Korean Journal of Life Science
Vol. 12. No. 5. 610~621, 2002

미생물 유전체의 *in silico* 분석에 의한 보존적 유전자 탐색

강호영 · 신창진 · 강병철 · 박준형 · 신동훈 · 최정현 · 조환규 · 차재호 · 이동근[1] · 이재화[1,2] · 박희경[3] · 김철민[†]

부산대학교 유전공학연구소부설 생물정보학센터
부산대학교 의과대학부설 부산지놈센터
[1]신라대학교 생명공학연구소
[2]신라대학교 공과대학 생명공학과
[3](주)에스제이하이테크 부설연구소

Investigation of Conserved Gene in Microbial Genomes using *in silico* Analysis

Ho-Young Kang, Chang-Jin Shin, Byung-Chul Kang, Jun-Hyung Park, Dong-Hoon Shin, Jeong-Hyeon Choi, Hwan-Gyu Cho, Jae-Hoo Cha, Dong-Geun Lee[1], Jae-Hwa Lee[1,2], Hee-Kyung Park[3] and Cheol-Min Kim[†]

Pusan Bioinformatics and Biocomplexity Research center, Pusan National University
Busan Genome Center, College of Medicine, Pusan National University
[1]Biotechnology Research Institute, Silla University
[2]Department of Bioscience and Biotechnology, Silla University
[3]Institute for Biomedical Research, SJ-Hightech Co.,Ltd.

Abstract

Conserved genes are importantly used to understand the major function in survival and replication of living organism. This study was focused on identification of conserved genes in microbial species and measuring the degree of conservation. For this purpose, *in silico* analysis was performed to search conserved genes based on the conservation level within microbial species. The ortholog list of COGs (Clusters of Orthologous Groups of proteins) in NCBI was used and whole genomes of 43 microbial species were included in that list. The distance value, derived from CLUSTALW multiple alignment program, was used as a descriptor of the conservation level of orthologs. It was revealed that 43 microbial genomes hold 72 conserved orthologs in common. The majority(72.2%) of the conserved genes was related to "translation, ribosomal structure and biogenesis" functional category. A GTPase-translation elogation factor(COG0050) was the best conserved gene from the distance value analysis. The 72 conserved genes, found in this research, would be useful not only to study minimal function genes but also new drug target among pathogens and to make a model of the virtual cell.

Key words – conserved gene, microbial genome, ortholog, COG, minimal function gene, *in silico* analysis

*To whom all correspondence should be addressed
Tel : 051-240-7736, Fax : 051-248-1118
E-mail : kimcm@pusan.ac.kr

서 론

현재 지구상에 존재하는 생물들은 화학적 진화와 생물학적 진화를 통하여 생성된 것으로 생각되고 있다. 생물학적 진화기간 동안 공통조상의 유전자(ancestral gene)는 종분화(speciation)와 복사(duplication)에 의하여 각 생물의 유전체(genome)에 분포하게 되었을 것이고, 생물체의 공통 생명현상과 각 분류단위에 특이한 생명현상을 띄게 되었을 것이다. 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 것이다[23].

이러한 기능을 내재하고 표현하는 유전자와 단백질에 대한 정보는 Watson과 Crick에 의해 DNA의 2중나선 구조가 밝혀진 이후 발달한 염기서열 분석과 아미노산서열 분석으로 서열(sequence)의 형태로 자료로 많이 확보하게 되었으며, 그들 중 일부는 기능이 밝혀져 있는 상태다. 이후 방대한 자료들을 조직적으로 추가, 관리, 공유할 수 있는 데이터베이스를 관리하는 GenBank, NCBI(National Center for Biotechnological Information), PIR(Protein Information Resource) 등이 생기게 되었다. 이들 데이터베이스를 이용하면서 유전자와 단백질의 서열과 기능을 통한 생물체의 진화학적 접근 등으로 생물상호간의 유연관계를 밝힐 수 있게 되었다. 방대해진 데이터베이스 처리에 관한 기법은 컴퓨터의 발달과 패를 같이 한다고 할 수 있을 것이다. 정보공학의 발전으로 생물정보학이 발달하게 되었으며 이를 이용한 비교유전체학(comparative genomics)으로 많은 생물체를 유전자 수준에서 관찰할 수 있게 되었다 [5,16]. 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.

최소기능유전자(minimal function gene)조합의 탐색을 위한 연구 [15,19]의 해결에는 유전자결손기법(gene knockout method) 등을 이용한 접근법이 사용되어 왔으나, 실제 기법상의 어려움과 분석상의 난점으로 인해 다양한 결과가 나오지 못하였다 [10,11,14,24]. 하지만 1995년 *Haemophilus influenzae*의 전체 유전체 염기서열이 최초로 밝혀진 직후 나온 *Mycoplasma genitalium*의 전체 유전체 염기서열에 대한 정보는 이러한 의문의 해소에 획기적인 전기를 가져오게 되었다[5]. 또한 *Haemophilus influenzae*와 *Mycoplasma genitalium*의 두 유전체를 이용하여 두 종에서 공통으로 존재하는 유전자를 밝히는 연구가 있었다 [16]. 이 선구적인 연구는 비교유전체학을 응용하여 보존적 유전자를 밝히는 방식으로 진행되었으며, 이를 바탕으로 최소기능유전자에 대한 많은 연구가 이어져 왔다[14,15,19].

Orthologs는 공통의 조상으로부터 종분화되어 서로 다른 종에 있는 유전자들의 집합으로 정의하며, 같은 ortholog 내의 구성원들은 서열의 유사성을 갖게 된다. 따라서 이러한 구성원들은 진화의 과정에 있어서도 같은 기능을 나타내게 된다. 한편 paralogs는 한 유전체내에서 복사(duplicatoin)로 생성된 유전자들을 총칭하는 용어이며, 새로운 기능들로 진화되어 구성원 서로간의 공통 기능은 거의 없게 된다[23].

