. Rev.* **51**, 221-271.
4. Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, and J. C. Venter (1995), The minimal gene complement of *mycoplasma genitalium, Science* **270**, 397-403.
5. Mushegian, A. and E. V. Koonin (1996), A minimal gene set for cellular life derived by comparision of complete bacterial genomes, *Proc. Natl. Acad. Sci. USA.* **93**, 10268-10273.
6. Lehoux, D. E., F. Sanschagrin, and R. C. Levesque (2001), Discovering essential and infection-related genes, *Curr. Opin. Microbiol.* **4**, 515-519.
7. Mushegian, A. (1999), The minimal genome concept, *Curr. Opin. Genet.* **9**, 709-714.
8. Reich, K. A.(2000), The search for essential genes, *Res. Microbiol.* **151**, 319-320.
9. 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.
10. 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.
11. Eisen, J. A. (1998), Phylogenomics: improvoing functional predictions for uncharacterized genes by evolutionary analysis, *Genome Res.* **8**, 163-187.
12. ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria.
13. http://www.ncbi.nlm.nih.gov/COG/.
14. 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, and C. -M. Kim (2002) Investigation of Conserved Gene in Microbial Genomes using *in silico* Analysis, *Kor. J. Life Sci.* **5**, 610-621.
15. Kimura, M. (1983), The Neutral Theory of Molecular Evolution, Cambridge University Press, London.
16. Durbin, R. S. R. Eddy, A. Krogh, and G. Mitchison (1998), Biological sequence analysis(Probabilistic models of proteins and nucleic acids, pp134-159 Cambridge University Press, London.
17. Saitou, N. and M. Nei (1987), The neighbor-joining method,a new method for reconstructing method for reconstructing phylogenetic trees, *Mol. Biol. Evol.* **4**, 406-425.
18. Givnish, T. J. and K. J. Systma (1997), Consistency, characters, and the likelihood of correct phylogenetic inference, *Mol. Phylogen. Evol.* **7**, 320-330.