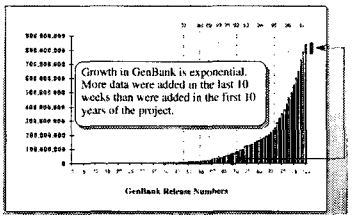
Bioinformatics와 Functional Genomics

김 승 목

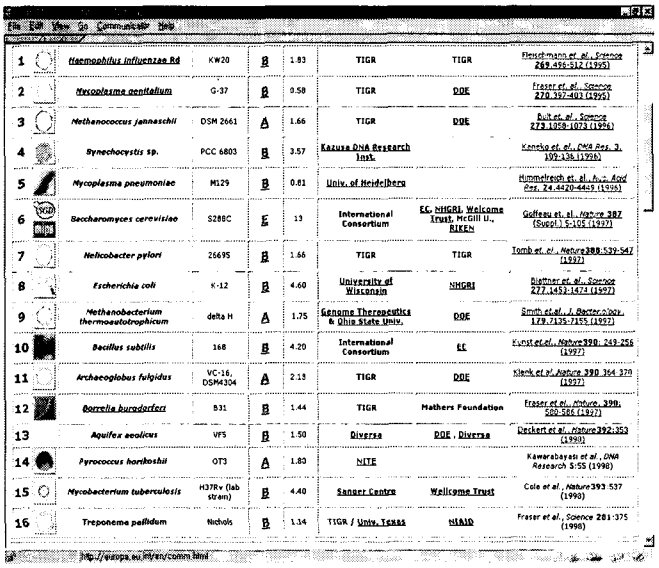
생명공학연구소 유전체사업단 유전체정보팀

1. Bioinformatics 시작하기

가 Bioinformatics의 기원?

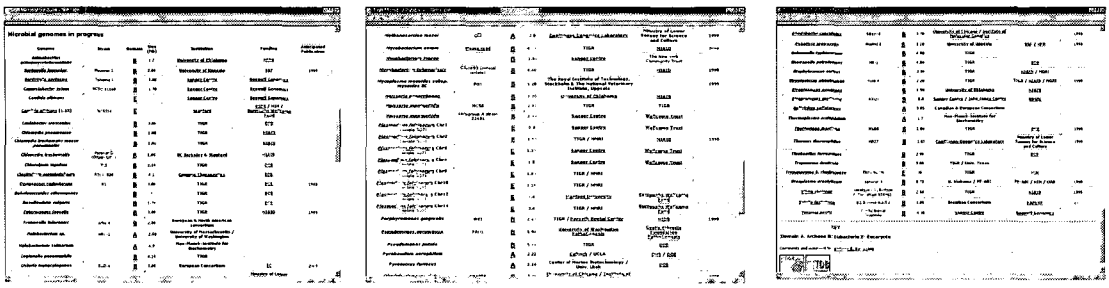


Bioinformatics는 우리말로 번역하여 생물정보학이라 칭하기도 한다. Bioinformatics의 아마도 Mendel이 유전법칙을 발견한 것에 기원을 두는 것이라 할 만하나 Bioinformatics라는 말이 생겨난 것은 1980년대 후반의 일이다. Bioinformatics에 대한 정의를 내리기에는 10여 년이 지난 지금도 아주 어려운 일이나 Bioinformatics라는 분야가 생겨나게 된 배경을 살펴본다면 어는 정도는 짐작이 가리라 생각한다. Bioinformatics 분야가 생겨난 시기는 미국을 중심으로 하여 Human Genome Project에 대한 논의가 한참 진행되고 있던 시기이며 컴퓨터와 Internet의 보급이 활발히 진행되던 시기와 일치하고 있다. 이를 바탕으로 Bioinformatics를 간략히 설명해 보면 Computer와 수학적인 방법을 이용해 실험적으로 얻어지는 다양한 생명현상과 관련된 사실을 분석, 종합하여 일반적인 생명법칙을 찾아내고자 하는 것이라 요약할 수 있다. Genome Project가 시작된 이후 분자생물학의 급속한 발전으로 인해 DNA나 단백질의 서열생산량은 증가 속도가 10개월에 2배 씩 증가하고 있는 추세이며, 대량으로 밝혀진 서열정보의 분석과 응용은 컴퓨터가 분자생물학자들에게 Pipet과 같은 실험 도구로 사용되는 시대가 오도록 하였다.