미생물들은 현재 지구 환경권 내에서 가장 큰 생명계(biosphere)를 형성하고 있으며, 다세포 생명체가 적응하지 못한 다양한 극한환경(extreme environment)에도 적응되어 그에 맞는 다양한 유전적 특질들을 지니게 되었다. 다른 많은 생명체들이 변화하는 환경에 적응하기 위해 다양한 진화적 산물들을 획득하는 과정에서 그들의 조상과의 공통분모를 잃어 가고 있을 때, 미생물(microorganism)들은 단세포로서 변화가 상대적으로 적은 미시적 환경속에서 그들 조상과의 공통점을 오랜 시간 유지해 왔다 [4]. 생명의 유지와 번식이라는 본질적인 기능은 최초의 조상으로부터 받은 유전자들에 계승되고 있으며, 이러한 유전자들은 미생물종 상호간에 많은 유사성을 지니고 있어서 비교유전체학(comparative genomics)을 이용한 유전자 수준에서의 관찰은 최소한 두 종 이상의 유전체에 모두 존재하는 유전자인 보존적 유전자(conserved gene)들을 찾아내는데 아주 유용할 것이다[9]. 보존적 유전자 집합 중 현재의 지구 환경에서 세포생명체의 생존과 번식에 필요한 최소한의 기능유전자들의 조합을 최소기능유전자(minimal function gene)라고 한다 [15]. 즉 모든 세포생명체의 보존적 유전자를 추출해 내면 기본적인 최소기능유전자 조합의 후보 유전자들을 얻을 수 있을 것이다[13,15,19]. COG (Clusters of Orthologous Groups of protein)는 ortholog들에서 유래된 단백질의 집합을 이르

는 말로 대개 유사한 구조와 기능을 갖는 것으로 알려져 있다 [7, 22]. 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 [23].

본 연구는 생물정보학적인 접근방법으로 염기서열이 알려진 모든 미생물이 가진 유전자집합(gene pool)중 필수적으로 유지되고 있는 유전자의 존재를 확인하고, 그 유전자들의 종류와 기능 그리고 보존성의 정도를 파악하고자 COG를 통한 접근법을 시도하였다.

재료 및 방법

재료

분석에 이용된 미생물 유전체(microbial genome)는 NCBI의 공개 서버로부터 추출하였다 [6]. 유전체들은 2002년 4월을 기준으로 총 75종으로 구성되어 있었으며, 모든 자료는 유전자 동정이 1차로 완료되어 있는 상태였다. 그리고 동정된 유전체 중 약 35~50% 정도의 유전자들은 기능이 알려지지 않은 ORF(open reading frame)로 알려져있다 [4]. 한편 미생물 유전자의 유사성에 관한 자료는 COGs에서 정리된 자료를 이용하였는데 [8] 이들은 43종의 미생물 유전체를 ortholog 그룹으로 분류하여, 총 77,069개의 유전자들을 3,852개의 단백질 그룹으로 분류해 놓았다 [8]. COGs의 유전자 목록을 조사하여 일부 미생물의 경우 7개의 유전자에 대한 자료가 COGs에서 누락된 것을 확인하고 NCBI의 자료를 이용하여 보강하였다[6]. Table 1은 실제로 분석한 자료인 43종의 미생물을 나타내고 있다.

분석 방법

분석 방법은 미생물 유전체 분석 작업 순서도 (Fig. 1)의 내용과 동일한 순서로 수행하였다. Figure 1의 각 분석 단계는 각각의 출력을 가지며 다음 단계로 출력을 전달한다. 여러 분석 단계 중에서 CLUSTALW를 이용한 다중서열비교를 통해 distance value를 계산했고 [12, 15], 자료의 분석과 정리에는 perl language (Practical Extraction and Report Language)를 사용했으며, 자료의 통계적 처리에는 통계 패키지인 SPSS(ver. 10.0)를 이용하였다.

각 분석단계별 과정은 다음과 같았다. NCBI의 공개 데이

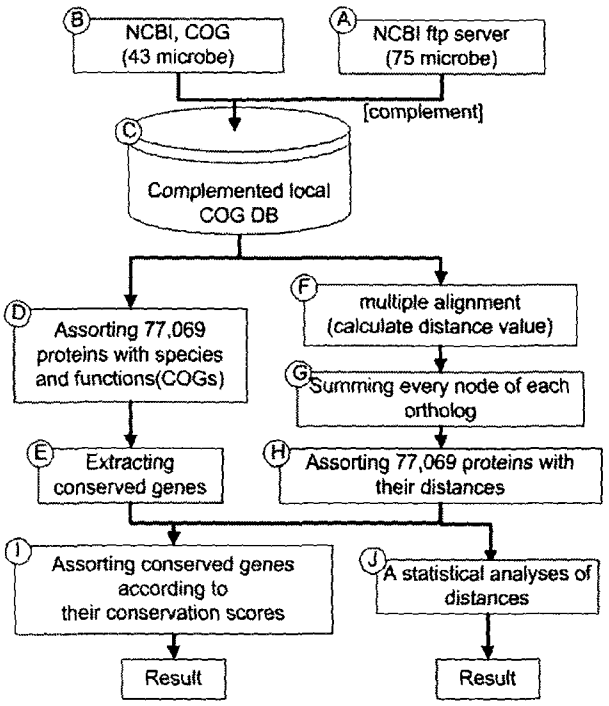


Fig. 1. Flow chart for analysis of microbial genomes. Total 43 genomes were analyzed ; 33 eubacteria, 9 archaebacteria and 1 eucaryote.

