◎◉◎ 총 설

Genome Annotation and Antimicrobial Target

글 _ 한원석[1], 윤창노[2] _ [1](주)나노믹스, [2]한국과학기술연구원

서 론

인간유전체 프로젝트 (1)를 포함한 수많은 염기서열 분석 프로젝트들로 인하여 National Center for Biotechnology Information (NCBI)에서 운영하는 GenBank (2)의 염기서열 수는 매 14개월 마다 두 배씩 증가하고 있으며 이러한 증가 추세는 한 동안 계속될 것으로 보인다. 그러나 축적되는 막대한 양의 데이터들이 생명과학 분야의 연구자들에게 쉽게 접근할 수 없거나 알 수 없는 형태로 보여 진다면 세계각지에서 벌이는 염기서열 분석의 노력이 무용지물이 될 것이다. 따라서 National Human Genome Research Institute (NHGRI, 3)를 포함하여 많은 연구기관에서는 사용자들이 쉽게 데이터에 접근하도록 컴퓨터 프로그램과 사용자 인터페이스 제작에 많은 노력을 기울이고 있다.

이 글에서는 우리들이 쉽게 접근할 수 있는 비교적 잘 구성된 Genome Annotation 용 데이터베이스 시스템을 소개하도록 네 부분으로 구성하였다. 본론의 첫 번째 부분에서는 사람의 유전자 이상에 의한 질병 데이터베이스인 Online Mendelian Inheritance in Man (OMIM, 4)과 생명과학자들에게 가장 많이 알려진 NCBI Entrez 시스템 (5), Annotation을 잘 강조한 University of California, Santa Cruz (UCSC)의 Genome Browser (6), 종합적인 Genomic Data를 제공하는 NCBI Map Viewer, EST와 유전자 서열 데이터를 제공하는 The Institute for Genomic Research (TIGR)의 Genome Databases와 Gene Indices (7)들을 소개하려고 한다. 두 번째 부분에서는 Post Genomic 입장에서 확장된 정

Table 1. Number of OMIM Entries (2004년 3월 현재)

	Autosoma	X-Linked	Y-Linked	Mitochondria	Total
Established genes or phenotype loci(*)	10645	575	46	37	11303
Phenotype descriptions(#)	1352	116	0	23	1488
Other loci or phenotypes (*no prefix*)	2226	156	2	0	2384
Total	14224	847	48	60	15179

의의 Function Annotation을 위한 데이터베이스 통합과 그 결과로부터 분석된 통계적인 데이터를 보게 될 것이다. 세 번째 부분에서는 앞에서 소개하였던 일반적인 목적으로 만들어진 데이터베이스 시스템과 달리 다양한 목적의 특별하게 제작된 데이터베이스들 (8)이 많이 출현하게 되는데 그 중에서 미생물을 대상으로 한 Genome Annotation System의 한 예를, 네 번째 부분에서는 Genome Scale에서의 단백질 3D 구조 모델링을 통한 Function Annotation System을 소개하도록 하겠다. 그리고 이러한 시스템들의 응용 예로서 단백질 엔지니어닝, 항생제 타겟과 내성에 대해서 살펴보는 것으로 끝을 맺으려고 한다.

본 론

가. Genome Browser

Online Mendelian Inheritance in Man (OMIM)

Johns Hopkins 대학의 McKusick 연구팀에 의해 만들어졌고 사람의 유전자 이상과 질병관계를 정리하여 발간했던 "Online Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders"란 제목을 가지는 책의 전자버전 (9,10)으로 유전학적 근거 하에서 매우 자세한 문헌정보 및 내용이 들어 있다. 또한 FTP site (11)를 운영하기 때문에 데이터베이스를 download하여 사용할 수도 있다. 보통의 컴퓨터와 인터넷 브라우저인 MS Explorer나 Netscape Navigator들을 필요로 한다. 이러한 OMIM이 제공하는 Browser로는 여러 가지가 있는데 그 중에서 중요한 것들을 살펴보면

1) OMIM Database : 관련 논문자료 정보
2) OMIM Gene Map : 질병관련 발현 유전자의 cytogenetic map의 위치
3) OMIM Morbid map : 질병에 의해 조직화된 질병 유전자 list들로,

탐색하는데 사용할 수 있는 query로는 chromosome [CH], allelic variant [AV], clinical synopsis [CS], EC number [EC], cytogenetic map location [GM], disorder [DIS], gene symbol [GN], 참고문헌에서 사용한 제목 또는 저자명 [RE] 등으로 매우 다양하다.

OMIM에서 사용하는 데이터베이스로는 NCBI의 Genes and Disease Database, Genome Database (GDB), Cardiff Human Gene Mutation Database (HGMD), Jackson Laboratory Mouse Genome Database 등이다.

다음은 FTP site에서 download 받을 수 있는 데이터들이다.

- omim.txt.z Complete OMIM text
- genemap OMIM Gene Map
- genemap.key OMIM Gene Map Key
- morbidmap OMIM Morbid Map
- genetable OMIM Gene Table : gene symbol, OMIM accession number

http://www.ncbi.nlm.nih.gov/Omim/Index/genetable.html

The NCBI Entrez System (12)

Entrez 웹 인터페이스는 http://www.ncbi.nlm.nih.gov/Entrez로 대부분의 NCBI webpage들이 연결되어 있다. 이 시스템은 20개의 데이터베이스들에 대해 간단한 text search를 통하여 들어갈 수 있는 통합 데이터베이스 시스템으로, 다루는 데이터들을 살펴보면 다음과 같다. GenBank, Protein Information Resource, SWISS-PROT, PDB, RefSeq들로부터 찾아낼 수 있는 DNA와 단백질 서열, NCBI taxonomy, genome, gene expression data, UniGene 속의 sequence cluster, UniSTS data, dbSNP, Molecular Modeling Database가 보유하는 단백질 구조, 단백질 도메인, PubMed, Pubmed Central, OMIM 속의 문헌자료들을 들 수 있다. 특히 PubMed는 1,270만개의 Medline 자료와 4,000여개 전문잡지들의 본문을 볼 수 있도록 해주고 있다. 최근에는 사용자의 편리성을 위해 NCBI website를 데이터베이스 목록에 넣어 바로 연결해 볼 수 있도록 하였다.

이 시스템의 장점은 데이터베이스 상호간의 연결이 자유로워 서열과 관련된 문헌자료, 단백질 서열과 대응되는 3D 구조, alignment와 annotation 정보까지 볼 수 있으며, 서열이나 초록정보의 similarity를 사용한 "neighbor"란 링크를 주어 관련된 데이터들로 빠르게 접속하도록 해 주고 있다. 또한 외부 데이터베이스로 연결시켜 주는"LinkOut"과 추가되는 링크 사이트들을 위한"Links"들이 매우 유용하다.

탐색하여 찾은 정보들은 여러 가지 format으로 보거나 download 받을 수 있으며, text 형태로 사용자 파일에 저장하거나 email로 보낼 수 있다. GenBank의 경우 일정 범위의 염기나 단백질 서열을 GenBank Flatfile, FASTA, XML, ASN.1 등의 format으로 보거나 download 받을 수 있고, Entrez Programming Utilities에서 제공하는 script를 사용하면 Entrez 데이터베이스들을 탐색하거나, 링크하거나, download 받거나 하는 작업들을 쉽게 수행하고 history 형태로 결과를 저장할 수도 있다.

UCSC Genome Browser

UCSC에서 계산한 annotation data 즉, gene prediction, mRNA alignment, EST alignment, 종간 homology, singlenucleotide polymorphism들을 볼 수 있는 브라우저 (13)로 여러 가지 scale로 표현되고 있다. 이 시스템에서 서열에 대한 annotation data를 보려면 두 가지 접속 방법을 사용할 수 있는데, 하나는 UCSC Genome Bioinformatics 홈페이지(http://genome.ucsc.edu)에 접속하여 제공되는 그래픽 웹브라우저를 통하는 것이고, 다른 하나는 그 연구팀에서 운영하는 FTP site (ftp://genome.cse.ucsc.edu/goldenPath)로부터 데이터 정보들을 직접 download 받는 것이다. 정보 탐색을 위한 입구로 BLAT (BLAST-like Alignment Tool)을 사용하며, 찾아진 정보들은 유전체 좌표 위치에 annotation track을 이미지 형태로 보여 준다. "Zoom in"과"zoom out" 버튼을 사용하여 10배까지 볼 수 있으며, GeneLynx, GeneCards, AceView, RefSeq, OMIM, LocusLink, PubMed 등의 외부 데이터베이스들과 링크를 시켜 놓아서 바로 관계된 자료들을 볼 수 있게 하였다. 다양한 용도에 따라 annotation 정보를 그래픽으로 표현하도록 하였기 때문에 사용자를 위한 UCSC Genome Browser User's Guide (14)를 운영하니 참고하면 편리하다.

The NCBI Map Viewer

이 viewer는 sequence, cytogenetic, genetic linkage, radiation hybrid map, annotated genomic sequence들을 보여 주는 인터페이스로 보통 NCBI 홈페이지에서 시작할 수 있다. BLAST (Basic Local Alignment Search Tool, 15-16) 페이지에 링크되어 있고 H. sapiens, M. musculus, R. norvegicus 들을 포함하여 19종의 유전체 정보를 다루고 있으며 제공되는 데이터는 cytogenetic map, physical map, 예측 gene model을 사용한 map, UniGene cluster에 연결시킨 EST alignment와 mRNA alignment들이다. 또한 LocusLink, Evidence Viewer, Model Maker와 링크를 시켜 놓아 매우 유용할 뿐만 아니라, genome assembly 조각들은 이 Viewer 의 "Download/View Sequence" 링크를 사용하여 GenBank나 FASTA format으로download 받을 수 있도록 해 주고 있다. 탐색할 수 있는 query로는 gene name, gene symbol, marker name, SNP identifier, accession number들이 있다.

TIGR Genome Databases

TIGR Genome Databases 웹페이지(http://www.tigr.org/tdb)에서 제공되는 genomic databases는 크게 Comprehensive Microbial Resource (CMR)과 Eukaryotic Resource로 구성 되어 있다.

