한국생물공학회지 제21권 제2호
Korean J. Biotechnol. Bioeng.
Vol. 21, No. 2, 144-150(2006)

메타게놈 유래 미규명 유전자의 발현에 관련된 특성분석

[1]박 승 혜 · [1,2]정 영 수 · [1]김 원 호 · [3]김 근 중 · † [1,2]허 병 기
[1]인하대학교 공과대학 생물공학과, [2]인하대학교 생물산업기술연구소, [3]전남대학교 자연과학대학 생물학과
(접수 : 2006. 2. 15., 게재승인 : 2006. 4. 14.)

Structural Characteristics of Expression Module of Unidentified Genes from Metagenome

Seung-Hye Park[1], Young-Su Jeong[1,2], Won-Ho Kim[1], Geun-Joong Kim[3], and Byung-Ki Hur† [1,2]
[1]Department of Biological Engineering, Inha University, Incheon 402-751, Korea
[2]Institute of Biotechnological Industry, Inha University. Incheon 402-751, Korea
[3]Department of Biological Sciences, Chonnam National University, Kwangju 500-757, Korea
(Received : 2006. 2. 15., Accepted : 2006. 4. 14.)

The exploitation of metagenome, the access to the natural extant of enormous potential resources, is the way for elucidating the functions of organism in environmental communities, for genomic analyses of uncultured microorganism, and also for the recovery of entirely novel natural products from microbial communities. The major breakthrough in metagenomics is opened by the construction of libraries with total DNAs directly isolated from environmental samples and screening of these libraries by activity and sequence-based approaches. Screening with activity-based approach is presumed as a plausible route for finding new catabolic genes under designed conditions without any prior sequence information. The main limitation of these approaches, however, is the very low positive hits in a single round of screening because transcription, translation and appropriate folding are not always possible in *E. coli*, a typical surrogate host. Thus, to obtain information about these obstacles, we studied the genetic organization of individual URF's (unidentified open reading frame from metagenome sequenced and deposited in GenBank), especially on the expression factors such as codon usage, promoter region and ribosome binding site (rbs), based on DNA sequence analyses using bioinformatics tools. And then we also investigated the above-mentioned properties for 4100 ORFs (Open Reading Frames) of *E. coli* K-12 generally used as a host cell for the screening of noble genes from metagenome. Finally, we analyzed the differences between the properties of URFs of metagenome and ORFs of *E. coli*. Information derived from these comparative metagenomic analyses can provide some specific features or environmental blueprint available to screen a novel biocatalyst efficiently.

Key Words : Metagenome, ORF, URF, expression factor, screening

서 론

생태계의 주된 구성요소의 하나로써 미생물은 토양, 하천, 해수, 동물 장내 등의 서식지를 비롯하여 심해저의 열수구, 고온의 온천수, 혹한의 남극 혹은 고농도의 염분이나 강산이나 강알칼리성 환경에 이르기까지 지구상의 어떤 환경에서도 존재한다(1). 이렇듯 미생물은 자연계 곳곳에 존재하면서 그 종과 기능의 다양성을 바탕으로 지구 생태계에서 중요한 역할을 담당한다. 따라서 자연계에 존재하는 미생물의 역할을 정확히 이해하는 일은 우리가 살고 있는 생태계를 이해하는 근간이라고 할 수 있다. 다양한 환경에 대한 미생물의 적응력은 생물산업의 핵심소재로서 이용될 수 있는 많은 가능성을 제시하지만, 배양을 통한 균주의 선별에 따른 한계로 인해 대다수의 미생물은 발견조차 되지 못한 채로 연구 대상에서 제외되고 있는 실정이다.

생태 환경에 존재하는 미생물의 배양가능성에 대한 연구에 의하면, 자연적으로 미생물이 선별-농축된 활성오니를 제외한 대부분의 토양, 담수, 해수 등의 환경에서는 1% 미만의 극히

† Corresponding Author : Department of Biological and Chemical Engineering, College of Engineering, Inha University, Incheon 402-751, Korea
Tel : +82-32-860-7512, Fax : +82-32-872-4046
E-mail : biosys@inha.ac.kr

제한된 종만이 배양 가능한 것으로 예측되고 있다(2). 생태계 미생물의 99% 이상이 현재의 기법으로는 순수하게 배양될 수 없다는 사실이 알려지면서 전통적인 선별법의 한계를 인식하게 되었다. 따라서 기존의 분리·배양 기법을 극복하고 미지의 미생물 자원을 이용하기 위한 방법으로 메타게놈(metagenom)이라는 환경유전체 연구분야에 관심을 갖게 되었다(3). 메타게놈은 "특정 환경에 존재하는 모든 미생물의 유전체 집합"으로 정의되나 최근에는 환경시료로부터 추출한 유전체 또는 유전자를 포함하는 클론을 총칭하기도 하며, 이와 관련된 일련의 연구를 메타게노믹스 (metagenomics)라고 한다(4).

메타게놈 관련연구는 99% 이상의 난배양성 혹은 비배양성 미생물을 활용하여 새로운 생물자원을 탐색하는 분야이고 다양한 환경에서 목적에 적합한 소단위 생태계를 선정하는 것으로 시작한다. 이 후 환경시료를 채취한 후, 메타게놈을 직접 추출하여 벡터에 클로닝하는 방법으로 미생물 유전자원을 확보한다(5). 필요에 따라서는 enrichment 기법을 사용하여 연구대상을 일정범위내로 한정시킴으로써, 유전자원을 비교적 용이하고 효과적으로 탐색하는데 이용한다(6).