터베이스로부터 전체 유전체가 공개된 75종의 미생물 자료를 수집하여 [3,6,20], 미생물 유전체로만 구성된 로컬 데이터베이스를 제작하였다 (Fig. 1-A). 그리고 COGs 데이터베이스의 공개파일전송(ftp) 서버로부터 총43종의 미생물 (archaebacteria 9종, eubacteria 33종, eucaryote 1종)에 대한 77,069개의 ortholog를 확보하였다 (Fig. 1-B) [8]. COGs 데이터베이스로부터 확보된 유전자 항목과 NCBI로부터 획득한 자료를 비교하여, 자료의 무결성을 검증하여 7개의 COGs 유전자 항목이 누락된 것을 확인하였으며, 미생물 유전체의 로컬데이터베이스를 이용하여 이를 수정하여 43종의 미생물 유전체에 대한 아미노산 서열을 추출하였다 (Fig 1-C). 이후 두 가지 분석을 수행하였다. 하나는 보존적 유전자의 항목을 배열화시키는 분석이었고 다른 하나는 distance value를 배열화시키기 위해 다중서열비교를 수행하는 분석이었다.

COGs 데이터 베이스의 자료는 단백질의 기능적 분류에 따라 정렬되어 있는데 이러한 1차원의 자료를 2차원 배열로 변환하여 종(species)과 ortholog에 따라 재정렬하여 개별 ortholog를 종을 기준으로 분류하였다 (Fig. 1-D). 분석

Table 1 : Studied 43 genomes derived from COGs database.

Phylogenetic Group		organism	Abbreviation	number of ortholog
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
		Pyrococcus horikoshii	Pho	1,442
		Thermoplasma acidophilum	Tac	1,258
		Thermoplasma volcanium	Tvo	1,268
Bacteria	Aquificales	*Aquifex aeolicus*	Aae	1,377
	Cyanobacteria	*Synechocystis*	Syn	2,369
	Firmicutes	*Bacillus halodurans*	Bha	3,032
		Bacillus subtilis	Bsu	3,030
		Lactococcus lactis	Lla	1,694
		Mycobacterium leprae	Mle	1,213
		Mycobacterium tuberculosis	Mtu	2,760
		Mycoplasma genitalium	Mge	396
		Mycoplasma pneumoniae	Mpn	441
		Streptococcus pyogenes	Spy	1,287
		Ureaplasma urealyticum	Uur	414
	Plactomyces/Chlamydia /Verrucommicrobium group	*Chlamydia pneumoniae*	Cpn	667
		Chlamydia trachomatis	Ctr	649
	Proteobacteria	*Buchnera sp. APS*	Buc	583
		Campylobacter jejuni	Cje	1,344
		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
		Helicobacter pylori J99	jHp	1,114
		Mesorhizobium loti	Mlo	5,390
		Neisseria meningitidis MC58	Nme	1,555
		Neisseria meningitidis Z2491	NmA	1,540
		Pasteurella multocida	Pmu	1,838
		Pseudomonas aeruginosa	Pae	4,698
		Rickettsia prowazekii	Rpr	723
		Vibrio cholerae	Vch	2,998
		Xylella fastidiosa	Xfa	1,687
	Spirochaetales	*Borrelia burgdorferi*	Bbu	714
		Treponema pallidum	Tpa	735
	Thermotogales	*Thermotoga maritima*	Tma	1,577
	Thermus/Deinococcus group	*Deinococcus radiodurans*	Dra	2,332
Eukaryota	Fungi	*Saccharomyces cerevisiae*	Sce	2,450

단계 D에서 정렬된 자료 중 43종의 미생물 모두에서 존재하는 보존적 ortholog들만 추출하였다 (Fig. 1-E). 분석 단계 C에서 분기한 또 다른 작업경로로서 distance value를 구하기 위하여 CLUSTALW 프로그램을 이용하여 다중서열검색을 실시하여 (Fig. 1-F) 각 유전자들에 관한 distance value를 담고 있는 '*.dnd 파일과 다중정렬된 염기서열을 담고 있는 *.aln 파일을 생성시켰다. 총 77,069개의 미생물 유전자들이 포함된 3,852개의 ortholog들에 대하여 perl language를 이용하여 일괄 처리(batch process)로 수행하였다. Distance value의 합을 모든 유전자들에 대하여 구하기 위해 '*.dnd' 파일을 읽어 개별 유전자의 distance value의 합을 유전자 이름과 함께 구하였다 (Fig. 1-G). 분석 단계 G에서의 결과를 각 종별로 분류하여 종과 유전자로 구성된 2차원의 배열로 하였다 (Fig. 1-H). 분석단계 E 와 H 각각에서 나오는 출력을 받아 43종 미생물 모두에 보존적인 유전자를 각 distance value에 따라 그리고 종을 기준으로 2차원 배열로 출력하였다 (Fig. 1-I). 분석단계 D부터 I까지의 모든 작업은 perl language를 사용하여 자동화시켰다.

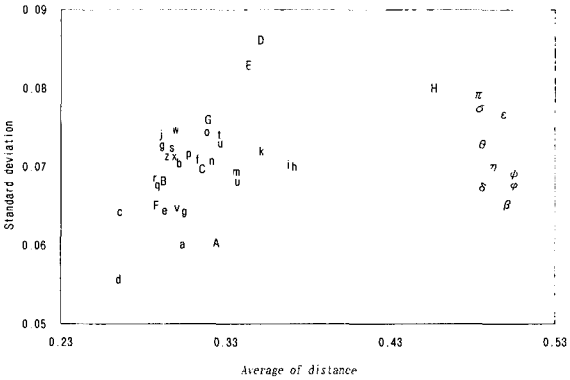
분석 단계 H로부터 나온 결과인 정리된 각 유전자들에 대한 distance value의 합을 SPSS 프로그램을 이용하여 통계처리하여 미생물 종별로 평균과 분산을 구하였다 (Fig. 1-J).