1. Comprehensive Microbial Resource (17,18)

미생물유전체분석정보를담고있는데이터베이스를 Omniome이라 부르며 2004년 초 현재 138종의 genome을 보유하고 있다. 이 중에 136종이 완성된 것이고 2종

이 미완성 된 것이며, 16종의 archaea와 122종의 bacteria genome들이다. 또한 이 데이터들은 외부의 NCBI genome, COG (cluster of ortholog groups), GenBank 들과 링크를 시켜 놓았다. 제공되는 탐색시스템의 몇 가지 특징적인 기능을 살펴보면 다음과 같다.

A. Multi-Genome Applications
   (1) Multi-Genome Query : Queries across all genomes in the database
   (2) Align Whole Genomes : Aligns any two genomes
   (3) Batch Download : Retrieves to your local disk a large number of sequences

B. Multi-Genome Analyses
   (1) Role Category Survey : Shows the number and percentage of genes in each genome for every role category in the CMR
   (2) Gene Ontology (GO) Survey : Shows the number of GO terms
   (3) Operon Predictions : Predictions of operons in microbial genomes

C. Multi-Genome Searches
   (1) Locus : Search based on Primary or TIGR annotation loci names/accession numbers, SWISSPROT/TrEMBL ACs, GenBank Accession or GI numbers or EcoCyc IDs

D. Multi-Genome Lists by Category
   (1) Genes by Role Category : Lists all genes in the CMR broken down into role categories.
   (2) Enzyme Commission Numbers : Lists all EC#s assigned to genes in the CMR

2. Eukaryotic Resource (19)

CMR과 같이 다양한 비교 search 기능이 없고 아래에 list 된 각 database 내에서 탐색할 수 있도록 BLAST server를 제공한다.

Databases :

(1) *Tetrahymena thermophila, Theileria parva, Arabidopsis thaliana,* Rice, Potato Functional Genomics 등
(2) Parasites Database; *Brugia malayi, Entamoeba histolytica, Plasmodium falciparum, Plasmodium vivax, Plasmodium yoelii, Toxoplasma gondii, Trichomonas vaginalis, Trypanosoma* brucei, *Trypanosoma cruzi, Schistosoma mansoni*

TIGR Gene Indices

TIGR Gene Indices 웹페이지 (http://www.tigr.org/tgi)는 동물 23종과 식물 22종을 포함한 약 67종에 대해 EST와 gene sequence 데이터들을 보여주고 있으며, GenBank의 dbEST 데이터를 사용하여 Gene Index Assembly 과정을 거친다. 현재 GenBank에서 보유하고 있는 1천만개 이상의 EST 데이터를 살펴보면 45% 정도가 human으로부터 얻어졌으며 75% 가량이 human, mouse 등을 포함하는 고등 포유류의 것 (20)이다. 이 시스템에서 제공하는 정보 중에는 annotation, genes, genome structure, genomic localization, orthologs, paralogs 등이 있으며 홈페이지에 있는 FTP link를 통해 데이터베이스를 download 받을 수 있다.

TIGR Gene Index Assembly 과정을 간단히 살펴보면, 먼저 NCBI의 dbEST 데이터베이스에서 EST sequences를 download 받아서 contaminating vector, adapter, mitochondrial, ribosomal sequence들을 제거하는 EST cleaning을 거치고, GenBank에서 parsing된 annotated coding sequence(CDS region)들을 받는다. 모든 EST와 gene sequence들을 대상으로 all-versus-all pairwise similarity search를 수행하여 그룹으로 묶는다. 이때 mgBLAST 프로그램을 사용하는데, 이것은 megaBLAST (21)의 minimum overlap length (default 40basepair)와 identity (default 95%) 등의 출력 option을 변형한 프로그램이다. 각 cluster 들은 Tentative

Consensus (TC) sequences로 모아지고 annotation 되며 TIGR website에서 제공된다.

나. Integration of Biological Databases

앞에서 보았던 다양한 웹기반 데이터베이스 시스템들을 우리는 쉽게 접근할 수 있고 필요한 정보들을 알아낼 수 있으며 자세한 문헌 탐색을 통해 실험 데이터들도 볼 수 있다. 대부분의 이러한 시스템들이 제공하는 정보는 원천 데이터에 가까운 것들이고 매우 잘 정리되어 있어서 관련 정보들을 최대한 찾아 주고 있다. 더욱이 BLAST (BLAST-like Alignment Tool) search tool을 가지는 UCSC Genome Browser를 포함하여 대부분의 시스템들이 BLAST search를 입구로 하여 정보를 찾아 들어가기 때문에 sequence homology를 사용한 유사 유전자 또는 단백질들의 정보를 모두 얻어내어 genome annotation 의 범위를 한층 확장시켜 주고 (약 40% annotation) 있다. 그리고 여러 종간의 유전자 정보를 이용한 comparative genomics를 거치기 때문에 비어 있는 공백을 많이 채워 넣을 수 (10-20 % annotation 증가 효과) 있게 해 준다.

Post-genomic View of Gene/Protein Function

이제 본격적으로 function annotation에 대해 생각해 볼 단계이다. 지금까지 우리는 유전자 또는 단백질의 function을 정의할 때 한 substrate에 대해 작용하여 product를 만들어 내는 molecular function의 의미로 생각해 왔다. 그러나 지금은 이 정의가 확장되어 유전자 또는 단백질이 한 상호작용 (interaction) 네트워크에 속하게 되어 구성원으로서의 역할을 하는 cellular function 또는 processing 이 그 의미를 갖게 된다. 그러나 앞에서 소개되었던 genome browser들이 제공하는 많은 function annotation 데이터들은 주로 sequence homology에 기반 하여 chemical 또는 molecular function에 대한 정보들을 주고 있다. 그래서 우리가 확장된 정의에 따른 function annotation 을 수행할 때는 functional

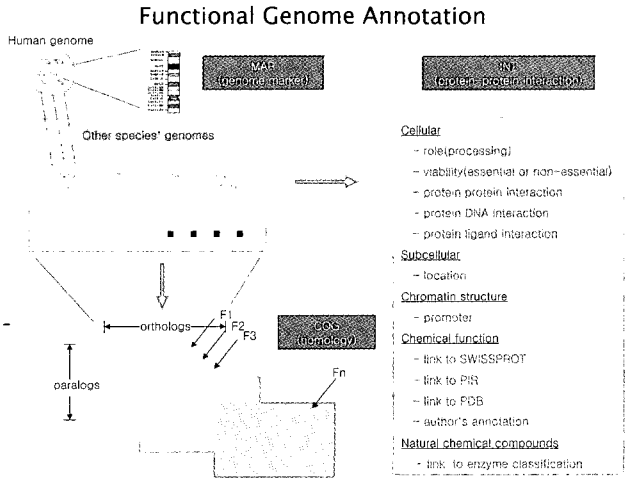


그림 1. Genome annotation 개념도

network을 사용하며, 자연스럽게 protein-protein interaction을 고려하게 된다. 따라서 genome annotation을 위한 데이터베이스들의 통합을 위해서는 protein-protein interaction data를 필수적으로 사용하게 된다.

Integration of Biological Databases for Genome Annotation

그림 1에서 genome annotation을 위한 개념도를 볼 수 있는데, human genome을 대상으로 한 그림이다. Marker를 통하여 chromosome 상의 위치를 파악하며, 접근 가능한 다른 종들의 genome database들을 사용한 comparative genomics를 수행하여 gene들 간의 orthologs와 paralogs 관계를 확인한다. 이 역할을 COG가 해주며, protein-protein interaction data를 사용하여 INT가 functional annotation을 수행한다. 그림에서 볼 수 있는 cellular role, protein/DNA/ligand interactions, location, promoter, protein structure 등을 포함하여 chromosome, contig number, cytogenetic map, disease, phenotype, metabolic과 regulatory pathways, phylogenetic profile, expression profile 등의 정보들을 볼 수 있다.

위의 개념도를 사용하여 NCBI 의 GenBank로부터 *H. sapiens*를 포함한 eukaryotes 14종과 *H. pylori*를 포함한 prokaryotes 75종, HIV virus를 포함한 virus

◈ 총 설

Table 2. Predicted operons with experimental evidences

Predicted operons^a	Known operons^b	GN^c	References
b2451 eutH eutJ cchB cchA b2459 b2461 b2462	*eutH cchB cchA*	st	Stojilkovic (27)
rnk citG citX citF citD citC	*citG citF citC*	kp	Schneider (28)
b1483 b1484 b1485 b1486 b1487 ydeV	*nikC nikB nikA*	bs	Jubier-Maurin (29)
agaV agaB agaC agaD agaI	*sorB sorC sorD*	lc	Yebra (30)
b1627 b1628 ydgO b1631 ydgO	*nqrE nqrD*	vc	Hase (31)
yccZ ymcA ymcB ymcC	*wbfB wbfD*	vc	Yamasaki (32)
b0829 b0830 b0831 b0832	*nikA nikB nikC*	bs	Jubier-Maurin (29)
yjbF yjbG yjbH yjbA	*wbfD wbfB*	vc	Yamasaki (32)
yhcI yhcJ nanA yhcK	*yhcK nanA*	ec	Oshima (33)
fba pgk epd tktA	*pgk gap*	cg	Schwinde (34)
vecC yecS fliY fliA	*gluA gluB*	cg	Kronemeyer (35)
spy astE b1745 b1747	*aruE aruB*	pa	Itoh (36)
fumB dcuR dcuS	*citB citA*	kp	Schneider (28)
yhdW yhdX yhdY	*aapJ aapQ aapM*	rl	Parker (37)
b2866 ygeT b2868	*ndhL ndhM ndhS*	an	Baitsch (38)
acrD yffB dapE	*yffB dapE*	ec	Oshima (33)
pgsA uvrC uvrY	*uvrC uvrY*	ec	Oshima (33)
ynhA ynhC ydiC	*ynhA ynhC*	ec	Oshima (33)
paaF paaG paaJ	*badK badI*	rp	Egland (39)

a The underlined names of genes in the predicted operons corresponded to the genes with the known operons.

b Known operons were collected from gene expression database and gene cluster database.

c Genome names.: an:*Arthrobacter nicotinovorans,* bs:*Brucella suis,* cg:*Corynebacterium glutamicum,* ec:*Escherichia coli,* kp:*Klebsiella pneumoniae,* lc:*Lactobacillus casei,* pa:*Pseudomonas aeruginosa,* rl:*Rhizobium leguminosarum,* rp:*Rhodopseudomonas palustris,* ss:*Synechocystis* sp, st:*Salmonella typhimurium,* and vc: *Vibrio cholera.*