1	*Haemophilus influenzae* Rd	KW20	B	1.83	TIGR	TIGR	Fleischmann et al., Science 269:496-512 (1995)
2	*Mycoplasma genitalium*	G-37	B	0.58	TIGR	DOE	Fraser et al., Science 270:397-403 (1995)
3	*Methanococcus jannaschii*	DSM 2661	A	1.66	TIGR	DOE	Bult et al., Science 273:1058-1073 (1996)
4	*Synechocystis* sp.	PCC 6803	B	3.57	Kazusa DNA Research Inst.		Kaneko et al., DNA Res. 3, 109-136 (1996)
5	*Mycoplasma pneumoniae*	M129	B	0.81	Univ. of Heidelberg		Himmelreich et al., Nuc. Acid Res. 24:4420-4449 (1996)
6	*Saccharomyces cerevisiae*	S288C	E	13	International Consortium	EC, NHGRI, Wellcome Trust, McGill U., RIKEN	Goffeau et al., Nature 387 (Suppl.) 5-105 (1997)
7	*Helicobacter pylori*	26695	B	1.66	TIGR	TIGR	Tomb et al., Nature 388:539-547 (1997)
8	*Escherichia coli*	K-12	B	4.60	University of Wisconsin	NHGRI	Blattner et al., Science 277:1453-1474 (1997)
9	*Methanobacterium thermoautotrophicum*	delta H	A	1.75	Genome Therapeutics & Ohio State Univ.	DOE	Smith et al., J. Bacteriology 179:7135-7155 (1997)
10	*Bacillus subtilis*	168	B	4.20	International Consortium	EC	Kunst et al., Nature 390: 249-256 (1997)
11	*Archaeoglobus fulgidus*	VC-16, DSM4304	A	2.18	TIGR	DOE	Klenk et al. Nature 390 364 370 (1997)
12	*Borrelia burgdorferi*	B31	B	1.44	TIGR	Mathers Foundation	Fraser et al., Nature, 390: 580-586 (1997)
13	*Aquifex aeolicus*	VF5	B	1.50	Diversa	DOE, Diversa	Deckert et al., Nature 392:353 (1998)
14	*Pyrococcus horikoshii*	OT3	A	1.80	NITE		Kawarabayasi et al., DNA Research 5:55 (1998)
15	*Mycobacterium tuberculosis*	H37Rv (lab strain)	B	4.40	Sanger Centre	Wellcome Trust	Cole et al., Nature 393 537 (1998)
16	*Treponema pallidum*	Nichols	B	1.14	TIGR / Univ. Texas	NIAID	Fraser et al., Science 281:375 (1998)

현재 Bioinformatics의 연구 방향은 크게 두 가지를 들 수 있다. 첫째는 Genome Project의 결과 산물인 서열 정보를 어떻게 효과적으로 관리하는 가이고 둘째는 이들 서열 정보를 이용하여 효과적인 연구방향의 정립과 생명현상을 설명하기 위한 법칙을 발견하기 위한 노력이다. Genome Project가 전 세계적으로 8년간 진행된 지금 이미 16종의 생물에 대한 전체 염기서열이 얻어진 상태이고, 약 70종의 생물에 대한 염기서열 규명이 진행되고 있다. Bioinformatics는 이를 바탕으로 Gene Prediction, Gene Family, Phylogenetic Relationship, Protein Function 등에 관한 연구를 진행하기 위해 새로운 전산학, 수학, 물리학의 이론과 기술을 접목하기 위한 시도를 활발히 진행하고 있다.