결 과

총 43종의 미생물 유전체에 대한 분석 작업 결과, 72개의 ortholog들이 보존적인 것으로 나타났다 (Table 2). Ortholog 중 단백질 합성에 관여하는 유전자들이 총 52개(72.2%)로 높은 비율을 차지하는 것을 확인하였다. DNA replication과 repair 그리고 recombination에 관여하는 COG가 3개로 4.2%였고 전사(transcription) 관련 COG가 4개(5.6%)로 나타났다. 기타 에너지생성, 세포분열, 핵산과 당의 운반에 관련되는 유전자들이 보존적인 것으로 나타났다. 그리고 정확한 기능을 알 수는 없지만 일반적인 기능의 유추가 가능한 보존적 유전자도 2개였다.

Table 3은 보존적 유전자를 distance value의 합에 따라 정리한 (FIg. 1-I) 결과로서 43종 미생물 모두에 대하여 보존적인 것으로 나타난 총 72개의 유전자들에 대하여 distance value를 모두 합산한 것이다. Distance value가 낮다는 것은 각 종들간의 유전자 염기서열 차이가 작다는 것으로 보존성이 높다는 것을 의미한다. 이때 특정 종내에서의 유전자 복제에 의한 paralog가 존재할 때에는 가장 보존성이 높게 나타난 유전자만을 기준으로 처리했다. 분석 결과 Translation elongation factor인 GTPase의 보존성이 가장 높은 것으로 나타났으며 리보솜단백질 등 해독(translation)에 연관된 유전자들의 보존성이 높은 것을 알 수 있었다. 그 외 전사(transcription)와 신호인식(signal recognition)에 관여하는 일부 유전자도 보존성이 높은 것으로 판명되었다.

분석에 이용된 43종 미생물에 대하여 개별 유전체(genome)를 72개의 보존적 유전자가 대별하는 COG들의 distance value가 갖는 평균과 분산으로 나타낸 결과가 Fig. 2에 나타나있다. 각 미생물이 보유하는 COG들의 distance



Archaea
β :*Aeropyrum pernix*
δ :*Archaeoglobus fulgidus*
ε :*Halobacterium sp. NRC-1*
η :*Methanobacterium thermoautotrophicum*
θ :*Methanococcus jannaschii*
π :*Pyrococcus abyssi*
σ :*Pyrococcus horikoshii*
φ :*Thermoplasma acidophilum*
Ψ :*Thermoplasma volcanium*

Bactreia
a :*Aquifex aeolicus*
b :*Synechocystis*
c :*Bacillus halodurans*
d :*Bacillus subtilis*
e :*Lactococcus lactis*
f :*Mycobacterium leprae*
g :*Mycobacterium tuberculosis*
h :*Mycoplasma genitalium*
i :*Mycoplasma pneumoniae*
j :*Streptococcus pyogenes*
k :*Ureaplasma urealyticum*
l :*Chlamydia pneumoniae*
m :*Chlamydia trachomatis*

Bacteria
n :*Buchnera sp. APS*
o :*Campylobacter jejuni*
p :*Caulobacter crescentus*
q :*Escherichia coli K12*
r :*Escherichia coli O157*
s :*Haemophilus influenzae*
t :*Helicobacter pylori 26695*
u :*Helicobacter pylori J99*
v :*Mesorhizobium loti*
w :*Neisseria meningitidis MC58*
x :*Neisseria meningitidis Z2491*
y :*Pasteurella multocida*
z :*Pseudomonas aeruginosa*
A :*Rickettsia prowazekii*
B :*Vibrio cholerae*
C :*Xylella fastidiosa*
D :*Borrelia burgdorferi*
E :*Treponema pallidum*
F :*Thermotoga maritima*
G :*Deinococcus radioduran*

Eukaryota
H :*Saccharomyces cerevisiae*

Fig. 2. Distribution pattern of microbial genomes by distance value.
X-axis represents average of distance (sum of distance value for each COG was divided by 72) and Y-axis shows the variance of distance value among 72 COGs of one genome. Almost of species were grouped with their phylogenetic relatives. Alphabet in the figure represents archaea(β ~ Ψ), Bacteria(a ~ G), and Eukaryota (H).

Table 2. Conserved genes and their function.

Gene name	Product	Functional category
COG0008	Glutamyl- and glutaminyl-tRNA synthetases	Translation, ribosomal structure and biogenesis
COG0016	Phenylalanyl-tRNA synthetase alpha subunit	
COG0018	Arginyl-tRNA synthetase	
COG0024	Methionine aminopeptidase	
COG0030	Dimethyladenosine transferase (rRNA methylation)	
COG0048	Ribosomal protein S12	
COG0049	Ribosomal protein S7	
COG0051	Ribosomal protein S10	
COG0052	Ribosomal protein S2	
COG0060	Isoleucyl-tRNA synthetase	
COG0072	Phenylalanyl-tRNA synthetase beta subunit	
COG0080	Ribosomal protein L11	
COG0081	Ribosomal protein L1	
COG0087	Ribosomal protein L3	
COG0088	Ribosomal protein L4	
COG0089	Ribosomal protein L23	
COG0090	Ribosomal protein L2	
COG0091	Ribosomal protein L22	
COG0092	Ribosomal protein S3	
COG0093	Ribosomal protein L14	
COG0094	Ribosomal protein L5	
COG0096	Ribosomal protein S8	
COG0097	Ribosomal protein L6	
COG0098	Ribosomal protein S5	
COG0099	Ribosomal protein S13	
COG0100	Ribosomal protein S11	
COG0102	Ribosomal protein L13	
COG0103	Ribosomal protein S9	
COG0124	Histidyl-tRNA synthetase	
COG0143	Methionyl-tRNA synthetase	
COG0162	Tyrosyl-tRNA synthetase	
COG0172	Seryl-tRNA synthetase	
COG0180	Tryptophanyl-tRNA synthetase	
COG0184	Ribosomal protein S15P/S13E	
COG0185	Ribosomal protein S19	
COG0186	Ribosomal protein S17	
COG0197	Ribosomal protein L16/L10E	
COG0198	Ribosomal protein L24	
COG0199	Ribosomal protein S14	
COG0200	Ribosomal protein L15	
COG0231	Translation elongation factor P/translation initiation factor eIF-5A	
COG0244	Ribosomal protein L10	
COG0255	Ribosomal protein L29	