메타게놈 연구는 일반적으로 유전자 조작에 많이 사용되는 플라스미드 벡터 혹은 보다 큰 크기의 유전자를 삽입할 수 있는 BAC, Cosmid 혹은 Fosmid 등의 벡터를 이용하여 다양한 라이브러리를 구축하고, 형질발현 혹은 유전자 상동성에 기초한 선별법을 적용하여 원하는 특성을 지닌 유전자를 선별하는 과정으로 진행된다(7). 이와 관련한 연구는 비교적 길지 않은 역사에도 불구하고, 4-hydroxybutyrate dehydrogenase(8), lipase/esterase(9), protease(10), amylase, DNase(11), polyketide synthase(12), agarase(13) 등의 효소 자원을 탐색하는 가시적인 연구 성과들을 거두었다.

메타게놈 라이브러리에 적용되는 선별법 중, 순수배양이 가능한 기존 미생물 유래의 유전정보를 이용하여 신규 유전체를 탐색하는 기법은 유전체의 다양성을 확보하기 어렵다는 문제점을 지닌 반면, 발현되는 활성에 근거한 선별법은 유전자 서열의 사전 정보가 없을지라도 다양한 유전체를 확보한 후 원하는 활성에 기초하여 신규유전자를 선별할 수 있다는 장점이 있다(14). 하지만 형질발현 근간의 기법은 cloning host, vector 및 발현시스템 (promoter, regulatory element 및 protein folding)에 따라 발현된 활성정도가 다르기 때문에 라이브러리 내에 많은 신규 유전체를 확보하고도, 유용유전자의 상당 부분이 선별되지 않는다는 심각한 문제점을 드러내고 있다(15).

메타게놈 라이브러리 구축에 일반적으로 이용되는 *E. coli*는 다른 숙주에 비해 상대적으로 적용하기 쉬운 발현계로 알려져 있음에도 불구하고, 메타게놈 자체가 수 만종의 미생물에서 유래된 유전자라는 특성 때문에, *E. coli* 발현계와 부합하지 않는 시스템을 가질 수 있어 원하는 clone의 확보에 어려움이 있을 것이라고 예측할 수 있다. 따라서 유전자의 발현에 근거한 선별법에 있어 언급한 한계점을 개선하기 위해서는, 메타게놈 유전자가 *E. coli* 숙주 세포내에서 발현할 가능성이 있는지 여부를 판단할 수 있는 근거를 마련해야 한다. 이를 위하여 메타게놈 유전자 자체의 특성과 *E. coli* 내에서 정상적으로 발현되는 유전자의 특성을 비교·분석할 필요성이 있다.

상기된 분석을 위해서는, Open Reading Frames (ORFs) 자체의 구조, 즉 유전자 혹은 오페론의 크기, 아미노산 조성 뿐만 아니라 유전자 발현의 조절부위로써 전사 및 번역에 관여하는 promoter region 과 ribosome binding site (RBS)의 특성을 분석해야 한다. 따라서 *E. coli* 를 숙주로 하여 메타게놈 유전자의 형질을 발현시킬 때 반드시 갖추어야 할 유전자의 최소 구조를 'Gene Expression Module'로 규정하고, 메타게놈 유래의 미규명 'Gene Expression Module'과 *E. coli* 내에서 정상적으로 발현되는 4100여 개의 ORF module을 생물정보학 기법을 이용하여 비교·분석하였다.

재료 및 방법

Metagenome DNA 및 단백질 서열정보 수집

NCBI (http://www.ncbi.nlm.nih.gov)의 GenBank 데이터베이스 중 환경시료로부터 분리된 유전체 중, 메타게놈 유래의 단백질로 추정되는 500 여개의 URFs(undefined open reading frames) 서열을 수집하였다(Fig 1). 수집된 서열정보는 BAC, cosmid, fosmid 벡터를 사용하여 구축된 메타게놈 라이브러리로부터 획득한 정보로 비교적 큰 DNA 들이다. 총 410,000 bp 길이의 염기서열에서 유전자 구성의 염기서열과 단백질 구성의 아미노산 서열을 각각 FASTA 형식으로 변환한 후 분석에 이용하였다.

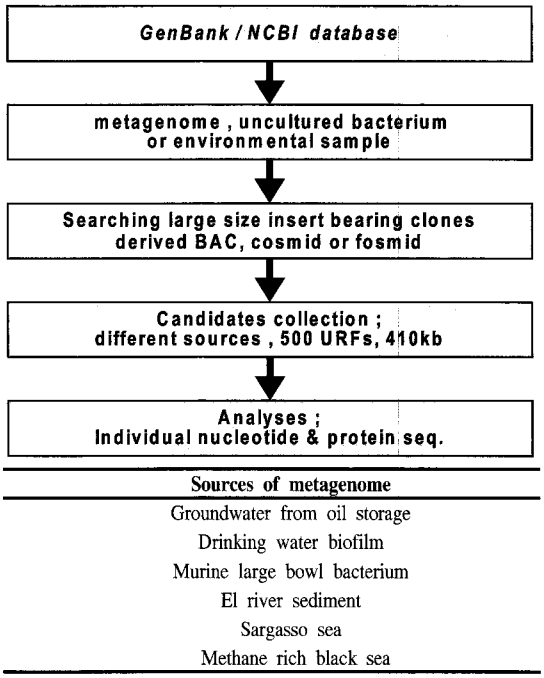

Figure 1. Flow chart of URFs data sets analyzed and sources of metagenome.