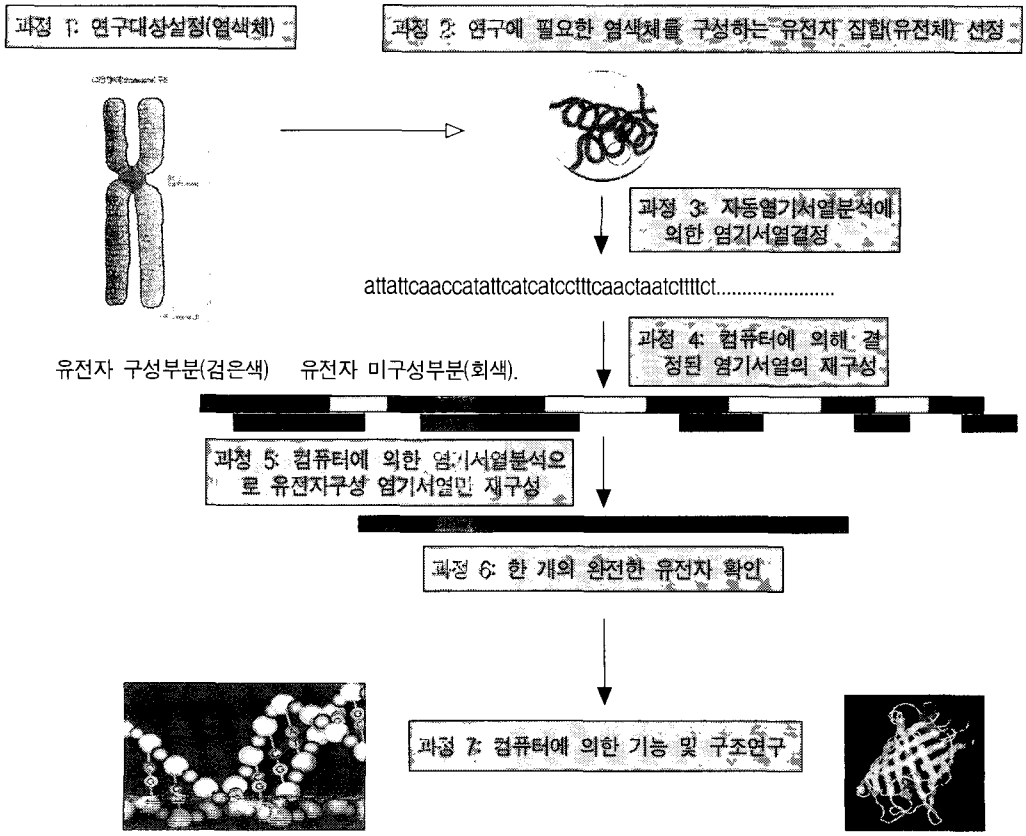
나. 유전체(Genome)연구란 무엇인가?

○ 유전체(Genome)는 생명체의 근원인 유전자(Gene)들의 집합체이다.

○ 이를 특히 유전체(Genome)이라 하는 이유는 각각의 유전자들이 무의미한 집합체를 이루는 것이 아니라 아직은 인간이 해석할 수 없는 유기적인 연관 관계를 갖는 집합체를 이루기 때문이다.

○ 유전체연구(Genome Project)의 궁극적인 목적은 유전자의 상관관계를 그 집합체인 유전체내에서 종합적으로 해석하기 위한 연구이다.

다. 유전체연구의 방법



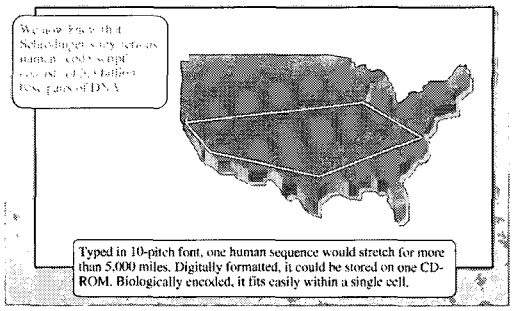
라. 기존 분자생물학 연구방법과 차이는?

ㅇ 연구의 재료나 실험방법의 기초기술에서는 동일하다.
ㅇ 유전체연구와 기존의 연구의 차이는 접근방법에 있다.
ㅇ 기존의 분자생물학 연구는 앞의 그림의 과정6부터 과정1로 을 거꾸로 수행한다고 할 수 있다.
ㅇ 즉 이미 유전자에 대한 정보를 알고 있는 상태에서 생물학적 기원과 작용에 대한 연구를 수행하나 유전체연구는 유전자의 기원인 염색체에서 대규모 염기서열분석을 통해 지금까지 알지 못하던 유전자를 찾기 위한 것이다.
ㅇ 또한 한 개의 유전자가 대상이 아니므로 앞 그림에서 과정3 이하는 유전체의 부분별로 수많은 연구진에 의해 병렬로 진행된다.
ㅇ 유전체연구의 성공여부는 동시에 얻어지는 많은 연구 결과를 얼마나 효율적으로 빠르게 검증하고 재구성하는가 하는 것이다.