Table 2. (continued)

Gene name	Product	Functional category
COG0050	GTPases - translation elongation factors	Translation, ribosomal structure and biogenesis /Amino acid transport and metabolism
COG0085	DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful)	Transcription
COG0086	DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn)	
COG0202	DNA-directed RNA polymerase alpha subunit/40 kD subunit	
COG0250	Transcription antiterminator	
COG0550	Topoisomerase IA	DNA replication, recombination and repair
COG0258	5'-3' exonuclease (including N-terminal domain of PolI)	
COG0592	DNA polymerase sliding clamp subunit (PCNA homolog)	
COG0037	Predicted ATPase of the PP-loop superfamily implicated in cell cycle control	Cell division and chromosome partitioning
COG0492	Thioredoxin reductase	Posttranslational modification, protein turnover, chaperones
COG0533	Metal-dependent proteases with possible chaperone activity	
COG0526	Thiol-disulfide isomerase and thioredoxins	Posttranslational modification, protein turnover, chaperones/Energy production and conversion
COG0201	Preprotein translocase subunit SecY	Cell motility and secretion
COG0541	Signal recognition particle GTPase	
COG0552	Signal recognition particle GTPase	
COG0636	F0F1-type ATP synthase c subunit/Archaeal/vacuolar-type H+-ATPase subunit K	Energy production and conversion
COG1109	Phosphomannomutase	Carbohydrate transport and metabolism
COG0125	Thymidylate kinase	Nucleotide transport and metabolism
COG0012	Predicted GTPase	General function prediction only
COG0073	EMAP domain	

value는 정규분포를 따르는 것으로 나타났다. Distance value의 평균값이 가장 낮은 미생물은 *Bacillus subtilis*로 (Fig. 2의 d) 0.2565였고 (표준편차 = 0.0056), 가장 높은 종은 archaea인 *Themoplasma volcanium*(Fig. 2의 Ψ)으로 0.4947(표준편차 = 0.0690)로 나타났다. 그리고 *Bacillus halodurans*(Fig. 2의 c)와 *Thermoplasma acidophilum*(Fig. 2의 φ)등 대부분의 종들이 근연종과 그룹화되어 있는 것을 알 수 있었다. 한편 archaea 그룹 모두가 distance value가 비교적 높은 곳에 위치하는 것으로 나타났다. 유일한 1종의 eukaryota(Fig. 2의 H)는 archaea보다 낮은 distance value를 나타냈으나 각 COG 변이의 표준편차는 높게 나타나는 것을 알 수 있었다.

고 찰

보존적 유전자 (conserved gene)

분석한 미생물 유전체들은 크기(size)나 유전자의 개수, ortholog, paralog의 존재 등에서 상호간에 복잡한 양상을 보이는 것을 확인할 수 있었다. Table 2에서 보존적 유전자로 선정된 총 72종의 COG들은 대부분 단백질의 합성에 관여된 것으로, 전체 보존적 유전자 중 단일분류로 52개(72.2%)에 이른다. 이러한 결과는 생명체의 유지와 번식에 있어 단백질이 차지하는 비중이 아주 높은 것을 알 수 있으며 오랜 시간의 진화과정에 있어서도 필수적인 기능으로 유지되어 왔음을 나타내는 것으로 사료된다. 전사(transcription), DNA 복제(replication) 및 복구(repair), 세포분열 조절, 에너지대사 등의 기능들도 보존적인 것으로 나타났지만 전체 COG중 그 비율은 높지 않은 것을 알 수 있었다. 생명의 가장 큰 특성을 물질대사(metabolism)와 자기복제(reproduction)로 보았을 때, 이러한 두 특성을 유지하기 위한 유전자들이 미생물들 간에 보존적인 것으로 판단할 수 있었고 특히 물질대사에 관련된 유전자들의 보존정도가 아주 높은 것으로 나타나 초기생명체의 주된 활동은 물질대사와 주로 관련되었을 것으로 추측해볼 수 있었다. 보존적 유전자 중 핵산의 운반과 생성에 관여하는 COG의 비율이 아주 낮은 것(Table 2)도 이를 뒷받침하는 결과라고 할 수 있을 것이다. 따라서 초기생명체의 유전체에서 물질대사에 관여하는 유전자들의 비중이 높았을 것이라고 유추할 수 있었다. Mushegian[16]은 *Haemophilus influenzae* 와 *Mycoplasma genitalium*을 이용한 연구에서 최소기능유전자(minimal function gene) 조합을 얻기 위해 두 종에 공통으로 존재하는 유전자들을 256개 추출하였다. 그는 최소기능유전자로 선정된 유전자 중에서 약 37.1% (95/256 유전자)가 단백질의 합성에 직접적으로 관여하는 것으로 파악하였다.

