COG 알고리즘으로 파악한 *Proteobacteria*의 보존적 유전자

이동근[1], 이진옥, 이재화
[1]신라대학교 마린바이오산업화지원센터, 신라대학교 생명공학과 발효공학연구실
전화 (051) 999-5748, FAX (051) 999-5636

Abstract

A COG (clusters of orthologous groups of proteins) algorithm, protein similarities among genomes, was used to detect conserved genes and to figure out their relationships within 42 procaryote, 33 *Bacteria* and 16 *Proteobacteria* All analyzed procaryotes shared 75 COGs. COG0195, COG0358 and COG0528 were only represented by the 42 procaryotes. Sixty-four 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. 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.

서 론

*Proteobacteria*는 purple non-sulfur bacteria라고도 하며 다양한 그룹으로 진정세균(*Bacteria*)에 속하는 그룹으로 총 1534 species로 이루어져 알려진 미생물종의 32.3%를 구성하는 것으로 알려져 왔다. 유용한 생체물질은 진화적으로 유연관계가 높은 생물 사이에 공통적으로 분포할 가능성이 높으므로 유연관계에 대한 이해는 중요하다고 할 것이다. 컴퓨터와 정보공학의 발전으로 생물정보학이 발달하여 이를 이용한 비교 유전체학 (comparative genomics)으로 생물체들을 게놈 수준에서 관찰할 수 있게 되었다[1]. Orthologs는 공통의 조상으로부터 종분화되어 서로 다른 종에 있는 유전자들의 집합으로 정의하며, paralogs는 한 유전체내에서 복사 (duplication)로 생성된 유전자들을 총칭하는 용어이다[2]. COG는 ortholog들에서 유래된 단백질의 집합을 이르는 말로 대개 유사한 구조와 기능을 갖는 것으로 알려져 있다[3]. 단백질의 유사도를 이용한 COG 접근법은 미지의 단백질에 대한 기능 추측 등이 가능하므로 생물공학적인 측면에서도 이용가능성이 높다고 할 것이다. 본 연구에서는 COG 알고리즘의 접근방식으

로 게놈염기서열이 알려진 미생물 중 42종의 procaryote, 33종의 *Bacteria,* 16종의 *Proteobacteria*가 가진 유전자집합 (gene pool)중 보존적으로 유지되고 있는 유전자의 존재를 확인하고자 하였다.

재료 및 방법

분석에 이용된 42종의 원핵생물 (procaryote) 유전체 (microbial genome)는 National Center for Biotechnology Information (NCBI)의 공개 서버에서[4] 미생물 유전자의 유사성에 관한 자료는 COGs에서 정리된 자료를 이용하였다[5]. 분석대상 원핵생물은 *Archaea*가 9종, *Bacteria*중 *Firmicutes*가 9종, *Proteobacteria*가 16종, 기타 8종이었다[6]. Table 1은 본 연구에서 분석한 16 *Proteobacteria*와 COG 자료를 나타내고 있다.

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

분석 방법은 강 등[6]의 방법을 이용하였다. COGs 데이터베이스의 공개파일전송 (ftp) 서버를 이용하여 로컬 데이터베이스를 제작한 후 비교대상 분류수준에서 공통적으로 관찰되는 COG들을 ancestral gene에서 유래된 보존적 유전자로 간주하였다[6].

결과 및 고찰

Fig. 1은 각 분류단계별로 보존적 유전자로 판명된 것을 COG web site의 기능분류 (functional category) 별로 정리한 그래프이고 그림속의 숫자는 각 분류단계와 기능분

류별로 보존적 COG의 개수를 나타낸다.