마. Genome Research와 Bioinformatics

ㅇ 유전체연구에서 컴퓨터의 기본적인 용도는 대규모 염기서열 결정 후 유전자를 재구성하고 유전자를 찾고자 하는 데에 있으나(앞의 그림에서 과정4와 과정5)
ㅇ 궁극적인 컴퓨터를 이용한 유전체정보연구는 유전체연구로 유전자가 밝혀진 이후로 유전자의 기능과 구조를 알아 각각의 유전자의 상관관계를 파악하는 데에 있으며 이 과정에서 초고속 고용량의 컴퓨터를 필요로 한다(과정6과 7).
ㅇ 즉 "유전체(Genome)란 우리가 전혀 알지 못하는 언어로 쓰여진 아주 정교한 생물학적인 프로그램이다"라는 것이다. 유전체(Genome)의 정보화된 의미는 단지 4개의 문자인 A, C, G, T로 표현되는 다양한 문자열(String)인 Nucleotide의 집합체라고 할 수 있으며 고등생물의 경우 전체 크기가 수 십억 문자열, 다시 말하면 수십만 Gigabytes로 이루어진 아주 정교한 프로그램인 것이다
ㅇ 유전체(Genome)와 유전자(Gene)의 관계를 컴퓨터에 비유하여 설명하자면 유전체(Genome)라는 하나의 거대한 프로그램은 유전자(Gene)이라는 작은 부프로그램(Subprogram)들의 집합체로 구성되어 있으며 이들 부프로그램(유전자)들은 보통의 컴퓨터 프로그램들과 마찬가지로 많은 매개변수와 통신망을 이용해 상호정보교환을 하며 계획된 정확한 결과를 내는 것이며 이것이 바로 생명현상이다.
ㅇ 현재 세계각국에서 진행중인 유전체연구(Genome Project)의 궁극적인 목적은 예기한다면 "유전체를 구성하는 각 유전자의 모든 염기서열을 규명하며 이들 염기서열에 대한 완벽한 기능을 확인(Annotation) 하는 것이다"라 할 수 있으며 유전체연구가 종결되었을 때 유전체(Genome)를 구성하는 모든 유전체의 각각의 기능과 작용에 대한 확실한 규명이 이루어짐을 뜻하고, 이런 의미에서 유전체연구는 마치 지금까지 인간이 본적도 사용해 본적도 없는 프로그램을 분석하여 그 프로그램을 완벽히 이해하는 것과 다름없다.
ㅇ 1980년대 후반 미국을 중심으로 인체유전체연구(Human Genome Project)를 계획하고 실제로 1990년부터 인체유전체연구에 돌입할 초기만 해도 많은 연구자들이 유전체의 방대함에 따른 유전체연구의 일차목표인 인체 유전체의 염기서열을 계획한 기간(2010년)까지 완료하는 것은 불가능할 것으로 생각했으며 더욱이 유전체를 연구하는데 과연 컴퓨터를 이용할 필요가 있는가? 또한 이용하더라도 과연 무슨 정보를 얻을 것인가에 대해서는 상당히 회의적이었다. 유전체연구의 초기에는 컴퓨터는 유전체연구에서 나오는 염기서열정보를 단순히 컴퓨터에 효율적으로 보관하는 데이터베이스 정도를 이용하는 것으로 생각했던 것이다.
ㅇ 그러나 미국을 중심으로 한 인체유전체연구의 일차연구를 종료한 1995년의 상황은 연구를 시작했어 던 5년 전과는 전혀 다른 상황이 벌어지고 있었다. 2010년까지 완료하기에도 불가능하리라 생각되던 인체의 모든 유전체의 염기서열 결정이 일차 년도 5년간의 염기서열 결정속도의 증가추세가 계속 유지될 경우 2010년까지 최대 7십억Base (인체유전체의 2배)의 염기서열결정이 가능할 것이란 예측을 하게되었으며 이 결과 인체유전체의 염기서열결정의 완료시

기도 2005년으로 5년이 앞당겨지게 되었다. 이러한 계획의 수정은 인체유전체연구를 계획하던 1980년대 말의 염기서열결정속도에 현재의 속도가 15배 이상 증가했으며 이 속도의 증가는 기술발달에 따라 가속되게 되기 때문이다.