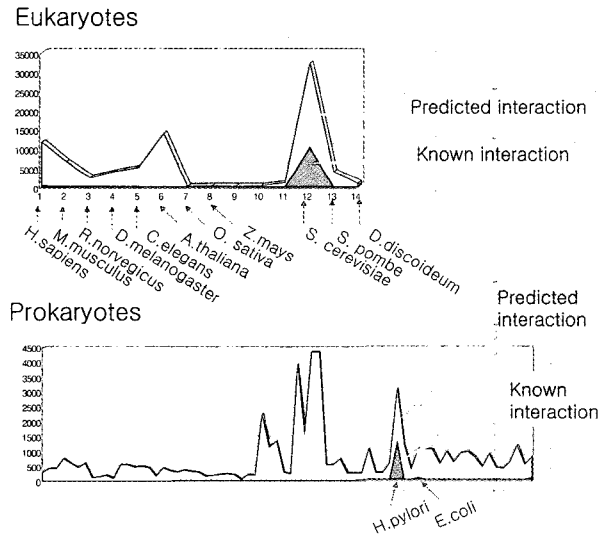


그림 2. Protein-protein interaction data 의 species별 분포도

816종 에 대해 genome annotation을 수행한 결과를 살펴보면, 약 5만 8천개 가량의 human gene을 대상으로 했을 때 comparative genomics와 protein interaction data (22,23,24)를 사용하여 약 35% 가량의 annotation 증가효과를 보아서 약 75% 정도의 genome annotation 이 가능했다. Species 간의 common function 수를 비교해 보면 human gene function 2만 8천개에 대해서 human-mouse간에 7600개, human-rat 간에 2800개, human-yeast 간에 850개, mouse 1만 6천개와 rat 4600개의 gene function에 대한 mouse-rat 간에는 2600개의 common function을 갖는다. Annotation 결과로부터 얻어낸 protein-protein interaction data의 species 별 분포 (그림 2)를 보면 human 11236개, mouse 6407개, *C. elegans* 4952개, yeast 21896개, *H. pylori* 1777개, *E. coli* 985개 등이고 실험 데이터를 갖고 있는 개

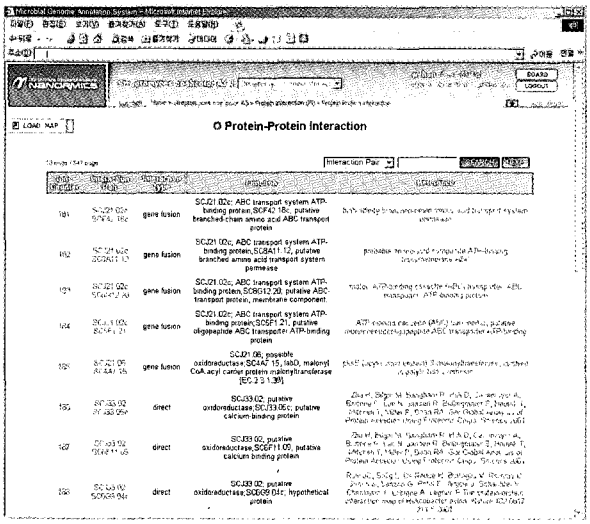
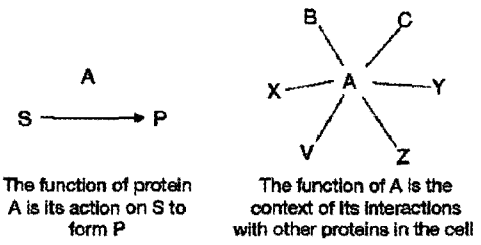


그림 3. Table of protein-protein interaction pairs

수는 2002년 10월 현재 human 396개, mouse 73개, *C. elegans* 122개, yeast 10437개, *H. pylori* 1271개, *E. coli* 144개였다.

그림 3에서 보여주는 table을 살펴보면, *Streptomyces coelicolor* A3 의 protein-protein interaction data들로 ABC transport system ATP-binding protein인 ID "SCJ21.02C"와 interaction partner 3개를 보여 주는데 그들의 function을 보면 ABC transport system permease, membrane component들인 것을 알 수 있으며, ID "SCJ21.06"인 oxidoreductase와의 partner로 malonyl CoA-ACP malonyltransferase인 fabD, ID "SCJ33.02" 의 oxidoreductase에 대한 interaction partner 3개는 calcium binding protein들로 확인된다.

Human genome에 대한 functional annotation 결과로부터 얻어낼 수 있는 정보들 중에는 target gene list를 들 수 있는데 그림 4에서 볼 수 있다. Neurological, endocrine, metabolic등의 phenotype list로 약 2500개, 582개의 virus로부터 acquired 된 human gene, 이들 중에 480개는 oncogene으로 확인되었다. 또한 551개의 bacteria로부터 human acquired된 gene들이 있는데 이들 중에는 357개의 문헌 데이터를 확인할 수 있었다.



질병 관련 tissue specific gene이 4075개, chromosome specific gene이 2303개 확인 가능하였다.

다. Microbial Genome Annotation System

특별한 목적을 위해 구축된 microbial genome annotation system 중에서 biosynthesis 용 gene cluster를 찾아줄 수 있는 시스템, RNGAS (25)을 소개하고자 한다. 이 시스템은 structural, functional, comparative genomics의 세부분으로 구성되어 있다. 주어진 full genome에 대해서 protein coding region이 확인되면 앞에서 보았던 일련의 annotation 과정을 거친 후 operon prediction (Table 2에서 일부 예를 보임) 이 이루어지고, 이 operon들 (26)을 바탕으로 global regulatory network을 구성한다. 이 network 의 목적은 전체 genome map에 모든 정보를 표시하고 기능적으로 연관성이 있는 gene들을 그룹으로 묶어서 전체적인 구성을 확인할 수 있게 해

Target Database

Phenotype DB

Neurological	205
Endocrine	131
Deafness	48
Cardiovescular	70
Ophthalmologic	210
Pulmonary	9
Gastrointestinal	12
Renal	20
Immunological	89
Hematologic	106
Coagulation_abnormalities	29
Malignancies	227
Skeletal_Development	51
Soft_Tissue	3
Connective_Tissue	14
Dermatologic	37
Metabolic-mitochondrial	963
Pharmacologic	12
Peroxisomal	10
Storage	95
Pleitropic_Developmental	35
Etc.	513

Human-acquisitions of Viruses genes

Total known oncogenes	759
Human acquisitions of Viruses genes	582
Known(60) or patented(420) oncogenes	~420
New possible oncogenes	162

Human-acquisitions of Bacterial genes

Human acquisitions of Bacterial genes	551
OMIM linked genes	357

Disease-linked tissue(4<) specific protein DB

Unigene DB	1147
Swiss-prot DB	3928
disease linked tissue specific proteins	4075

Disease-linked chromosome(2<) specific protein DB

chromosome specific proteins in human	23757
disease linked chromosome specific proteins	2303

그림 4. Target database (2002년 10월 현재)

준다. 알 수 있는 정보로는 primary와 secondary metabolic encoding genes, regulon, modulon, stimulon, protein-protein interactions들이 있다.

Genome structure 모듈은 gene structure와 regulatory machinery를 볼 수 있도록 하기 위해 일반적인 gene prediction 프로그램으로 얻어진 protein coding region을 사용하며 gene expression과 regulation에 관련된 정보를 제공하는 것이 목적이다. 또한 이 module은 CONTIG, BlastX 결과, ORF, amino acid sequence 정보들을 실시간으로 탐색할 수 있도록 구성되어 있다.

Gene cluster 모듈은 대상 genome의 기능적으로 관계있는 operon들을 묶어 주어서 1차와 2차대사 경로, 그리고 신호 조절 네트워크 등의 구성을 위해 사용하며, operon, phylogenetic profile, metabolic pathway, gene expression profile, protein-protein interaction 등의 정보를 다룬다. 이러한 기본적인 정보들은 논문 등의 문헌정보로부터 대부분 얻어지고 있으며, 잘 정리된 데이터들은 대상 genome에 gene cluster의 형태로 align된다. 이렇게 align된 대상 genome은 다른 모든 알려진 prokaryote genome과 같이 비교, 표현된다. 사용되는 비교 genome의 개수는 약 100개 이상이고 브라우저 형태로 보여 준다.

Protein interaction 모듈은 protein-protein interaction 데이터를 보여 주는데 이러한 상호작용을 조절할 수 있는 물질 개발을 하는데 매우 중요할 뿐만 아니라, target protein을 찾는데 사용될 수 있다. 여기서는 새로운 genome이 대상일 경우 실험으로 확인이 되어 있지 않기 때문에 되도록 가능성이 높은 protein interaction 데이터를 주도록 하고 있다. 그래서 이러한 데이터들의 신빙성을 높이기 위해서 많은 양의 실험 문헌정보를 통한 데이터베이스 구축을 요하고 이러한 실험 데이터를 기반으로 하여 INTEROLOG와 ROSETTA 방법을 사용하고 있다.