단백질 발현과 관련한 분석항목의 결정

일반적으로 유전자의 정보가 단백질로 발현되기까지는 Fig. 2의 과정을 거치게 된다(22). 메타게놈이 *E. coli* 숙주 내에서 발현되기 위해서는 숙주세포에 존재하는 발현계에 적합한 특성을 가지고 있어야 한다. 따라서 메타게놈과 *E. coli*의 유전자 특성, 발현될 단백질의 특성, 그리고 전사와 번역에 관련된 요

소들의 특성을 비교 분석의 주요 항목으로 선정하였다. 유전자 특성으로는 G+C 함량 및 codon usage를, 단백질의 특성으로는 단백질의 길이와 분자량, 등전위점의 분포 및 아미노산의 조성을, 전사과정에 관련된 요소로는 프로모터 영역의 염기서열을, 번역에 관련된 요소로는 리보솜 결합 부위의 염기서열을 고려하였다.

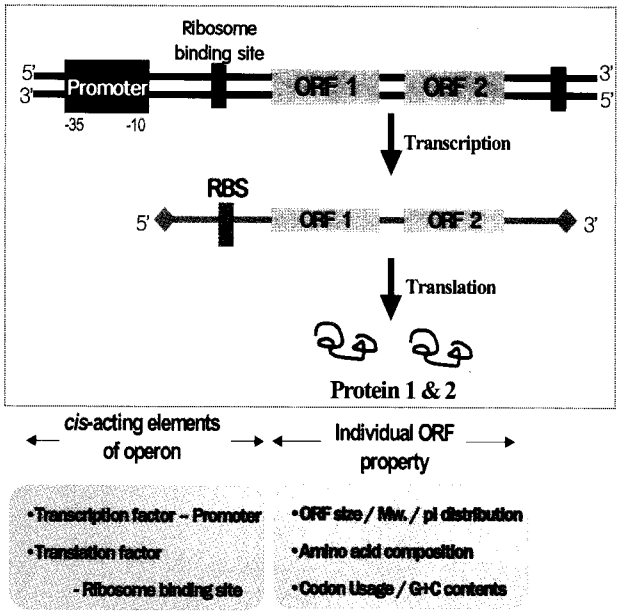

Figure 2. The minimal set of gene expression module in *E. coli* and expression factors analyzed.

미규명 유전자의 특성 분석

Sequence Manipulation Suite (SMS, http://bioinformatics.org/sms/index)는 염기 서열과 아미노산 서열에 관련된 다양한 분석기법이 활용가능한 웹 기반의 생물정보학 프로그램이다(23). 서열 분석을 위한 기초 작업과 간단한 수치 계산에 이 프로그램을 활용하였으며 DNA Stats 프로그램을 이용하여 G+C %를 계산하였다. GenBank에서 수집한 메타게놈 유래 유전자에서 추정된 단백질의 길이와 조성, 분자량 및 이론적 pI 값의 분포 특성은 ExPASy (Expert Protein Analysis System)의 ProtParam 도구 (http://kr.expasy.org/tools/protparam.html]를 이용하여 계산하였다(24).

코돈사용 (Codon usage)에 대한 분석은 단백질을 구성하고 있는 아미노산에 부합되는 표준 유전자코드 빈도수를 계산하여 각 아미노산에 대한 synonymous codon 별로 상대적인 분포를 나타내었다. 이때 사용한 프로그램은 EvolvingCode이다(25). Table 1에 나타낸 간단한 수식을 이용하여 상대적인 수치인 Relative Synonymous Codon Usage (*RSCU*) 값과 ω value를 계산하였고, 이를 *E. coli* 의codon usage와 비교하였다.

전사 및 번역에 관련한 유전서열의 분석

단백질 합성 조절부위의 비교를 위해 *E. coli*의 일반적인 promoter 영역과 리보솜 접합점 (RBS)의 보존서열을 기준으로 메타게놈 내에 존재하는 조절부위의 패턴을 분석하였다. 메타게놈 서열내의 프로모터 영역 예측을 위해 Neural Network Promoter Prediction (www.fruitfly.org/seq_tools/promoter) 프로그램을 활용하였다(26). 원핵세포에 국한시키고 score cutoff를 0.61 이상으로 조정하여 메타게놈 서열상의 프로모터를 예측하였다. 단백질의 발현에 영향을 미치는 보존서열로써 RBS 부위의 분석을 위해 WebLogo(A Sequence Logo Generator, http://weblogo.berkeley.edu) 프로그램을 이용하였다(27). 이를 활용하여 메타게놈 프로모터의 보존양상과 RBS의 다중배열 결과를 나타내었고 *E. coli* 프로모터 및 RBS의 보존서열과 비교하였다.

Table 1. The example of RSCU and *w* calculation. Codon usage frequencies and associated metrics for genes of *E. coli*.