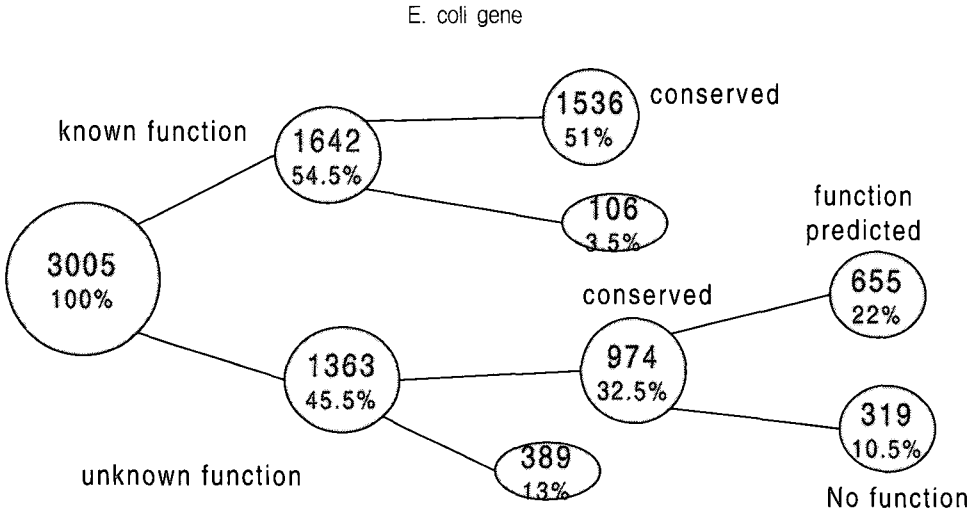


year	per base cost	budget	year	cumulative	percent completed
1995	$0.50	16,000,000	10,774,411	10,774,411	0.33%
1996	$0.40	25,000,000	23,843,771	34,618,182	0.96%
1997	$0.30	35,000,000	39,281,706	73,899,888	2.15%
1998	$0.20	50,000,000	84,175,084	158,074,972	4.71%
1999	$0.15	75,000,000	165,350,188	323,425,160	9.81%
2000	$0.10	100,000,000	336,700,337	660,325,497	20.01%
2001	$0.05	100,000,000	673,400,673	1,333,726,150	40.42%
2002	$0.05	100,000,000	673,400,673	2,007,126,824	60.82%
2003	$0.05	100,000,000	673,400,673	2,680,527,497	81.23%
2004	$0.05	100,000,000	673,400,673	3,353,928,171	101.63%

o 더욱이 초기 유전체연구에서의 컴퓨터이용의 주된 목적이었던 데이터베이스의 구축도 예상치 못했던 자료의 증가추세**(1997년 8월 현재 GenBank의 자료 수는 1,053,000,000 bases, 1,611,000 sequence, 1996년 10월 약 652,000,000 base, 1,021,000 sequence)**로 인해 데이터베이스의 구성과 관리, 그리고 전세계에 걸친0 염기서열 데이터베이스 이외의 다양한 데이터베이스(Mapping Database, Gene의 Function과 관련된 Database, 또는 인체유전자 이외의 다른 종의 Database 등)들과의 연계방법 등이 유전체연구에서 컴퓨터이용의 중요성을 더욱 증대시키고 있으며 현 기술로는 유전체연구에 병목현상을 초래하리라 예상하고 있다.

o 2005년의 인체유전체의 염기서열결정을 완료할 것으로 예상되는 현 시점 에서 유전체연구를 수행하는 많은 연구자들이 많은 관심과 우려를 하는 것은 현재의 컴퓨터이용 기술로 그 방대한 연구결과를 분석하고 정리하여 유전체연구의 결과를 의료, 산업 등의 응용분야에 적절히 활용할 수 있는가 하는 것이다.

o 실 예로 프랑스를 중심으로 수행된 이스트(Yeast) 유전체의 염기서열 결정은 1996년에 완료되었으나 그 유전체를 구성하는 유전자가 예상했던 것보다도 훨씬 많을 뿐 아니라 유전자를 찾기는 했으나 대부분의 유전자의 기능에 대해서는 아직 알지를 못하고 있으며 컴퓨터를 이용한 정보처리기술로 이를 해결하기 위한 방법을 찾는 것이 일 단계 유전체연구 이후에 우선적으로 해야할 일로 인식되고 있으며 이를 "Paradigm Shift"라고 표현하고 있다.