Snel 등은 21종의 미생물들을 이용한 최소기능유전자조합의 연구를 수행하였다 [15, 재인용]. 이들의 연구 결과에서 다수의 종을 동시에 고려하면 최소기능유전자조합의 연구에 오히려 역효과가 생길 수도 있다는 사실이 알려지게 되었다. 즉 유전자의 기능대체현상(gene displacement)이 있는 경우 분석에 어려움이 따른다는 것을 유추하였다. 분석대상 종들의 수를 늘리면, 특정 유전자의 기능이 다른 유전자에 의해 대체된 필수 요소가 탈락하는 경우가 자주 발생할 수 있다는 것이었다. 그러므로 여러 종들을 동시에 고려할 때 기능이 대체된 유전자들을 고려하지 않는다면, 결과는 단지 순수한 염기서열상의 보존적 유전자를 의미하는 것이 된다고 보고하였다 [15]. 이는 기능은 유사하지만 서열이 다른 유전자(non-orthologous gene)들은 단순한 염기서열로만 비교할 때 탈락될 가능성이 높다는 것을 의미한다. 이러한 이유로 본 연구결과와 다른 연구자들의 결과를 비교한 Table 4를 보면 Snel은 본 연구보다 적은 21개의 유전체를 분석하고도 60개의 보존적 유전자를 검출한 것으로 나타났다. 이러한 결과는 COG와는 다른 Snel의 ortholog 선별방식에 의한 non-ortholog들의 탈락에 따른 결과라고 유추할 수 있었다. Snel의 연구에서도 단백질의 합성(기능 분류 중 "Translation, ribosomal structure 그리고 biogenesis에 해당)에 관여하는 유전자는 보존적 유전자 60개중 44개로 (73.3%) 본 연구의 결과(72.2%)와 비슷한 것을 알 수 있었고 리보솜단백질에 관계된 유전자들의 보존정도가 가장 높은 것으로 나타났다[15].

연구대상 미생물 모두에 존재하는 각 ortholog에 대하여 보존성의 정도를 가름하는 척도로 이용된 distance value의 합에 따라 정렬하였다 (Table 3). 가장 수치가 낮은 것은 단백질 합성의 조절에 관여하는 GTPase-translation elogation factor (COG0050)로 distance value의 합이 9.83905 였고, 가장 큰 수치를 가진 것은 ATPase로 작용할 것으로 판단되어지는 Predicted ATPase of the PP-loop superfamily implicated in cell 32cycle control (COG0037)로 distance value의 합은 18.6566였다. 즉 GTPase-translation elongation factors (COG0050)는 43종의 미생물에서 가장 강하게 보존되고 있는 유전자라고 할 수 있었다. Figure 3은 이렇게 구분되는 두 개(COG0050, COG0037)의 단백질 그룹을 이용한 phylogenic tree를 나타내고 있다. 보존성이 높은 COG0050의 경우 archaea들은 높은 유사성을 보이며 내부적으로 응집하면서 다른 미생물들과 확실히 구별되는 그룹을 형성하고 있었다. Eucaryote인 *S. cerevisiae*가 eubacteria 그룹에 속하는 결과가 나온 것은 horizontal gene transfer에 의한 현상이라고 추측할 수 있었다. 즉 이는 evolution의 부정확한 bifurcating과 일치하는 것이었다. Cell cycle에 관여하는 COG0037의 경우는 *Neisseria* 속이 다른 eubacteria와 높은 유연관계를 보이지 않고 오히려 archaea와 eucaryote에 더 유연관계가 높은 것으로 나타났다.

Table 3. Ranking of conserved genes by conservation score (Sum of distance value in each genome). The lesser value of conservation score represents that their orthologs were more conserved among microorganisms studied.

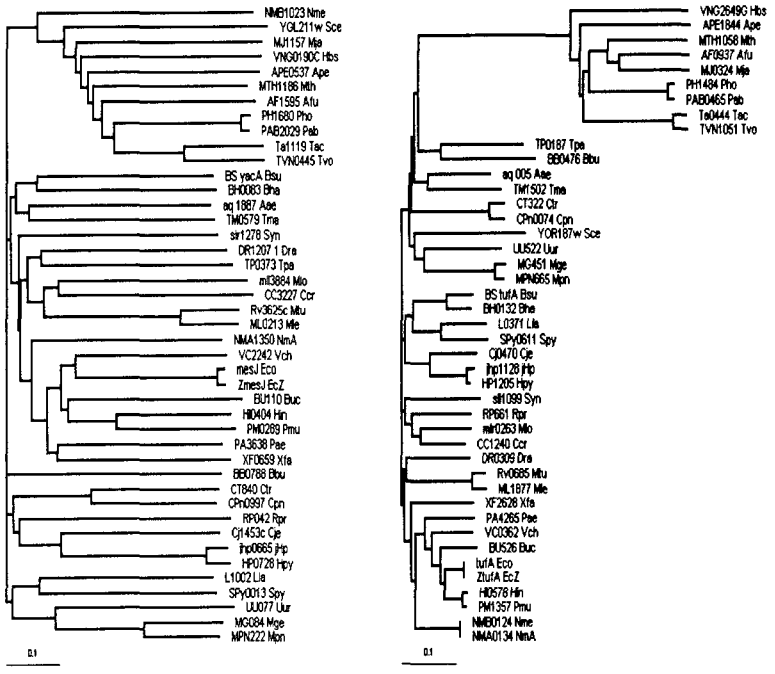
Gene name	conservation score	Product
COG0050	9.83905	GTPases - translation elongation factors
COG0093	11.04697	Ribosomal protein L14
COG0048	11.05022	Ribosomal protein S12
COG0185	11.91402	Ribosomal protein S19
COG0100	12.16104	Ribosomal protein S11
COG0080	12.16935	Ribosomal protein L11
COG0480	12.24023	Translation elongation and release factors (GTPases)
COG0051	12.27034	Ribosomal protein S10
COG0099	12.47138	Ribosomal protein S13
COG0090	12.81681	Ribosomal protein L2
COG0094	12.82743	Ribosomal protein L5
COG0049	12.96233	Ribosomal protein S7
COG0361	13.33831	Translation initiation factor IF-1
COG0103	13.35715	Ribosomal protein S9
COG0186	13.36017	Ribosomal protein S17
COG0012	13.38658	Predicted GTPase
COG0541	13.60601	Signal recognition particle GTPase
COG0102	13.84370	Ribosomal protein L13
COG0197	13.89479	Ribosomal protein L16/L10E
COG0081	13.96644	Ribosomal protein L1
COG0086	14.03151	DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn)
COG0184	14.05673	Ribosomal protein S15P/S13E
COG0092	14.09749	Ribosomal protein S3
COG0085	14.11366	DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful)
COG0552	14.16934	Signal recognition particle GTPase