	Codon	Codon_Num	RSCU	*w*
Phe	UUU	78743	1.1636839	1
	UUC	56591	0.8363161	0.7186798
Leu	UUA	51320	0.8531943	0.3038754
	UUG	45581	0.760448	0.2698937
	CUU	42704	0.7124407	0.2520505
	CUC	35873	0.5984851	0.2124108
	CUA	15275	0.2548396	0.0904462
	CUG	168885	2.8175632	1

$RSCU_{Phe\text{-}UUU} = 2 \times Codon\_Num_{Phe\text{-}UUU} / (Codon\_Num_{Phe\text{-}UUU} + Codon\_Num_{Phe\text{-}UUC})$

$w_{Phe\text{-}UUU} = Codon\_Num_{Phe\text{-}UUU} / Max(Codon\_Num_{Phe\text{-}UUU}, Codon\_Num_{Phe\text{-}UUC})$

E. coli 유래 ORF의 특성 분석

메타게놈라이브러리 구축에 전형적으로 사용되는 *E. coli*의 ORF 와 관련한 발현인자와 메타게놈 URF의 특성을 비교하기 위해, *E. coli*에서 발현되는 4100여 종의 ORF 와 발현조절 서열 특성을 분석하였다. 분석항목은 ORF에 의하여 합성되는 단백질의 크기와 분자량, 등전위점의 분포, 아미노산 조성, ORF 의 G+C 함량 및 코돈사용이다. 또한 유전자의 전사와 번역에 관여하는 350개의 프로모터의 염기서열과 리보솜 결합부위의 염기서열도 비교항목으로 선정하였다. 기본적으로는 NCBI Genome Category와 Genome Profile Database (GPDB, http://gpdb.life.nthu.edu.tw/GPDB/)(28)에 구축된 정보를 이용하였고, 분석항목에 따라서는 메타게놈 분석에서 언급한 SMS, Evolving code 및 WebLogo 프로그램을 활용하여 분석하였다.

결과 및 고찰

메타게놈 유래 단백질의 분자량 및 pI 분포 경향

우선 메타게놈 유래의 URF로부터 합성되는 단백질의 아미노산 잔기수를 대장균 내의 ORF 들의 평균길이와 비교하였다. 유래된 환경에 따라 조금씩 차이를 보이지만 대체적으로 200~400개 가량의 아미노산으로 구성되어 있음을 확인할 수 있었다. 일반적으로 알려진 원핵생물의 평균단백질의 길이와 비슷한 분포로서 메타게놈 URF의 80%가 200~500개 정도의 아미노산으로 구성된 단백질을 합성하는 것으로 조사되어졌다(Fig. 3A). 이러한 결과는 메타게놈 유래의 유전자가 주로 원핵생물에서 유래되었고, 대략 20~60 kDa 분자량 분포를 지니며, 적어도 물리적 측면에서 메타게놈의 URF와 *E. coli* 의 ORF가 서로 유사한 특성을 가지고 있음을 예상할 수 있다. Fig.

3(B)는 단백질 분자량의 분포를 메타게놈과 대장균을 비교한 그림이다. 두 가지 결과에 의하면, 메타게놈의 URF에 의하여 합성되는 단백질의 총괄특성이 *E. coli*의 ORF로부터 합성되는 단백질의 총괄특성과 분자량 및 아미노산 길이 측면에서 유사한 분포를 나타내고 있음을 알 수 있다.

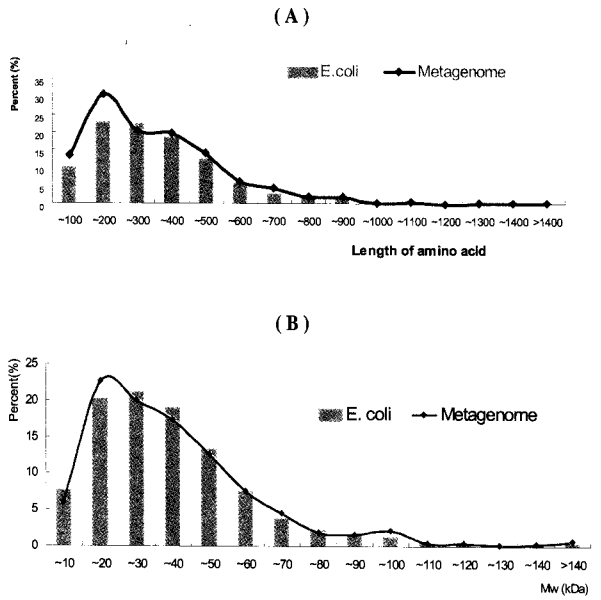

Figure 3. The average distribution of the length (A) and molecular weight (B) of proteins deduced form metagenome genes. The thick line indicate that of metagenome.

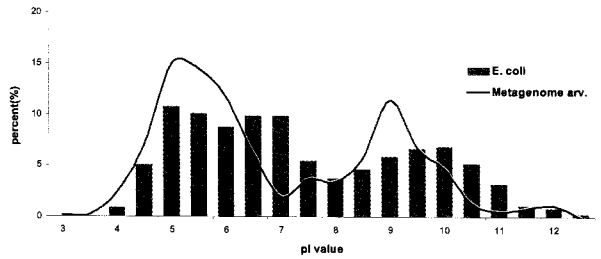

Figure 4. The distribution of theoretical isoelectric point (pI) of proteins deduced from genes of metagenome.