Gene ontology 모듈에서는 전형적인 genome annotation 과정을 통하여 chemical과 cellular function을 얻어내는

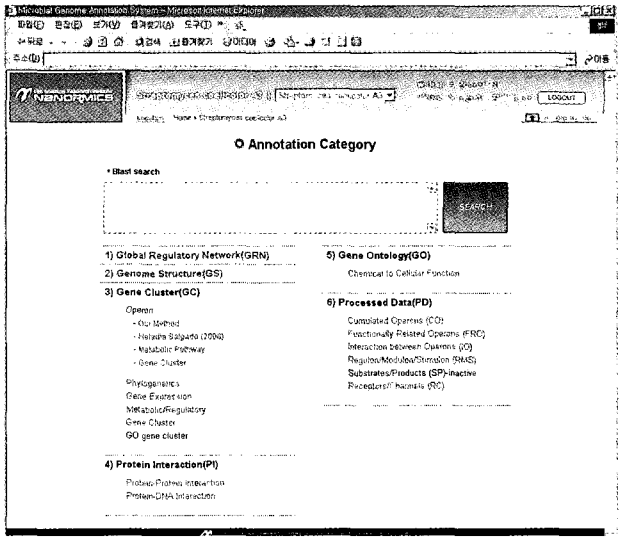


그림 5. 초기메뉴에서 보여 주는 모듈들과 입력창

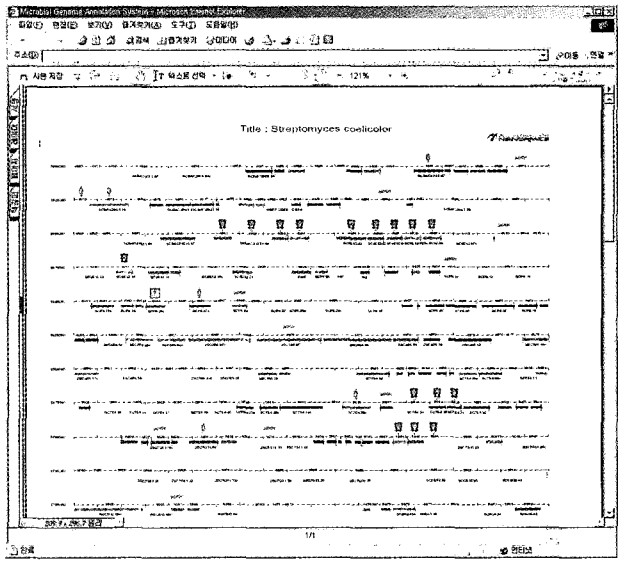


그림 6. Folic acid biosynthesis 관련 gene들 중에서 *folP* (dihydropteroate synthase) gene 이 있는 physical map 부분

데, 가장 일반적으로 사용하는 Gene Ontology International Consortium (GO) 방법과 Clusters of Orthologous Group (COG) 방법 등을 기본으로 한다. 이 시스템을 위해 사용되는 중요한 protein function들로는 polyketide & non-ribosomal peptide synthesis, receptors &

channels, regulatory proteins, phosphotase, DNA/RNA binding, metal binding 등이 있다. Function annotation 모듈로부터 얻어낸 정보로는 cumulated operons, functionally related operons, interaction between operons, substrates & product, receptors & channels 등이다. 그림 5에서 이러한 모듈들을 볼 수 있다. 이 초기 입력창에서 관심 있는 서열을 집어넣고 blast를 실행시키면 homology가 있는 서열들이 결과로 나오게 된다. 다른 탐색방법으로는 gene ontology 입력창에서 찾고자 하는 키워드를 넣게 된다. 한 예로 folic acid biosynthesis 관련 gene을 찾기 위해서 ontology search 창에"folic acid"를 집어넣으면 관련 gene들의 결과를 보여 주며 이 중에서 관심 있는 gene을 클릭하면 genome physical map에서의 위치와 주변 지도를 볼 수 있다. 그림 6에서 보여 주는 것은 folic acid biosynthesis 관련 gene들 중에서 *folP*(dihydropteroate synthase) gene 이 있는 physical map 부분이다. 네 번째 중간 부분에서 *folP*를 볼 수 있다. 또한 다른 100여종의 genome들과 comparative genomics를 한 결과를 그림 7에서

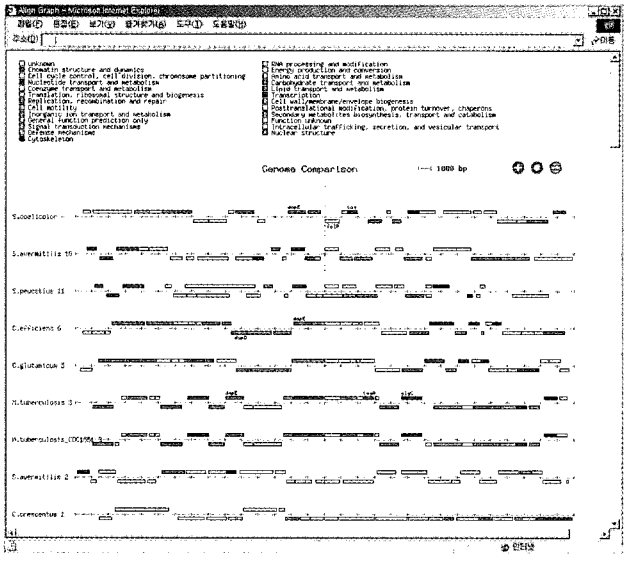


그림 7. Folic acid biosynthesis 관련 gene들 중에서 *folP* (dihydropteroate synthase) gene 이 있는 genome structure 부분으로 comparative genomics 결과를 보여줌

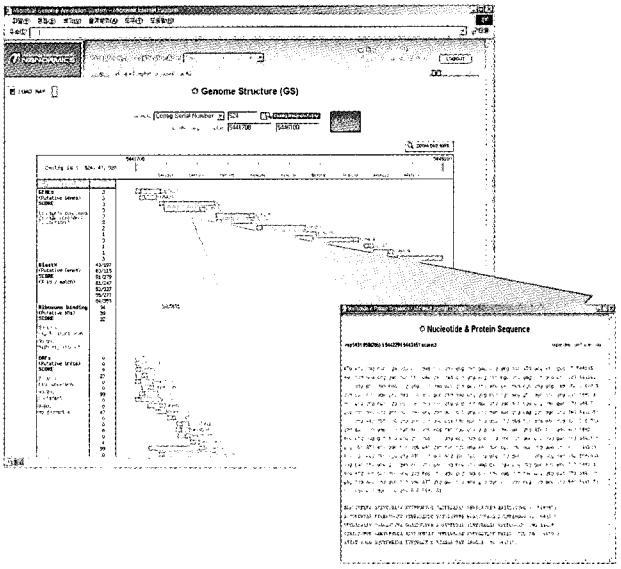


그림 8. Folic acid biosynthesis 관련 gene들 중에서 *folP* (dihydropteroate synthase) gene 이 있는 physical map 에서의 위치와 sequence 들을 보여줌

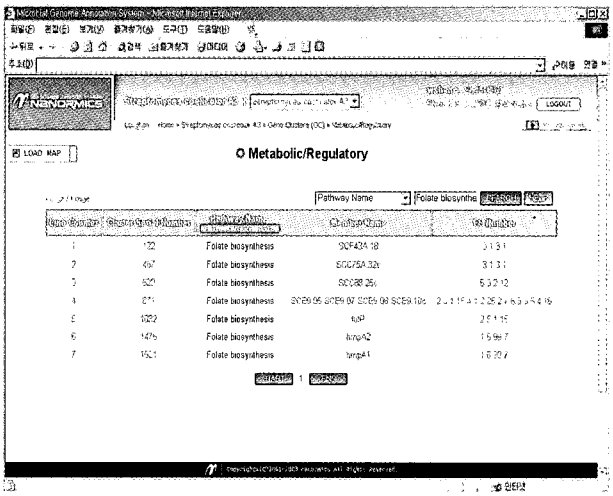


그림 9. Folic acid biosynthesis 관련 gene들을 사용한 metabolic pathway 재구성

볼 수 있으며, 그림 8에서는 염기와 아미노산 서열을 볼 수 있다. 따라서 그림 9에서와 같이 이러한 정보들을 바탕으로 하여 metabolic pathway를 재구성할 수 있게 된다. 그림 10은 actinorhodin과 prodiginine biosynthesis 관련 gene cluster들의 배치를 보여 준다.

◈ 총 설

Table 3. NanoModelDB Statistics

Species	#of genes	#of High Quality Models(%coverage)	#of Medium Quality Models(%coverage)	#of Low Quality Models(%coverage)	Total#of Models (%coverage)
1. Homo sapiens	45157	14156(31.35%)	1905(4.22%)	13788(30.53%)	29849(66.10%)
2. Mus musculus	39234	11172(28.48%)	1852(4.72%)	12405(31.62%)	25429(6481%)
3. Rattus norvegicus	45713	12593(27.55%)	1730(3.78%)	14421(31.55%)	28744(62.88%)
4. Brachydanio rerio	2175	1050(48.28%)	106(4.87%)	636(29.24%)	1792(82.39%)
5. Drosophila melanogaster	22295	5440(24.40%)	986(4.42%)	7774(34.87%)	14200(63.69%)
6. Caenorhabditis elegans	21453	3632(16.93%)	793(3.70%)	7369(34.35%)	11794(54.98%)
7. Arabidopsis thaliana	38480	9269(24.09%)	1422(3.70%)	12564(32.65%)	23255(60.43%)
8. Oryza sativa	17680	3170(17.93%)	603(3.41%)	5182(29.31%)	8955(50.65%)
9. Plasmodium falciparum	8062	1380(17.12%)	250(3.10%)	3450(42.79%)	5080(63.01%)
10. Dictyostelium discoideum	2434	649(26.66%)	132(5.42%)	892(36.65%)	1673(68.73%)
11. Saccharomyces cerevisiae	6820	1434(21.03%)	248(3.64%)	2298(33.70%)	3980(58.36%)
12. Schizosaccharomyces pombe	5380	1298(24.13%)	180(3.35%)	1872(34.80%)	3350(62.27%)
13. Encephalitozoon cuniculi	1929	356(18.46%)	79(4.10%)	708(36.70%)	1143(59.25%)
14. Escherichia coli	17275	3943(22.82%)	551(3.19%)	4929(28.53%)	9423(54.55%)
외 79종					
TOTAL GENES	454875	111572(24.53%)	19067(4.19%)	148540(32.66%)	279179(61.37%)