미생물 종에 대한 유전체 수준의 분석

미생물 유전체에 대한 유전자 수준의 분석법은 많이 나와 있으나 유전체 전체를 통괄하여 분석하는 방법은 그다지 많지 않다. 염기서열 비교에서 대상의 길이를 제한하는 방식에 따라 global alignment와 local alignment로 나누고, 비교하는 서열의 수에 따라 단순서열비교와 다중서열비교로 나눈다. 그러나 이러한 분석방식들은 유전자 수준에서의 비교라는 한계점을 가진다. Global alignment는 개발 초기에는 유전체 전체나 분석할 대상 전체에 대한 통괄적인 시각을 얻기 위해 설계되었으나, 유전자 단위의 보존성이 전체 유전체의 보존성보다 높다는 사실이 알려지면서 local alignment만이 현재는 주로 사용되고 있다 (FASTA 혹은 BLAST) [1,18]. 최근에 미생물 유전체에 대한 통괄적 시각을 얻기 위해 유전체 상호간의 consensus gene mapping에 의한 분석법이 많이 도입되고 있다. 미생물 종에 따라 모든 유전자 단위로 구해진 distance value들은 개별 요소적 관점에서는 그 유전체의 특성을 대변한다고 말할 수 없다. 하지만 유전자 표본의 수가 많아질수록 그 유전체의 특성을

Table 4 : Comparison of minimal function genes with various approach.

Conserved gene sets deduced by various approaches.			
Numbers of genome	**Estimated number of genes/proteins**	**Probable cause of underestimation or overestimation**	**Year/Ref.**
2 (*Haemophilus/ Mycoplasma*)	256	Extensive elimination of paralogs in small genome of *Mycoplasma* parastic lifestyle resulting in multiple auxotrophies in both species; unknown number of gene displacements in addition to several detected cases. (Bacteria-specific solutions for some functions)	1996/[13]
21	60	No consideration of non-ortholog displacement	1999/[14]
7	~320	Superkingdom-specific pathways, such as DNA replication, repair, lipid biosynthesis.(Unknown number of parallel solutions for the same biochemical function)	1997/[20]
1 (*B.subtilis*)	300-560	Extensive paralogy in large bacterial genome	1995/[8]
1 (*S.cerevisiae*)	~1000	Large number of essential genes involved in eukaryote-specific housekeeping	1999/[22]
43	72	Not sufficient consideration of non- orthologous gene displacement	2002/ This study



(A) COG0037 (B) COG0050

Fig. 3. Phylogenetic tree of 43 microorganisms studied in the respect of COG0037 (A) and COG0050 (B). Gene name, ahead of each abbreviation of microorganism, represents the gene name in that microorganism.

가깝게 표현한다고는 말할 수 있을 것이다. 하지만 서로 공유하지 못하는 유전자까지 추가하여 분석하는 것은 아직 널리 받아들여지지 않으므로 본 연구에서는 분석대상 미생물 모두가 보유하는 보존적 유전자 72개가 나타내는 단백질에 대한 distance value의 합을 이용하여 distance value 자료의 통계치와 한 미생물 유전체내에서 각 유전자의 distance value의 평균과 분산에 따른 분포(Fig. 2)를 분석하였다. Distance value는 정규화된 유전자의 보존 정도로서 유전체 전체 수준에서의 보존성의 정도를 나타내며, 특정 미생물의 distance value의 분산은 그 미생물이 가진 일부 유전자들의 유전자 변이(진화)등을 설명하는 지표로 사용될 수 있을 것이다 [12]. Figure 2에서 각 미생물 종들은 유연관계가 가까운 종끼리 그룹화되어 있는 것을 알 수 있었다. 총 9종의 고세균(archaeabacteria)이 distance value가 높은 하나의 그룹을 형성하는 것으로 나타났고 진핵미생물인 *S. cerevisiae* 가 archaea 그룹과 인접한 곳에 분포하는 것으로 나타났다. 진정세균(eubacteria)의 경우는 distance value는 상대적으로 낮은 위치에 archaebacteria와는 다른 그룹을 형성하는 것을 볼 수 있었다. 또한 *Escherichia coli*를 비롯한 대다수의 유전체들이 기존의 미생물 종 분류상의 근연종과 유사하게 그룹화되어 본 연구에서 사용한 분석알고리즘의 결과가 기존의 분류결과와 크게 다르지 않음을 확인할 수 있었다.