등전위점 (pI)의 이론치는 단백질 활성과 구조에 영향을 미치는 전하에 관련된 항목으로, pI의 분포 경향을 통하여 단백질의 활성과 미생물의 서식 환경에 유리한 pH 조건을 부분적으로 예측할 수 있다. *E. coli* 유래 단백질의 pI 값 분포는 중성 pH를 중심으로 5와10 사이에서 넓게 퍼져있다. 이와는 달리 메타게놈 유래 단백질의 pI 값 분포는 전체적으로 산성 (pH 5)과 알칼리성 (pH 9, pH 12)에서 집중되는 분포 경향을 보여 주었다(Fig. 4). 이는 메타게놈이 유래된 서식환경이 *E. coli*가 정상적으로 생장할 수 있는 조건에 비해 순화되지 않은 환경임을 예상할 수 있으며, 이러한 조건에서 요구되는 아미노산 잔기의 존재 여부를 부분적으로 예측할 수 있다.

아미노산 조성분석

메타게놈 유래의 단백질에 대한 pI 값의 분포가 *E. coli*가 합성하는 단백질에 대한 pI 값의 분포와 뚜렷한 차이를 나타내므로, 단백질을 구성하는 아미노산의 분포 또한 상당한 차이를 나타낼 것이다. Table 2는 *E. coli*와 메타게놈에 의하여 합성되는 단백질의 아미노산 조성비를 상대적으로 나타낸 도표 그림이다. 이 결과에 의하면 메타게놈의 URF에 의하여 합성되는 단백질의 아미노산 조성비는 *E. coli* 유래의 단백질에 대한 아미노산의 조성비와 뚜렷한 차이를 나타내었다. 따라서 메타게놈 유래의 URF를 *E. coli* 내에서 발현시킬 경우에는, 아미노산 조성 차이가 단백질 발현 및 활성에 미치는 영향을 고려하여 숙주세포로 *E. coli*를 선택하는 것이 적합한지를 분석해야 할 필요성이 있음이 제시되었다.

Table 2. Comparison of amino acid composition

Amino acid	*E. coli*	Metagenome	*metagenome* / *Ecoli*
L	10.5	9.82	0.93
A	9.4	9.55	1.02
G	7.2	7.57	1.04
V	7.0	7.17	1.02
I	5.9	6.18	1.05
S	5.7	6.25	1.09
E	5.7	6.42	1.13
Q	5.7	3.25	0.57
R	5.5	5.77	1.05
T	5.3	5.27	0.99
D	5.1	5.48	1.08
P	4.4	4.70	1.07
K	4.3	5.32	1.23
N	3.9	3.68	0.95
F	3.9	3.98	1.03
Y	2.8	2.73	0.97
M	2.7	2.32	0.84
H	2.2	2.08	0.93
W	1.5	1.10	0.73
C	1.2	1.22	1.06

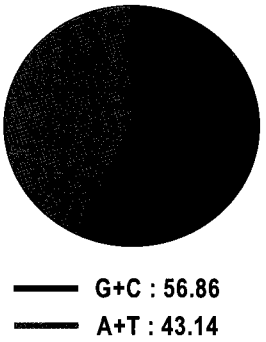

Figure 5. The average G+C content of metagenome.

G+C 함량 분석

미생물의 유전체를 구성하는 DNA에서 3중 수소결합을 이루고 있는 G 와 C 염기는 미생물의 서식 환경에 따라 함량이 조금씩 다른 것으로 알려져 있다. 일반적으로 극한 조건의 환경에서 서식하는 미생물에서 높은 G+C 함량을 보이는 경향성이 있다. *E. coli*의 G+C 함량은 전체의 50.98%로 A+T의 구성비와 큰 차이가 없는 것으로 알려져 있다. 메타게놈 유래 URF의 G+C 함량을 분석한 결과, 세분하면 메타게놈이 유래된 환경별

Table 3. Codon usage(*w*) of predicted genes form metagenome and *E. coli*.

		E. coli	Metagenome			E. coli	Metagenome
Gly	GGU	1.000	0.383	Ser	UCU	0.027	0.49
	GGC	0.650	1		UCC	0.027	0.62
	GGA	0.004	0.368		UCA	0.023	0.658
	GGG	0.013	0.317		UCG	0.489	0.754
Ala	GCU	1.000	0.314		AGU	0.769	0.494
	GCC	0.123	1		AGC	1.000	1
	GCA	0.537	0.405	Threo	ACU	0.036	0.421
	GCG	0.425	0.633		ACC	0.046	1
Val	GUU	1.000	0.411		ACA	0.242	0.536
	GUC	0.067	0.589		ACG	1.000	0.665
	GUA	0.508	0.246	Cys	UGU	0.364	0.6
	GUG	0.237	1		UGC	1.000	1
Leu	UUA	0.014	0.296	Asn	AAU	0.046	1
	UUG	0.015	0.402		AAC	1.000	0.958
	CUU	0.025	0.248	Gln	CAA	0.098	0.852
	CUC	0.034	0.325		CAG	1.000	1
	CUA	0.002	0.138	Asp	GAU	0.424	0.813
	CUG	1.000	1		GAC	1.000	1
Ile	AUU	0.198	0.704	Glu	GAA	1.000	1
	AUC	1.000	1		GAG	0.196	0.699
	AUA	0.002	0.508	Lys	AAA	1.000	1
Met	CUG	1.000	1		AAG	0.212	0.758
Pro	CCU	0.081	0.469	Arg	CGU	1.000	0.301
	CCC	0.003	0.815		CGC	0.406	1
	CCA	0.157	0.563		CGA	0.002	0.329
	CCG	1.000	1		CGG	0.001	0.502
Phe	UUU	0.204	0.834		AGA	0.001	0.538
	UUC	1.000	1		AGG	0.001	0.362
Trypto	UGG	1.000	1	His	CAU	0.244	0.846
Tyr	UAU	1.000	0.996		CAC	1	1
	UAC	0.676	1				