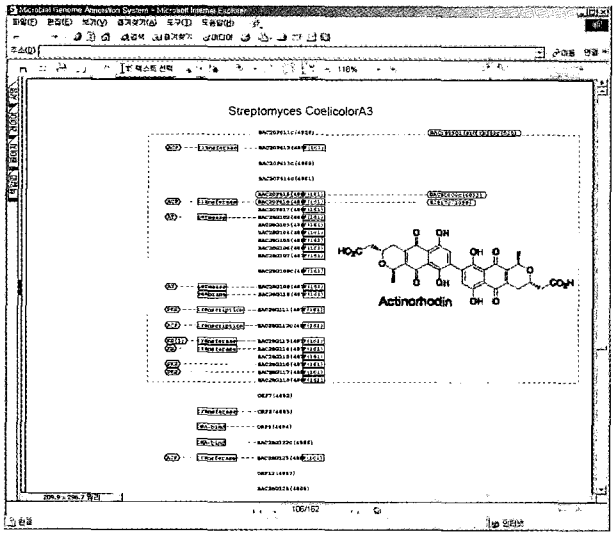


그림 10a. Actinorhodin biosynthesis 관련 gene cluster 의 배치도

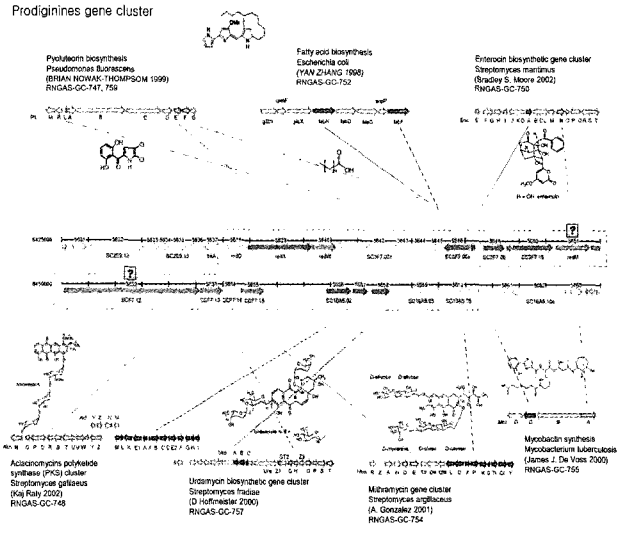


그림 10b. Prodiginines biosynthesis 관련 gene cluster 의 배치도

라. Structure Browser for Genome Annotation

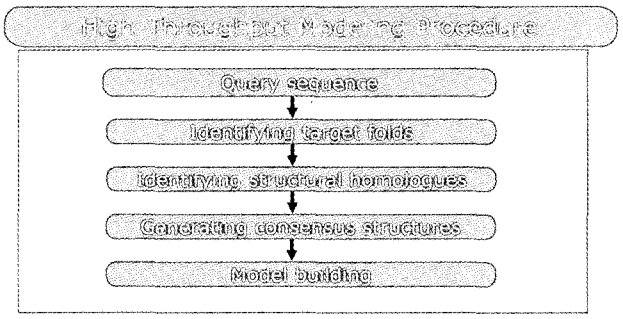
Genome annotation의 범위를 확장하기 위해서 여러 가지 방법들을 동원할 수 있는데 그 중에 단백질 구조를 사용하는 것도 한 방법이 되고 있다. 보통 unknown function의 gene annotation을 위해서 sequence homology를 사용하여 유사 유전자나 단백질들을 찾아내고 이들의 annotation 정보를 이용하거나 protein interaction data를 통하여 functional network의 구성원들을 추정하고 이들의 annotated function들을 이용하게 되는데, 이들 방법들은 대략 전체적인 sequence나 정해진 범위

의 sequence들을 사용하게 되며 homology의 정도를 변화시켜 가면서 annotation 범위를 조절할 수 있다. 그러나 기본적으로 sequence homology에 기반 하기 때문에 낮은 homology의 sequence를 가지지만 비슷하거나 같은 function을 하는 유전자나 단백질을 찾아내기는 어렵다. 이러한 점을 보완할 수 있는 방법 중에 구조를 이용한 방법도 그 한 가지가 될 수 있다.

먼저 genome scale의 protein model을 대량으로 만들어 내는 것이 필요하고, 만들어진 model을 사용하여 구조적으로 유사하거나 같은 부분을 갖는 known protein을 확인하여 찾아낸 후, 그 protein의 function 관련 정보를 annotation에 이용하는 것이다. 이러한 일을 할 수 있는 system들 중에 한가지인 NanoModel DB system (40)을 예로 하여 살펴보면 사용가능한 query로는 gene name, Swiss-prot entry name/accession number, TrEMBL entry name/accession number들이고, 찾아주는 정보는 model structure (quality information, coordinate), functional classification of model (SCOP), domain/family/super family information, query template alignment information, function of template들이다. 이 시스템에서 갖고 있는 protein model database는 High-Throughput Modeling Procedure를 거쳐서 구축되는데, 다음의 과정을 거친다.

1) 주어진 서열에 대하여서 유사한 서열을 가지는 가능한 많은 서열을 모아 multiple sequence alignment 하고 구조 형성
2) 같은 fold를 가지는 구조를 clustering
3) 같은 fold의 구조를 겹쳐서 low-resolution weighted consensus structure 형성
4) 구조의 variable region의 구조 모델링 및 구조 최적화

위의 과정을 거쳐서 만들어진 데이터베이스는 Table 3에서 볼 수 있는 바와 같이 2003년 10월 현재 *H. sapience*, *M. musculus* 등을 포함한 full genome 93종, 약 45만개의 gene들에 대해서 61%에 달하는 약 27만개의 protein



High-Throughput Modeling Procedure

model을 보유하고 있다. 이 Table에서 high quality model은 매우 정확한 model로 조금만 더 refinement 과정을 거치면 drug design을 위한 protein model 로 사용될 수 있는 정도이고, medium과 low quality model 들은 functional annotation을 할 수 있는 정도로 볼 수 있다. 보유 model을 종별로 보면 *H. sapiens*의 경우 약 66%이고, *M. musculus* 64%, *E. coli* 54%, *P. aeruginosa* 64%, *H. pylori* 61%, *M. tuberculosis* 60%, *S. coelicolor* 64% 등임을 알 수 있다. 알고자 하는 gene name이나 Swiss-Protentry name/accession number, TrEMBL entry name/accession number들로 보유 여부를 확인할 수 있다. 그림 11은 보유하고 있는 93종 전체 genome의 protein

Functional distribution of model structures of total 93 species

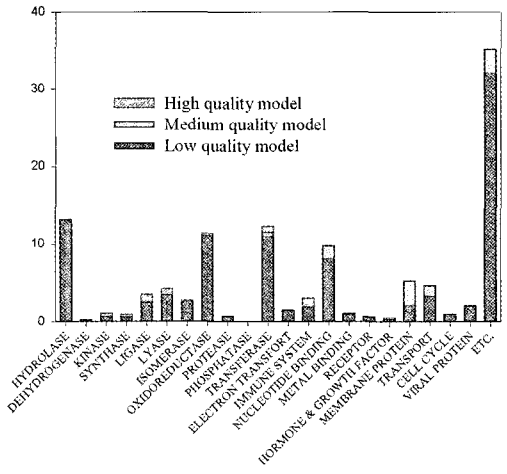


그림 11. 단백질 모델들의 functional distribution

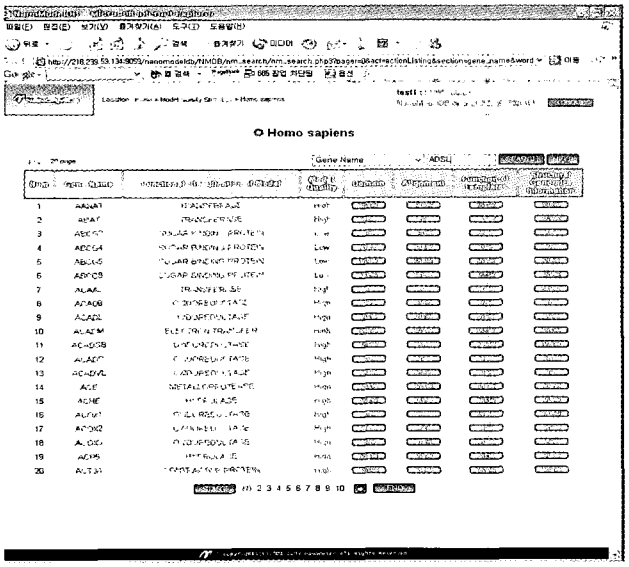


그림 12. Human genes 들의 3D structure 모델 목록

model 들에 대해서 annotation 된 function들을 크게 분류하여 분포를 본 것으로 전체 model로 보았을 때 hydrolase, oxidoreductase, transferase, nucleotide binding 등의 function 들이 큰 분포를 가지며, high quality model 개수로 보았을 때는 ligase, lyase, transferase, immune system, kinase, synthase 등의 분포가 비교적 큰 것을 알 수 있다.

브라우저를 통하여 model structure와 function을 확인하는 과정을 살펴보면, Table 3과 같은 species 별 model의 분포 목록을 보게 되는데 보고자 하는 species를 클릭 하여 선택된 종의 gene들 목록 (그림 12)이 나타나면 search 창에서 gene name, Swiss-Prot, TrEMBL entry name 등을 입력한다. 그림 13에서 gene name "ADSL"을 입력했을 때의 결과를 볼 수 있다. 이 화면에서 볼 수 있는 항목은 function classification, model quality, domain, alignment, function of template, structural genomics information 등이 있다 (그림 14).

마. 항생제 타겟의 선정

필수 유전자(Essential gene)는 세포의 생존에 반드시 필요한 유전자를 말한다. 이들 필수 유전자에 의해 발현

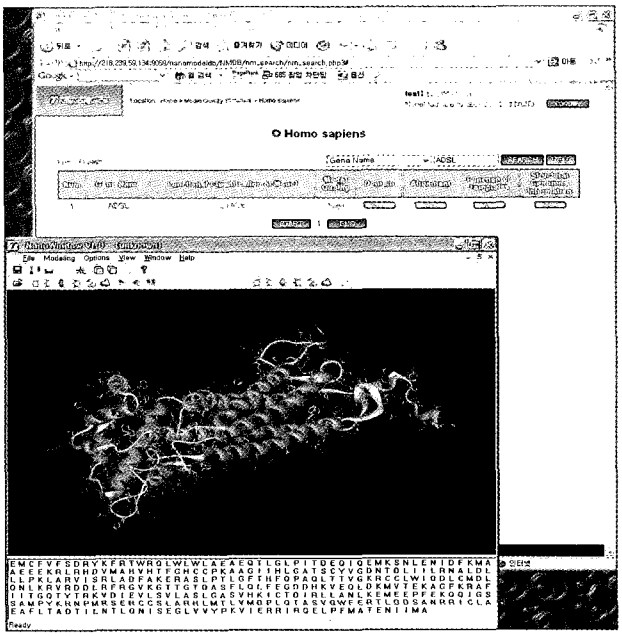


그림 13. Gene name "ADSL"의 functional annotation을 위한 3D 모델 구조의 한 예

되는 단백질들은 물질대사의 기본적인 기능을 수행하거나 외부의 자극으로부터 세포를 지키는 등의 세포 생존에 관여한다. 필수 유전자는 많은 생물학적 궁금증에 해답을 줄 수 있는 실마리를 제공할 뿐만 아니라 여러 가지 용도로 활용할 수 있다. 특히, 새로운 항생제 개발을 위한 타겟으로 주목받고 있다. 필수 유전자를 알아내는 여러 실험적인 방법들이 꾸준히 고안되고 있으며 생물정보학을 활용한 방법들도 제시되고 있다. 생물정보학을 활용한 방법들은 앞에서 제시한 Genome Annotation System들을 기반으로 하여 만들어 지게 된다.