Tatusov[22]는 COGs를 각 단백질군(protein family)에 대한 진화적 관점의 분석을 위한 편리한 체계라고 언급하면서 COGs의 응용가능성을 미지의 단백질의 기능 추측, 한 유전체에서 검출하지 못한 보존적 그룹에 대한 체계적 연구, 광범위 항생제의 연구 가능성 등이라고 하였다. 본 연구에서 이용된 보존적 유전자를 찾아내는 접근방법 등은 다양한 응용 분야를 가지며, 전체 유전체 혹은 일부의 염기서열이 밝혀진 어떠한 종에도 적용이 가능할 것으로 사료된다.

요 약

미생물 유전체(genome)들 사이의 보존된 유전자 (conserved gene)를 밝히는 것은 생명의 본질을 이해하는데 있어 다양한 의미를 갖는다고 할 수 있을 것이다. 본 연구에서는 보존적 유전자를 찾아내고, distance value를 이용하여 구한 보존성의 정도 C(conservation score)를 이용하여 종간의 유전자 변이의 정도를 단백질 관점에서 분석하였다. 분석에 사용된 자료는 COGs 데이터베이스의 총 43종의 미생물 유전체들이며, 이들은 총 77,069개의 유전자들을 포함하는 3,852 개의 ortholog들로 구성되어있었다. 분석 결과 43종의 미생물 유전체에 대하여 총 72개의 유전자들이 보존적인 것으로 나타났으며, 이들 중 72.2%인 52종의 유전자가 단백질 합성에 관련되는 것으로 나타났다. 이들 보존적 유전자들에 대하여 보존성의 정도 C를 계산하여 보존성의 순위를 얻었으며, 가장 잘 보존된 유전자는 GTPase-translation elogation factor (COG0050)로 나타났다. 그리고 72개의 보존적 유전자가 나타내는 COG 모두를 이용한 분석결과 고세균(archaea)과 진정세균(bacteria)이 각각 독자적인 그룹을 형성하는 것을 관찰하였다. 본 연구의 결과에서 도출한 72개의 보존적 유전자는 생명체의 본질적 기능에 중요한 역할을 담당하는 것으로 사료되었고, 생명체의 진화과정에서 이 유전자들이 보존된 이유와 기능적 연계에 대한 생물학적 연구에 기초 자료를 제공할 것으로 판단되어진다.

감사의 글

본 연구는 2001년도 과학기술부 '지역기술개발용역사업'에 의해 수행되었으며, 이에 감사드립니다.

참 고 문 헌

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST, a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402
2. Choi, J. H.,Jung, H. Y., Kim, H. S. and Cho, H. G. 2000. PhyloDraw,A phylogenetic tree drawing system. *Bioinformatics* **16**, 1056-1058
3. Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G. 1998. Biological sequence analysis (Probabilistic models of proteins and nucleic acids. *Cambridge University Press* 134-159
4. Fraser, C. M., Eisen, J. A. and Salzberg, S. L. 2000.

Microbial genome sequencing. *Nature* **406**, 799-803

5. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. M., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, O. M., Phillips, T. A., Merrick, J. M., Tomb, J.-F., Dougherty, D. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. M., Smith, H. O., Hutchison III, C. A. and Venter. J. C. 1995. The minimal gene complement of *mycoplasma genitalium*. *Science* **270**, 397-403
6. ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria
7. Henikoff, S., Greene, E. A., Pietrokovski, Bork, S., Attwood, T. K. and Hood, L. 1997. Gene Families: The Taxonomy of Protein Paralogs and Chimeras. Science **278**, 609-614
8. http://www.ncbi.nlm.nih.gov/COG/
9. Huynen, M. A. and Bork, P. 1998. Measuring genome evolution. *Proc. Natl. Acad. Sci. USA.* **95**, 5849-5856
10. Itaya, M. 1995. An estimation of minimal genome size required for life. *FEBS Lett.* **362**, 257-260
11. Judson, N. and Mekalanos, J. J. 2000. TnAraOut, A Transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.* **18**, 740-745
12. Kimura,M. 1983. The neutral theory of molecular evolution. *Cambridge University Press*
13. Koonin, E. V. and Mushegian. A. 1996. Complete genome sequences of cellular life forms,glimpses of theoretical evolutionary genomics. *Curr. Opin. Genet. Dev.* **6**, 757-762
14. Lehoux, D. E., Sanschagrin, F. and Levesque, R. C. 2001. Discovering essential and infection-related genes. *Curr. Opin. Microbiol.* **4**, 515-519
15. Mushegian, A. 1999. The minimal genome concept. *Curr. Opin. Genet.* **9**, 709-714
16. Mushegian, A. and Koonin, E. V. 1996. A minimal gene set for cellular life derived by comparision of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA.* **93**, 10268-10273
17. Needleman, S. B. and Wunsch, C. D. A. 1970. General method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453
18. Pearson, W. R. and Lipman, D. J. 1988. Improved Tools for Biological Sequence Analysis. *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448
19. Reich, K. A. 2000. The search for essential genes. *Res. Microbiol.* **151**, 319-320
20. Saitou, N. and Nei, M. 1987. The neighbor-joining method,a new method for reconstructing method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406-425
21. Smith, T. F. and Waterman, M. S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197
22. Tatusov, R. L., Galperin, M. Y., Natale, D. A. and Koonin, E. V. 2000. The COG database,a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33-36
23. Tatusov, R. L., Koonin, E. V. and Lipman, D. L. 1997. A genomic perspective on protein families. *Science* **278**, 631-637
24. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Lian, H., Anderson, K., Andre, B., Bangham, R.,Benito, R., Boeke, J. D, Bussey, H. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901-906

(Received September 16, 2002; Accepted October 15, 2002)