로 다소간의 차이는 보이지만, 전체적으로 G+C 함량이 A+T 함량보다 높은 경향을 보이는 것으로 확인하였다(Fig. 5). 특이할만한 점은 같은 서식환경 유래의 샘플이라 할지라도 이를 좀 더 세분화하여 분석해보면 극명한 차이를 보이는 결과도 확인하였다. 이는 메타게놈의 다양성에 관련이 있는 지표로서 분석에 사용된 유전체가 다양한 균주로부터 유래되었음을 의미하는 것이다.

코돈사용 분석

모든 생물체는 단백질 합성과정에서 mRNA의 염기서열에 따라 고유한 아미노산을 결정한다. 4종류의 염기로 20 가지의 아미노산을 결정하여야 하므로, 3개의 연속된 염기가 하나의 코돈을 구성하며 총 64개의 코돈이 형성된다. 따라서 각각의 아미노산에 대응하는 코돈은 1개 이상이 된다. 이러한 동종코돈(synonymous codon)은 아미노산 합성에 무작위로 사용되는 것이 아니라, 유전체에 따라 서로 다른 편향성 (codon bias)을 나타내며, 유전자의 코돈 활용 패턴은 단백질로 번역되는 수준과 상호 연관성이 있는 것으로 보고되었다(30, 31).

코돈사용은 단백질 합성 시 각 아미노산에 대응하는 동종코돈의 사용빈도수로 나타낸다. 코돈사용의 상대값 (Relative Synonymous Codon Usage, RSCU)인 ω 값을 분석하면 Table 3과 같다. 보다 자세한 분석에 의하면, 동일한 아미노산에 대한 코돈바이어스가 메타게놈과 *E. coli*에서 서로 다른 패턴을 나타내었다. 또한 사용빈도가 가장 높은 코돈과 가장 낮은 코돈도 서식지와 *E. coli*에 따라서 상당한 차이를 나타내었다. 특히 아미노산의 조성비가 크게 다른 단백질을 합성하는 경우, 코돈바이어스 패턴은 물론 사용빈도가 가장 높은 코돈 및 사용빈도가 가장 낮은 코돈 사이에 현격한 차이를 나타내었다. 이러한 결과로부터 메타게놈 유전자가 *E. coli* 숙주 내에서 발현되는 경우 발현 수준에서 상당한 차이를 나타낼 것으로 예측할 수 있었다.

전사 및 번역 인자의 특성 분석

궁극적으로 *E. coli* 내에서 메타게놈 유래의 단백질이 정상적으로 발현되기 위해서는 단백질 발현에 영향을 미치는 조절부위가 *E. coli*와 유사한 패턴을 가지고 있어야 한다. 따라서 메타게놈 염기서열로부터 프로모터 영역을 예측하고 리보솜 결합 부위 (RBS)의 보존서열을 분석한 후 *E. coli*와 비교하였다.

단백질 합성의 조절부위는 대부분의 경우 ORF의 5'말단에서 발견되며, 전사과정에서 RNA 중합효소와 결합하는 조절부위에는 두 개의 중요한 보존서열을 가지고 있다. 전사 개시점을 기준으로 상류방향 쪽의 -10, -35 부위를 프로모터라고 부르며 이 영역에서는 A-T 쌍이 많이 보존되어 있다(Fig. 6A). 메타게놈에서 예측된 프로모터 영역을 WebLogo로 다중배열하여

나타내어보면 Fig. 6B와 같은 패턴의 결과를 얻을 수 있다. TATA box라고 불리우는 -10 영역의 서열은 비교적 보존이 잘 되어있는 반면, -35 영역의 서열은 일정한 보존서열 경향성을 보이지 않는 것으로 보아, 대장균내의 RNA 중합효소가 인식하기 어려운 서열이 많이 존재함을 예측할 수 있었다. 보다 세분화된 분석에서도 같은 경향을 보여주었다.

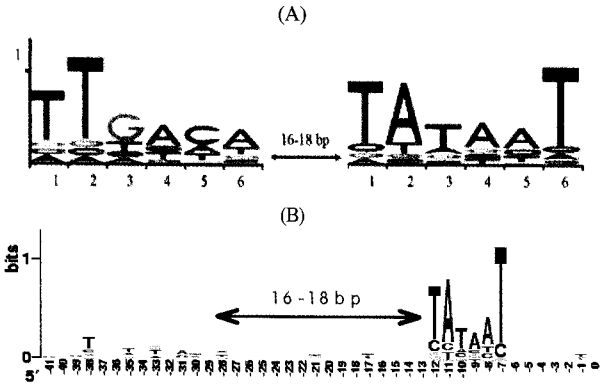

Figure 6. Structural analyses of transcriptional regulatory region of genes from metagenome ((A) In prokaryotes, the DNA sequence just upstream of the transcription start point contains two conserved regions, -35 and -10, recognized by RNA polymerase. (B) The conserved sequences in metagenome).