필수 유전자를 알아내기 위한 실험적인 방법들

많은 박테리아 유전체의 염기 서열이 밝혀지고 있으며 전체 유전자를 대상으로 불활성 실험이 가능해지면서 박테리아의 생존에 필수적인 유전자들이 밝혀지고 있다.

필수 유전자를 알아내기 위한 생물정보학적 방법들

이미 알려진 필수 유전자 정보를 활용하거나 Genome

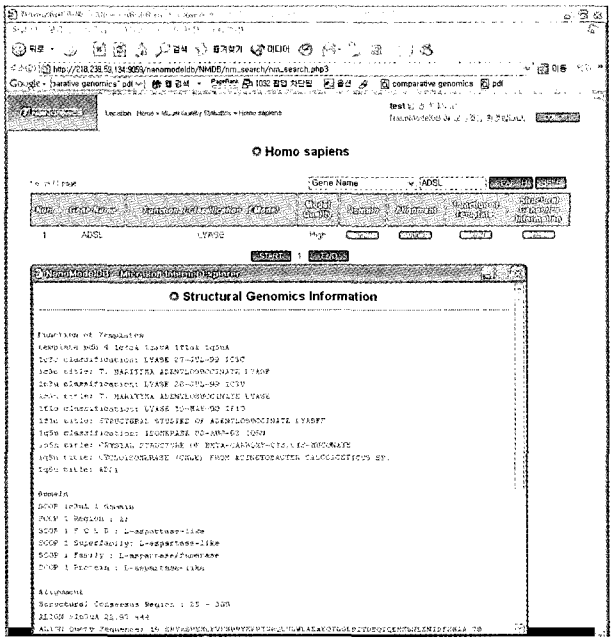


그림 14. 그림13 모델의 functional annotation 내용

Annotation System으로 축적된 많은 생물학적 지식들을 활용하면 직접 실험을 하지 않고도 필수 유전자들을 예측할 수 있다. 또한 실험적인 방법은 실험조건의 제한으로 인하여 모든 유전자에 대해서 필수 여부를 확인할 수 없지만 생물정보학적 방법을 활용하면 실험적으로 찾지 못한 필수 유전자들도 찾을 수 있다.

1. 염기 서열이 밝혀진 유전체의 필수 유전자를 알아내기 위해서 기존에 알려진 필수 유전자 정보들을 활용할 수 있다. 여러 종에서 실험으로 밝혀진 필수 유전자들과 서열 유사성이 있는 유전자들을 찾아서 필수 유전자로 정의한다.
2. 필수 유전자들은 여러 종의 박테리아에서 공통적으로 발견되는 성향을 보인다. 필수 유전자는 세포의 생존에 반드시 필요한 기능을 수행하기 때문에 진화론적으로 유지가 되지만 그렇지 않은 유전자들은 각각의 박테리아가 처한 환경에 따라 진화한다. 따라서 여러 종의 박테리아에서 공통적으로 발견되는 유전자를 필수 유전자로 정의하는 것이 가능하다.
3. 세포 생존에 필수적인 물질대사 경로를 구성하는 모든 유전자들을 항생제 타겟으로 정의한다. 필수 유전자가 포함된 물질대사 경로를 필수 물질대사 경로로 정의하고 이를 구성하는 모든 유전자들을 필수 유전자일 가능성이 있다고 정의한다. Fatty acid biosynthesis, Folic acid biosynthesis, MEP pathway, Peptidoglycan biosynthesis, Shikimate pathway, Lipopolysaccharide biosynthesis, Oxidative phospory-lation, Urea cycle and metabolism of amino groups, Aminosugars metabolism, Pantothenate and CoA biosynthesis, Thiamine metabolism 등을 주로 이용하고 있으며 이들 물질대사 경로들은 Genome Annotation System에 의해서 재구성되고 확장 또는 수정되게 된다. Fatty acid biosynthesis (그림 15)를 구성하는 필수 유전자들을 표 5에서 볼 수 있으며 사람의 유전자와 많이 다르기 때문에 항생제 타겟으로 주목받고 있다.

필수 유전자를 알아내기 위한 실험 방법

박테리아	실험 방법
Staphylococcus aureus	antisense RNA techniqu
Mycoplasma genitalium	transposon mutagenesis
Haemophilus influenzae	high-density transposon mutagenesis
Vibrio cholerae	mariner-based transposon
Saccharomyces cerevisiae	genetic footprinting
Mycoplasma genitalium	transposon mutagenesis
Haemophilus influenzae	transposon mutagenesis

항생제 타겟의 확인 (validation)

1. 항생제 타겟으로 정의된 박테리아의 유전자와 숙주인 인간의 유전자 사이에 유사성이 있어서는 안 된다. 만약 항생제 타겟과 유사성이 있는 유전자가 인간의 세포에 있으면 독성이나 부작용을 일으킬 수 있기 때문이다.
2. 항생제의 오남용으로 인하여 내성균의 출현과 증가 속도가 급등하고 있다. 항생제 내성 문제는 전 세계

◈ 총 설

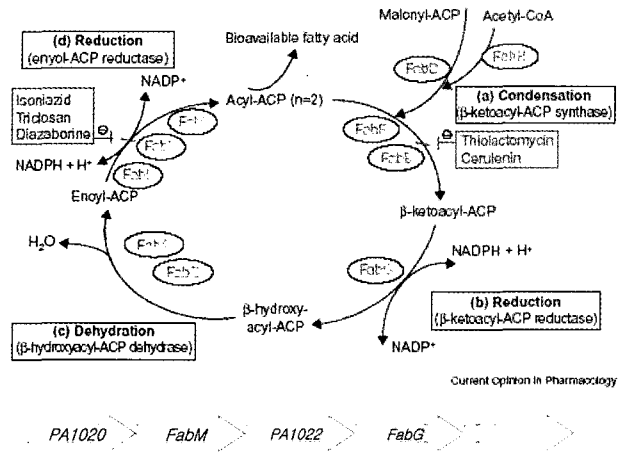


그림 15. Fatty acid biosynthesis pathway and gene cluster

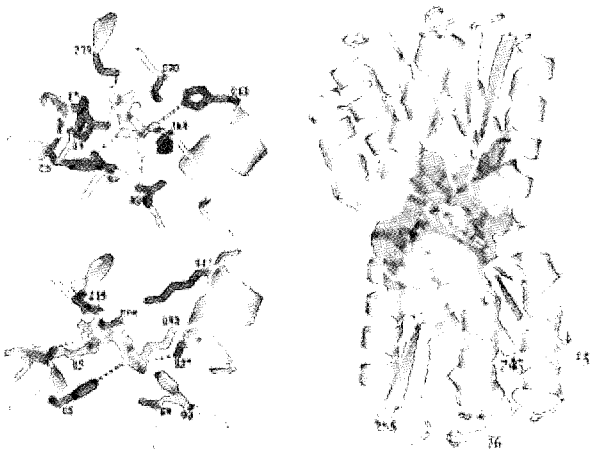


그림 16. RBP (Ribose Bonding Protein)를 변형하여 만들어진 TIM (triose phosphate isomerase)의 활성을 갖는 새로운 단백질

적인 문제이며 새로운 항생제 개발의 근본적인 이유이기도 하다. 항생제 타겟 후보들 중에서 내성의 가능성이 없거나 적은 타겟을 미리 예측하여 항생제 개발 과정의 효율을 증가시킬 수 있다.

3. 여러 종의 박테리아를 대상으로 한 항생제를 만들 것인지, 특정한 박테리아만을 대상으로 한 항생제를 만들 것인지에 따라서 타겟을 정의할 수 있다.

위에서 제시한 Genome Annotation System들을 이용하여 축적된 유전자들의 기능과 세포내의 물질대사 과정 등의 생물학적 지식들을 기반으로 항생제 타겟을 선정하고 확인하게 된다.

바. 단백질 엔지니어링

많은 종류의 항생제, 항암제, 호르몬제, 정밀화학제품 등이 유기합성 방법으로는 합성이 어렵거나 합성에 소요되는 경비가 비싸기 때문에 경제성이 없다. 따라서 이들 물질을 생산하기 위하여 기존의 단백질을 변형하여 새로운 단백질을 만들어내기 위한 연구가 수행되어져 왔다. 그러나 전적으로 실험에만 의존하는 전통적인 단백질공학 방법으로는 다룰 수 있는 변이체 (변형단백질)의 수에 한계가 있으며 막대한 시간과 경비가 소요된다. 최근 이러한 단점을 극복하고 원하는 특성을 갖는 단백질을 자유롭게 만들어내기 위하여 컴퓨터상에서 (*in silico*) 단백질의 3차 구조 정보를 사용하여 단백질 엔지니어링을 수행하기 위한 연구가 진행 중에 있다. 즉 가능한 모든 경우의 변형된 단백질을 만들어내고 scoring function을 이용하여 활성유무를 예측하며 실험적인 방법으로 확인하게 된다. 이러한 개발 과정의 대부분을 컴퓨터상에서 수행하게 되므로 시간 및 비용을 획기적으로 단축시킬 수 있을 것으로 기대된다. 그림 16은 컴퓨터를 이용한 단백질 엔지니어링의 한 예로서 RBP (Ribose Bonding Protein)를 변형하여 TIM (triose phosphate isomerase)의 활성을 갖는 새로운 단백질을 만들어낼 수 있음을 보여 주고 있다 (42).