사. 서열검색에 쓰이는 대표적인 프로그램

Blast (Basic Local Alignment Search Tool) :
blastn - nucleotide database is compared with both strands of a nucleotide query
blastp - protein database is compared with a protein query
blastx - nucleotide query is translated in six frames and compared to protein database
tblastn - protein query is compared to a nucleotide database, translated in six frames
tblastx - nucleotide query, translated in six frames, is compared to a nucleotide database, translated in six frames

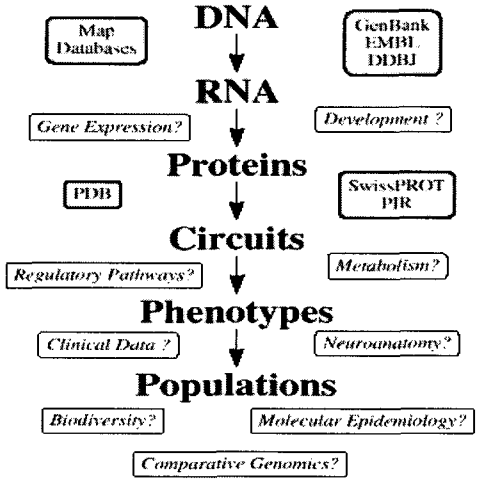
BLAST_ORG : BLAST report parser filtered by organism
ALUBLAST : input filter for repetitive regions such as ALU and MER sequences
CAP (Consistent Alignment Parser) : produce alignment blocks
BLANCE : BLAST report summary

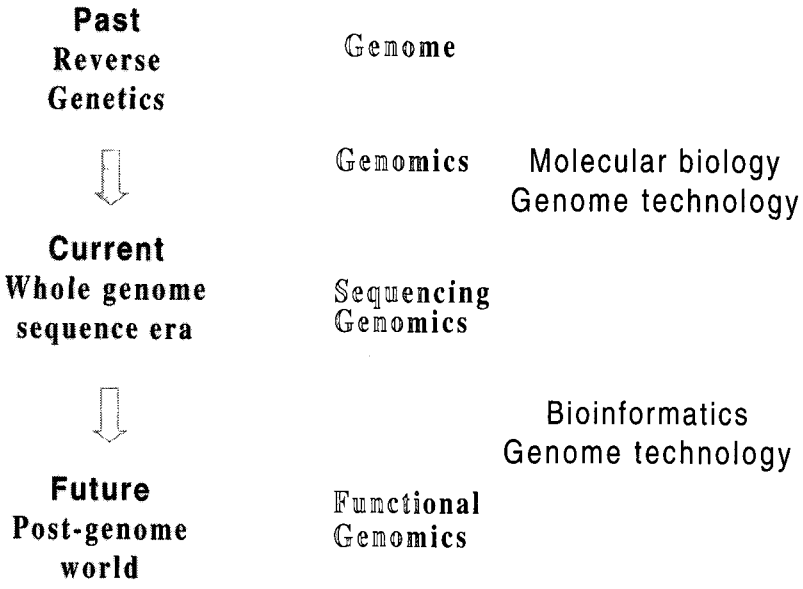
FASTA :
fasta - compare a protein sequence to a protein sequence database or a DNA sequence to a DNA sequence database using FASTA
ssearch - compare a protein sequence to a protein sequence database using Smith-Waterman
tfasta - compare a protein sequence to a DNA sequence database, translating each DNA database sequence in all six frames

2. 유전체연구의 세계적 추세

가. Functional Genomics 란?