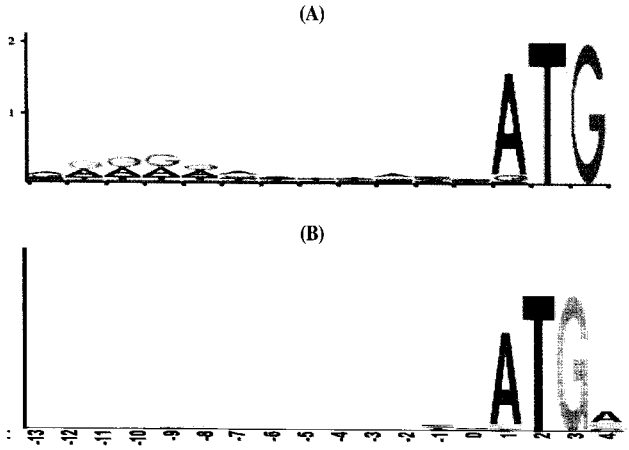

Figure 7. Sequence analyses of ribosome binding region of genes from metagenome ((A) In prokaryotes, the DNA sequence just upstream of the translation start point contains a conserved region referred to Shine-Dalgano sequences. (B) Non-conserved sequence of predicted RBS region of gene from metagenome).

리보솜 결합 부위는 AUG, GUG를 비롯한 시작코돈으로부터 7~9개의 염기서열 앞에 위치하며 시작코돈과의 거리와 염기조성에 따라 mRNA가 단백질로 번역되는 효율에 영향을 미치는 보존서열이다. 일반적으로 원핵세포의 경우 이 영역에 퓨린 계열의 염기서열이 보존되어 리보솜을 구성하는 16S rRNA 3'-말단의 서열 (UCCU)과 상보적인 서열을 이루게 된다 (Fig. 7A). 메타게놈 URF에 대한 시작코돈의 상류 영역을 수집하여, RBS 영역의 보존서열을 분석한 결과 Fig. 7B와 같은 경향을 나타내었다. 메타게놈이 수많은 미생물 유래의 유전체라는 점을 감안하면 *E. coli*와 다른 형태의 RBS 보존서열이 존재할 가능성이 있을 것 이라고 예상할 수 있다. 주변 환경에 따른 경향성을 파악하기 위하여, 세분화된 서식지별로 메타게놈의 RBS 보존서열을 분석해본 결과도 *E. coli*와 다른 형태의 보존서열을 보이거나, 특정한 보존서열이 관찰되지 않는 범주가 많음이 확인되었다.

따라서 메타게놈으로부터 유용한 유전자원을 선별하는 연구에서는, 위의 분석결과와 같은 RBS의 차이점으로 인하여 유발되는 문제점을 심도있게 고려하여야 하며 경우에 따라서는 대장균유래 RBS를 지닌 특정한 벡터 (예를 들면 Expression vector) 에 비교적 적은 크기의 유전자 단편을 삽입하여 라이브러리를 구성하는 것이 유리할 것으로 판단되어졌다.

요 약

본 연구는 메타게놈 유전자 특성과 *E. coli* 에서 정상적으로 발현되는 유전자 특성을 생물정보학 기법으로 비교·분석하고 그 결과를 메타게놈 선별 연구에 활용하고자 하는데 그 목적을 두었다. 이를 위하여 메타게놈 유래의 URF 와 숙주세포로 이용되는 *E. coli*의 ORF에 대한 염기구조, 발현되는 단백질의 크기 및 분자량, 아미노산의 구성 및 코돈사용은 물론 전사와 번역에 관여하는 프로모터 부위와 리보솜 결합부위의 보존서열 특성을 비교·분석하였다.

메타게놈과 *E. coli*가 합성하는 단백질의 크기와 분자량은 매우 비슷한 경향을 보였으나, 아미노산의 조성비, G+C 함량 및 코돈사용에서는 매우 다른 경향을 나타내었다. 특히 전사와 번역에 직접적으로 관여하는 프로모터와 RBS 영역에서의 DNA 보존서열이 상당부분 부합되지 않아 *E. coli*에서 메타게놈의 발현율이 현저히 낮을 것으로 예측할 수 있었다. RBS와 같이 유전자 발현에 필수적인 조절인자가 메타게놈과 *E. coli*에서 큰 차이를 나타내는 문제점은 메타게놈으로부터 유용한 유전자원을 탐색하는 연구에서 심도있게 개선하여야 할 사항이다. 부분적으로는 라이브러리 구축에 사용되는 벡터 및 숙주의 개량을 통하여 위의 문제를 극복할 수도 있을 것이다.

감 사

이 연구는 한국학술진흥재단 지정 인하대학교 중점연구소 (KRF-2004-005-D00006)의 지원에 의하여 수행되었으며, 이에 감사드립니다.

REFERENCES

1. Whitman, W. B., D. C. Coleman, and W. J. Wiebe. (1998), Prokaryotes: The Unseen Majority, *Proc Natl Acad Sci.* **95**, 6578-6583.
2. Amann, R. I., W. Ludwig, and K. H. Schleifer (1995), Phylogenetic identification and in situ detection of individual microbial cells without cultivation, *Microbial. Rev.* **59**, 143-169.
3. Voget, S., C. Leggewie, A. Uesbeck, C. Raasch, K. E. Jaeger, and W. R. Streit (2003), Prospecting for novel biocatalysts in a soil metagenome, *Appl. Environ. Microbiol.* **69**, 6235-6242.
4. Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M.