결 론

지금까지 보아온 것처럼 genome annotation system은 이제 functional network에 기반한 annotation을 수행하고 있으며 이러한 효과는 통계적인 데이터로부터 알 수 있었다. 즉, 일반적인 sequence homology를 사용할 경우 약 40% 정도의 annotation이 가능하나 comparative genomics를 추가하면 species에 따라 10-

Table 4. Triclosan의 FabI와 Fabk에 대한 Sensitivity데이터(41)

Organism	Sequence identity (%)*		Triclosan MIC+
	FabI	Fabk	(microgran/ml)
Escherichia coli	100	–	0.25
Staphylococcus aureus	43	–	0.01
Salmonella pneumoniae	–	100	2.0
Clostridium acetobutylicum	–	58	3.0
Enterococcus Faecalis	47	68	10
Mycobacterium tuberculosis	33	31	100
Pseudomonas aeruginosa	69	33	>1,000

* Identity to FabI and Fabk is shown.

\+ Minimum inhibitory concentration (MIC) value

20% 정도 증가하고, protein interaction data를 더하여 활용하면 역시 10-20% 정도 증가하며, 여기에 protein 3D model을 더 추가하여 사용하면 약 5-10%의 증가 효과를 가져 올 뿐만 아니라, 이미 다른 방법으로 annotation된 데이터들에 대한 재확인의 효과를 갖는다. 현재 앞에서 살펴본 많은 website들에서처럼 충분히 좋은 genome annotation 정보들이 일반에게 제공되기 때문에, 이제는 biosynthesis용 gene cluster, target gene finding 등 사용자의 특별한 목적에 맞도록 제작된 specified genome annotation system들의 필요성이 부각되고 있다. 이러한 시스템을 이용하면 원하는 기능을 수행하는 유전자들, 즉 functional network을 찾고자 할 때, 한 개의 구성원이라도 알고 있을 경우 이 유전자가 속해 있는 gene cluster들과 이 cluster와 network을 이루는 cluster들을 확인하여 목적하는 기능을 구현할 수 있을 것이다.

한 예로 Table 4는 *E. coli*의 fatty acid biosynthesis 과정 중 한 단계를 담당하는 essential gene인 *fabI*을 목표로 하여 개발된 Triclosan의 sensitivity 데이터를 보여 주는데, *E. coli*와 *S. aureus*의 경우 좋은 효과를 보이는 반면, 표에 있는 다른 species들에서는 효과를 보지 못했다. Triclosan 개발 당시에는 그 이유를 알지 못하다가 나중에 알려지게 되었지만 다른 species들 경우에는 *fabI* gene이 essentiality를 유지하지 못하고 있음을 볼 수 있게 된다. 그림 18에서 보면 *fabK*가 *fabI*의 molecular function을 대신하며, 이러한 *fabK*를 *S. pneumoniae*를 비롯한 다른 species들이 갖고 있음을 확인하게 된다. 그리하여 Table 4의 결과들에 대한 설명이 가능하게 되었는데, 만일 위와 같은 genome annotation system에서 미리 확인해 보았다면 표의 결과들을 먼저 예상할 수 있었을 것이다. 실지로 이 system을 사용하여 확인해 본 결과 *S. pneumoniae*는 *fabK*, *fabM*, *fabZ*, *fabF* 들이 한 개의 cluster를 구성하여 *E. coli*의 fatty acid biosynthesis 과정 중 그림 18의 단계에 대응되는 기능을 수행한다는 것을 예상하게 해 주었다. 더욱이 *fabI*와 *fabK* gene들은 sequence homology를 사용해서는 유사 구조나 기능을 확인할 수 없었으나, 3D model을 비교하였을 때는 그림 17에서 볼 수 있는 것처럼 functional domain structure가 매우 근접하여 같은 기능을 함을 알 수 있었으며, *fabI*를 목표로 개발 되었던 Triclosan이 결합하는 부위를 살펴봄으로써 왜 같은 기능을 하는 *fabK*에는 결합할 수 없는지를 구조적으로 확인할 수 있었다. 따라서 *S. pneumoniae*의 *fabK*가 소속된 gene cluster가 fatty acid biosynthesis를 위한 functional network의 한 구성원이라는 것을 protein interaction data를 이용함으로써 알 수 있었으며, *fabK*와 *fabI*가 같은 기능을 한다는 사실은 3D model을 비교하여 functional structure의 유사성을 확인함으로써 알게 되었고, 두 gene들의 target으로서의 가능성을 가늠할 수 있게 되었다.

또 다른 예로 그림 19에서 *M. tuberculosis*의 fatty acid elongation system 과정의 일부를 볼 수 있는데, 일반적으로 알려진 같은 biosynthesis 과정을 비교해 놓았다. 여기서 점선으로 betahydroxyacyl-ACP를 Enoyl-ACP로 dehydration하는 기능을 갖는 아직 알려지지 않은 dehydratase를 표시하고 있는데, *fabA*와 *fabZ*에 해당된다. 이 빈 공간을 어떻게 채울 수 있을까라는 질문이 나오게 된다. 우선 알고 있는 *fabD*, *fabG*, *fabH*,

◈ 총 설

Table 5. essential[a] and non-essential genes in fatty acid biosynthesis pathway

FabH EC: 2.3.1.41	FabF EC:2.3.1.41	FabB[a] EC:2.3.1.41	FabG EC:1.1.1.100	FabA[a] EC:4.2.1.60	FabZ[a] EC:4.2.1-	FabI, FabK EC:1.3.1.9	FabD EC:2.3.1.39
PA0999	PA1373	PA1609	PA1023	PA1610	PA3645	PA1806	PA0214
PA3333	PA2965		PA1344			PA1024	PA2968
			PA1470				
			PA2967				
			PA3387				
			PA3507				
			PA3511				
			PA4079				
			PA4148				
			PA4389				
			PA5524				

그림 17. FabI 와 FabK 의 functional domain structure 들간의 구조 비교: 진한색 리본, FabI ; 연한색 리본, FabK

kasA, *kasB*들을 사용하여 이들을 구성원으로 하는 gene cluster들을 확인하고 functional network을 구성한 후, comparative genomics를 수행하면 대응되는 다른 species들의 cluster들을 찾게 된다. 이렇게 얻어진 cluster들의 구성원들에 대한 annotation을 확인하고 functional domain structure를 FabA, FabZ와 비교하게 되면 유사한 구조를 갖는 MtxA를 찾게 되는데 구조 비교한 것을 그림 20에서 볼 수 있다. MtxA는 isomerase와 dehydratase로 annotation 되어 있어서 *M. tuberculosis*의 fatty acid elongation system의 빈 공간인 dehyration

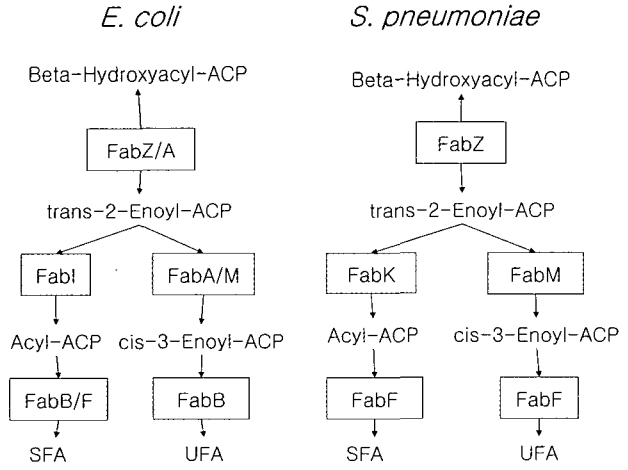


그림 18. *FabI*와 *FabK*의 역할 비교

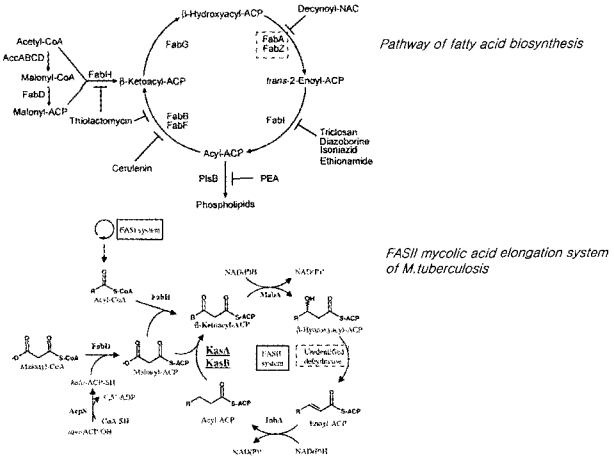


그림 19. *M. tuberculosis pathway*의 빈 공간

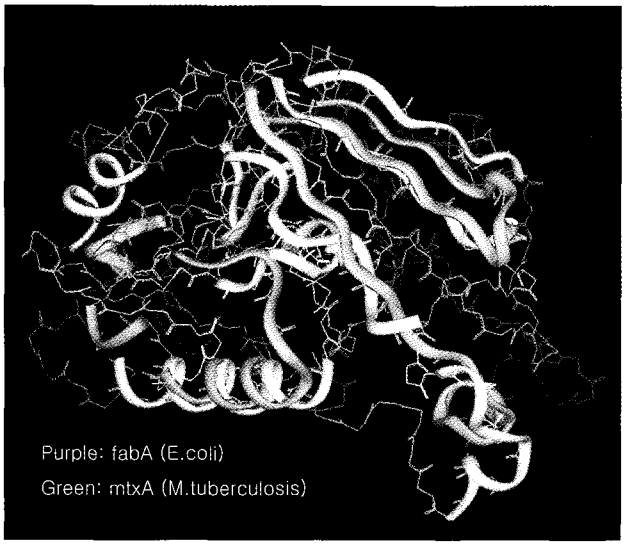


그림 20. FabA와 같은 functional domain structure를 갖는 mtxA 유전자의 구조비교: 진한색 리본, FabA ; 연한색 리본, mtxA