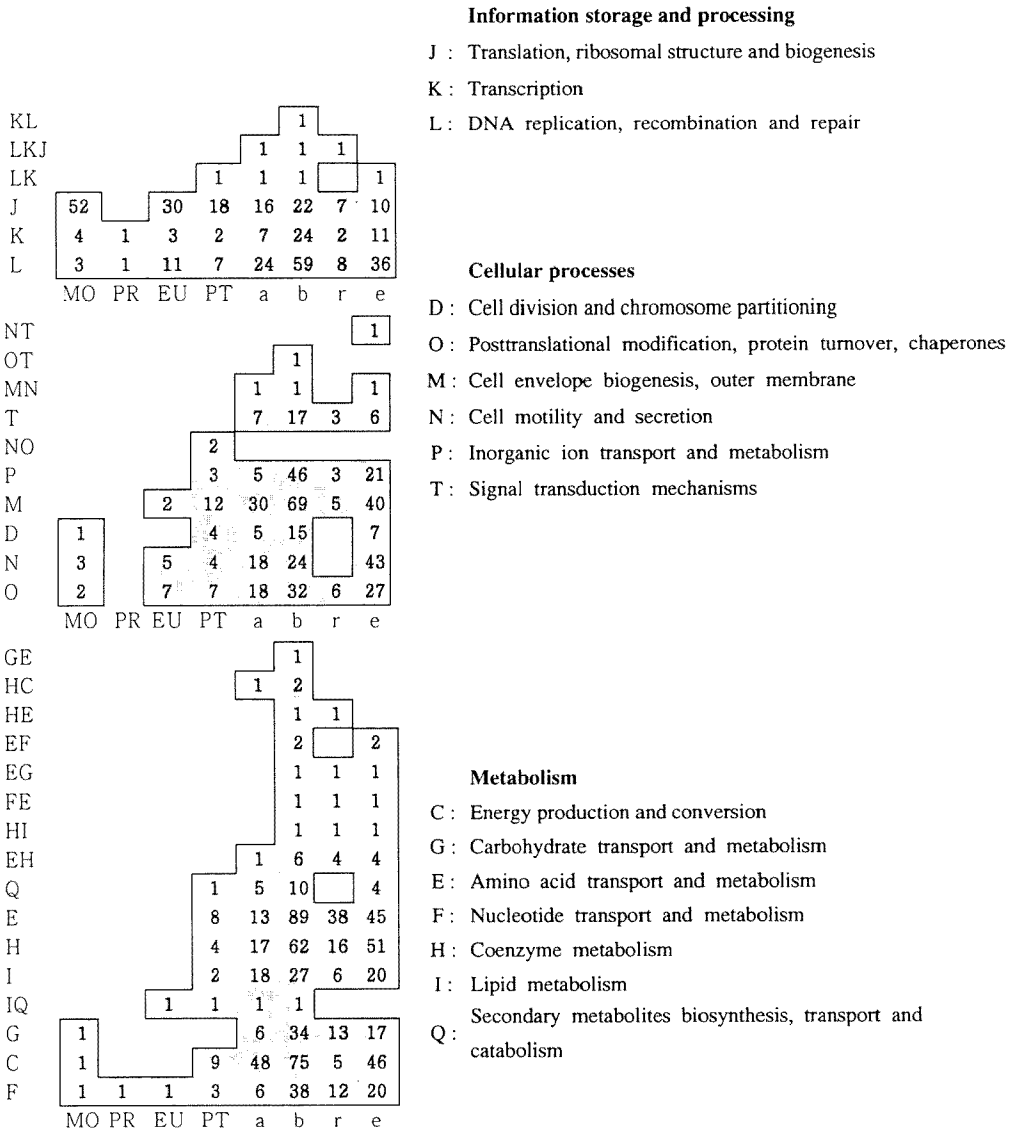

Fig. 1. Conserved functions according to phylogenetic group. Definition of functions were followed as COGs. Unknown functions were omitted. Abbreviations in horizontal form represent phylogenetic group (MO; 42 procaryote and 1 eucaryote, PR; 42 procaryotes, EU; 33 eubacteria, PT; 16 *Proteobacteria*,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

Procaryote 42종은 72종 COG외에 3가지 COG 즉 COG0195 (transcription elongation factor), COG0358 (bacterial type DNA primase), COG0528 (uridylate kinase)이 추가로 보존적인 것이었다. *Bacteria* 33종에서는 64개의 COG가 추가되어 총 139개의 보존적 유전자를 보였다. 물질대사에 관여하는 4 계열 (E, H ,I, Q)의 15 COG 그룹이 *Proteobacteria*에서 보존적이었으며 *Proteobacteria* 각 그룹사이에서 가장 많은 차이가 나는 것도 metabolism에 해당하는 기능분류 (functional category)로 (Fig. 1) beta 그룹이 351종, gamma 그룹은 98종, alpha 그룹이 116종, epsilon 그룹이 212종의 COG가 보존적인 것으로 나타났다. COG의 응용 장점은 기존의 자료와 유전자 서열만으로 미지의 유전자에 대한 기능 추측이 가능하다는 것이므로 본 연구에서 적용된 것과 같은 방법으로 극한 미생물 등에서 유용한 효소를 탐색하는 일이 충분히 가능할 것으로 사료된다.

요 약

Clusters of orthologous groups of proteins (COG) 알고리즘을 이용하여 42 종의 원핵생물, 33종의 진정세균, 16종의 단백세균 수준에서의 보존적 유전자 (conserved gene)를 파악하였다. 분석대상 원핵생물 모두 75종의 COG 즉 보존적 유전자가 관찰되었다. COG0195, COG0358 그리고 COG0528은 원핵생물에서만 관찰되었고 64종류의 보존적 유전자가 33종의 *Bacteria*에서 관찰되었다. 각 분류단계를 특징짓는 새로운 COG의 추가를 확인하였고 각 단백세균 그룹은 독자적인 COG 레퍼토리를 소유하였으며 물질대사에 관련된 보존적 유전자는 beta 그룹이 다른 그룹에 비해 다양한 것을 확인하였다. 본 연구는 *Proteobacteria*의 기원과 진화적 유연관계를 파악하는데 도움을 줄 뿐만 아니라 향후 세균분류학과 생명공학에 필수적인 유용유전자 탐색 등에서도 충분한 연구가치가 있는 것으로 사료되었다.

References

1. Mushegian, A. and E. V. Koonin, A minimal gene set for cellular life derived by comparision of complete bacterial genomes (1996), *Proc. Natl. Acad. Sci. USA.* **93**, 10268-10273
2. Tatusov, R. L., E. V. Koonin, and D. L. Lipman, A genomic perspective on protein families (1997), *Science* **278**, 631-637
3. 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
4. ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria
5. http://www.ncbi.nlm.nih.gov/COG/
6. 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, Investigation of Conserved Gene in Microbial Genomes using *in silico* Analysis (2002), *Korean Journal of Life Sci.* **5**, 610-621