Goodman (1998), Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products, *Chem. Biol.* **5**, 245-249.
5. Lorenz, P., K. Liebeton, F. Niehaus, and J. Eck (2002), Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space, *Curr. Opin. Biotechnol.* **13**, 572-577.
6. Cowan, D., Q. Meyer, W. Stafford, S. Muyanga, R. Cameron, and P. Wittwer (2005), Metagenomic gene discovery: past, present and future, *Trends Biotechnol.* **23**, 321-329.
7. Galvao, T. C., W. W. Mohn, and V. D. Lorenzo (2005), Exploring the microbial biodegradation and biotransformation gene pool, *Trends Biotechnol.* **23**,497-506.
8. Henne, A., R. Daniel, R. A. Schmitz, and G. Gottschalk (1999), Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate, *Appl. Environ. Microbiol.* **65**, 3901-3907.
9. Henne, A., R. A. Schmitz, M. Bomeke, G. Gottschalk, and R. Daniel (2000), Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli, Appl. Environ. Microbiol.* **66**, 3113-3116.
10. Gupta, R., Q. K. Beg, and P. Lorenz (2002), Bacterial alkaline proteases: molecular approaches and industrial applications, *Appl. Microbiol. Biotechnol.* **59**, 15-32.
11. Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman (2000), Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms, *Appl. Environ. Microbiol.* **66**, 2541-2547.
12. Seow, K. T., G. Meurer, M. Gerlitz, E. Wendt-Pienkowski, C. R. Hutchinson, and J. Davies (1997), A study of iterative type II polyketide synthases, using bacterial genes cloned from soil DNA: a means to access and use genes from uncultured microorganisms, *J. Bacteriol.* **179**, 7360-7368.
13. Voget, S., C. Leggewie, A. Uesbeck, C. Raasch, K. E. Jaeger, and W. R. Streit (2003), Prospecting for Novel Biocatalysts in a Soil Metagenome, *Appl. Envirn. Microbiol.* **69**, 6235-6242.
14. Martines, A., S. J. Kolvek, C. L. Yip, J. Hopke, K.A. Brown, I. A. MacNeil, and M. S. Osburne (2004), Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts, *Appl. Environ. Microbiol.* **70**, 2452-2463.
15. Uchiyama, T., T. Abe, T. Ikemura, and K. Watanabe (2005), Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes, *Nat. Biotechnol.* **23**, 88-93.
16. Martinez, A., S. J. Kolvek, C. L. Yip, J. Hopke, K. A. Brown, I. A. MacNeil, M. S. Osburne, K. Watanabe, Y. Kodama, N. Hamamura, and N. Kaku (2002), Diversity, abundance, and activity of archaeal populations inoil-contaminated groundwater accumulated at the bottom of an underground crude oil storage cavity, *Appl. Environ. Microbiol.* **68**,3899-3907.
17. Schmeisser, C., C. Stockigt, C. Raasch, J. Wingernder, K. N. Timmis, D. F. Wenderoth, H.-C. Flemming, H. Liesegang, R. A. Schmitz, K. E. Jaeger, and W. R. Streit (2003), Metagenome survey of biofilms in drinking-water networks, *Appl. Environ. Microbiol.* **69**, 7298-7309.
18. Walter, J., M. Mangold, and G. W. Tannock (2005), Construction, analysis, and beta-glucanase screening of a bacterial artificial chromosome library from the large-bowel microbiota of mice, *Appl. Environ. Microbiol.* **71**, 2347-2354.
19. Hallam, S. J., N. Putnam, C. M. Preston, J. C. Detter, D. Rokhsar, P. M. Richardson, and E. F. DeLong (2004), Reverse methanogenesis: testing the hypothesis with environmental genomics, *Science* **305**, 1457-1462.
20. Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith (2004), Environmental genome shotgun sequencing of the Sargasso Sea, *Science* **304**, 66-74.
21. Kube, M., A. Beck, A. Meyerdierks, R. Amann, R. Reinhardt, and R. Rabus (2005), A catabolic gene cluster for anaerobic benzoate degradation in methanotrophic microbial Black Sea mats, *Syst. appl. Microbiol.* **28**, 287-294.
22. Snyder, L. and W. Champness (2003), Molecular Genetics of Bacteria, 2nd ed. ASM press. Washington, D. C., USA
23. Stothard, P. (2000), The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences, *BioTechniques* **28**, 1102-1104.
24. Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch (2005), Protein Identification and Analysis Tools on the ExPASy Server; pp. 571-607. *In* John M. Walker (ed.), The Proteomics Protocols Handbook. Humana Press, N. J., USA.
25. EvolvingCode by Freeland Lab. Biological Sciences Department at UMBC http://www.evolvingcode.net/codon/cai/cai.php
26. Reese, M. G. (2001), Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome, *Comput. Chem.* **26**, 51-56.
27. Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner (2004), WebLogo: A sequence logo generator, *Genome Research* **14**, 1188-1190.
28. Institute of Bioinformatics and Structural Biology, NTHU-PCLyu's Lab http://gpdb.life.nthu.edu.tw/GPDB/index.php
29. Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyew, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield (2004), Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature* **428**, 37-43.
30. Gouy, M. and C. Gautier (1982), Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acids Res.* **10**,7055-7074.
31. Ikemura, T. (1981), Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, *J. Mol. Biol.* **151**, 389-409.