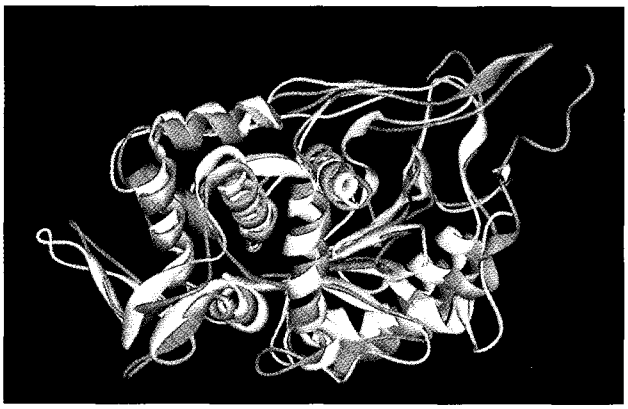


그림 21. 항생제 타겟 PBP (penicillin-binding protein : light blue)와 항생제 내성 유전자 PBP2a (methicillin resistance protein : grey)의 3차원 구조 비교

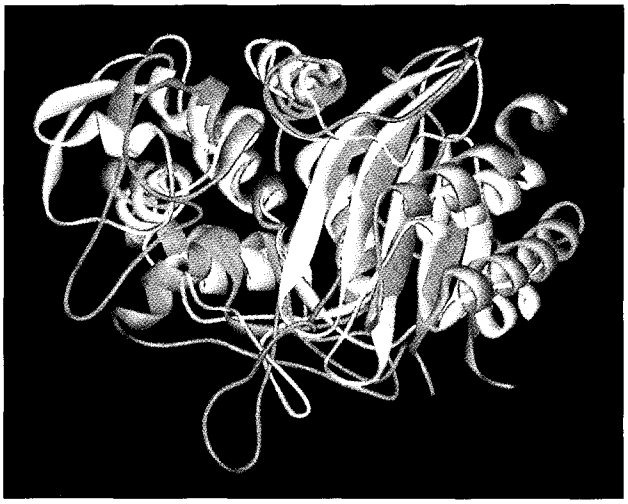


그림 22. 항생제 타겟 PBP (penicillin-binding protein : light blue)와 항생제 비활성화 효소 BlaZ (β-lactamase : grey)의 3차원 구조 비교

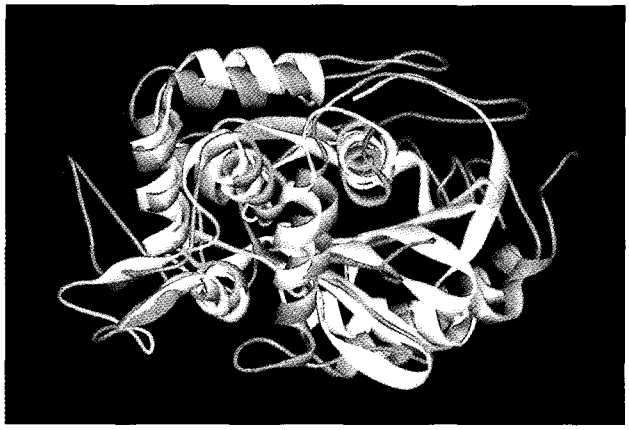


그림 23. 항생제 타겟 PBP (penicillin-binding protein : light blue)와 항생제 인지 관련 유전자 BlaR1 (senor protein : grey)의 3차원 구조 비교

단계에서 fabA의 역할을 수행할 가능성이 있음을 추정하게 해줄 뿐만 아니라, 그 다음 단계인 trans-2-enoyl-ACP를 cis-3-enoyl-ACP로 바꾸어 주는 isomerase 역할의 가능성도 열어 두고 있다. 실지로 *E. coli*에서는 FabA가 다음 단계인 isomerase 기능도 FabM과 함께 수행하고 있다.

마지막으로 한 가지 예를 더 소개하고 이 글을 마치고자 한다. 우리나라에서 '슈퍼 박테리아' 로 알려져 있는 MRSA (Methicillin-Resistance *S. aureus*)는 항생제에 강한 저항성을 가지고 있어서 감염되어 특정 병이 생기면 웬만한 항생제를 먹어도 쉽게 치료가 되지 않는다. MRSA의 내성은 주로 항생제를 비활성화 시키는 효소(β-lactamase)나 항생제에 영향을 받지 않는 유전자 때문이라고 알려져 있다. MRSA는 항생제를 인지하는 유전자를 가지고 있어서 항생제가 들어오면 신호

◈ 총 설

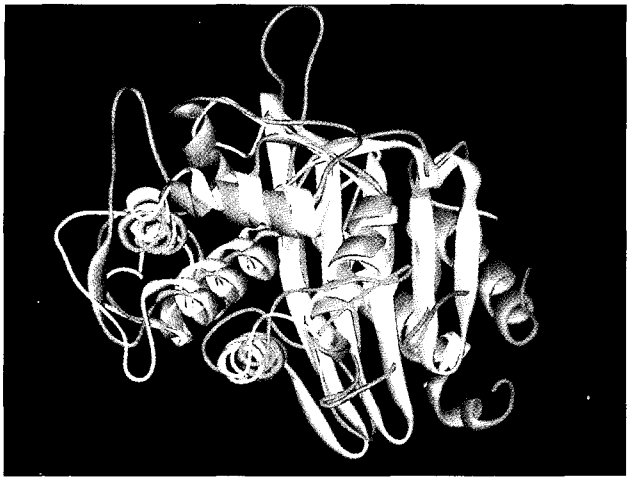


그림 24. 항생제 타겟 PBP (penicillin-binding protein : light blue)와 항생제 인지 관련 유전자 MecR1(senor protein : grey)의 3차원 구조 비교

전달 과정을 거쳐서 내성 관련 유전자들을 발현하게 된다. 그림 21~24는 페니실린의 타겟인 PBP(penicillin-binding protein)과 항생제 내성 관련 유전자들의 3차원 구조를 비교한 그림들이다. 항생제 타겟은 기능상으로 내성 관련 유전자들과 연관성이 있다. 세포내에서의 기능이나 항생제와 같은 화합물과의 결합 등에서 유사성을 가지고 있다. 이들 유사성은 유전자의 염기 서열이나 3차원 구조의 유사성을 이용하여 확인하거나 예측할 수 있다.

본 총설은 저자들이 "분자세포생물학뉴스지"("Trends in Bioinformatics for Genome Annotation", 분자세포생물학뉴스, 2004년 3월호 16권 1호, pp. 20-33)에 출판하였던 내용을 기본으로 하여 최근의 몇 가지 응용 예와 항목을 더 추가하여 작성하였음을 알려드립니다.

Reference

1. Collins F.S., A. Patrinos, E. Jordan, *et al.* 1998. *Science* 282, 682-689.
2. GenBank, National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov
3. National Human Genome Research Institute (NHGRI) http://www.genome. gov
4. Online Mendelian Inheritance in Man (OMIM) http://www.ncbi.nlm.nih.gov /omim
5. Entrez Website, National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov/Entrez
6. University of California at Santa Cruz (UCSC) Genome Browser http://genome.ucsc.edu
7. The Institute for Genome Research http://www.tigr.org/tdb/tgi
8. Baxevanis A.D. 2002. *Nucleic Acids Res.* 30, 1-12.
9. Hamosh A., A.F. Scott, *et al.* 2002. *Nucleic Acids Res.* 30, 52-55.
10. National Center for Biotechnology Information (NCBI) website http://www.ncbi.nlm.nih.gov
11. FTP site for OMIM http://ncbi.nih.gov/repository/OMIM
12. Wheeler D.L., *et al.* 2004. *Nucleic Acids Res.* 32, D35-D40.
13. Kent W.J., C.W..Sugnet, *et al.* 2002. *Genome Res.* 12, 996-1006.
14. The UCSC Genome Browser User's Guide http://genome.ucsc.edu/ goldenPath/help/hgTracksHelp.html
15. Altschul S.E., *et al.* 1990. *J. Mol. Biol.* 215, 403-410.
16. Altschul S.E., *et al.* 1997. *Nucleic Acids Res.* 25, 3389-3402.
17. Peterson J.D., *et al.* 2001. *Nucleic Acids Res.* 29, 123-125.
18. http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
19. http://www.tigr.org/tdb/euk
20. http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html
21. Zhang Z., *et al.* 2000. *J. Comput. Biol.* 7, 203-214.
22. DIP, Database of Interacting Proteins http://dip.doe-mbi.ucla.edu
23. BIND, Biomolecular Interaction Network Database http://www.bind.ca
24. MIPS, Munich Information Center for Protein Sequences http://www.mips.biochem.mpg.de
25. RNGAS, Regulatory Network based Genome Annotation System http://www.nanormics.com
26. Zheng Y. 2002 *Genome Res.* 12, 1221-1230.
27. Stojilkovic I. 1995. *J. Bacteriol.* 177, 1357-1366.
28. Schneider K. 2002. *J. Bacteriol.* 184, 2439-2446.
29. Jubier-Maurin V. 2001. *J. Bacteriol.* 183, 426-434.
30. Yebra M.J. 2000. *J. Bacteriol.* 182, 155-163.
31. Hase C.C. 1999. *Proc Natl Acad Sci* USA 96, 3183-3187.
32. Yamasaki S. 1999. *Gene* 237, 321-332.
33. Oshima T., *et al.* 2002. *Mol. Microbiol.* 46, 281-291.
34. Schwinde J.W. 1993. *J. Bacteriol.* 175, 3905-3908.
35. Kronemeyer W. 1995. *J. Bacteriol.* 177, 1152-1158.
36. Itoh Y. 1997. *J. Bacteriol.* 179, 7280-7290.
37. Parker G. 2001. *Microbiol.* 147, 2553-2560.
38. Baitsch D. 2001. *J. Bacteriol.* 183, 5262-5267.
39. Egland P.G. 1999. *J. Bacteriol.* 181, 2102-2109.
40. NanoModel Database System http://www.nanormics.com
41. Heath R.J. and C.O. Rock. 2000. *Nature* 406, 145-146.
42. Dwyer M.A. 2004. *Science* 304, 1967-1771.

약 력

한 원 석

- 1997. 2 고려대 화학과 (학사)
- 1999. 2 고려대 대학원 화학과 (석사)
- 2001. 8-현재 (주)나노믹스 근무

윤 창 노

- 1980. 2 연세대 화학과 (학사)
- 1982. 2 한국과학기술원 화학과 (석사)
- 1985. 2 한국과학기술원 화학과 (박사)
- 1985. 2-1997. 2 한국과학기술연구원 도핑콘트롤센터
- 1997. 3-현재 한국과학기술연구원 생체대사연구센터