Human Genome Sequence는 약 30억 Base (3GBytes) 정도로 알려져 있으며, 염기서열 결정이 완료되는 시점을 2003년으로 계획하고 있다 (TIGR의 Crag Venter는 2년 안에 끝내겠다고 공언하기도 함). 인체의 Genome 염기서열이 결정되면 이로부터 우리가 얻을 수 있는 유전자(Gene)는 75,000 ~ 100,000 정도로 추정되며 이는 전체 염기서열의 5% 이하일 것으로 추정되고 있다. 바로 이 것이 Genome 연구가 염기서열 결정으로 끝나는 것이 아니라 바로 새로운 Genome 연구의 시작이 되는 이유이며, 이를 Post-Genome Project라고 부르고 있다. Genome 연구는 바로 Genome을 이루고 있는 유전자들의 상관관계를 이해하여 생명현상을 규명하자는 것으로 것으로부터 출발하였고 바로 지금 이 순간이 진정한 Genome Project를 시작하는 새로운 출발점이라 할 수 있으며 Functional Genomics의 시작이다.

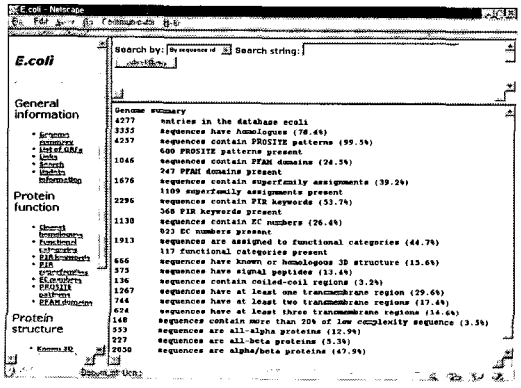


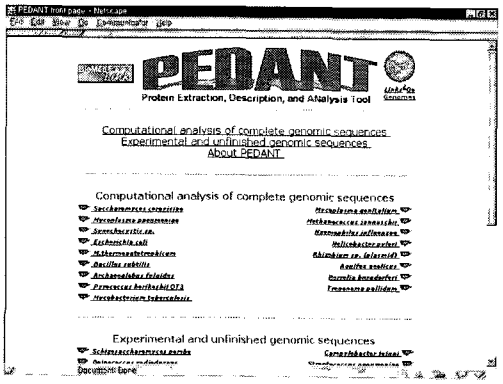


나. Functional Genomics를 위한 Bioinformatics의 접근방법

Genome Project에서 Bioinformatics의 지금까지의 기여도는 Sequencing Genomics (또는 Structural Genomics) 분야의 데이터베이스 구축과 1차적인 분석프로그램의 제공을 꼽을 수 있다. Genome Project가 Functional Genomics의 단계로 접어들면서 Bioinformatics는 서열 자료로부터 가공된 2차 데이터베이스의 구축과 염기서열이던 단백질서열이던 간에 기능을 함께 검색할 수 있는 시스템 개발에 주력하고 있다. 일 예로 독일의 Munich Information Center for Protein Sequence (MIPS)에서 개발한 PEDANT (Protein Extraction, Description and ANalysis Tool; http://pedant.mips.biochem.mpg.de)의 경우 지금까지 할 수 있는 모든 분석방법을 통해 Genome Sequence로부터 Protein의 구조-기능에 관련된 모든 정보를 얻어볼 수 있도록 되어있다.

또 다른 접근방법은 지금까지 연구된 종들의 Genome Sequence를 중심으로 알려진 유전자의 기능별 분류와 각 종간의 유전자의 기능별 비교를 통해 새로운 유전자를 찾기 위한 노력이다. 그 첫 번째 예로 다음과 같은 Yeast 유전자의 기능별분류 (Functional Catalogue)를 들 수 있다. 이와 유사한 연구로 우리가 눈여겨봐야 할 것은 KEGG





(Kyoto Encyclopedia of Genes and Genomes)으로 지금까지 알려진, 그리고 알려질 모든 종의 Genome Sequence를 기초로 Metabolic Pathway에 대한 통합 지도를 만들고 궁극적으로는 어떠한 종에 관계없이 적당한 Genome Sequence를 입력하면 거기에 어떤 기능을 하는 몇 개의 유전자가 있는지를 알 수 있는 시스템을 만들려고 하고 있다.

Functional Catalog of *Saccharomyces cerevisiae*

METABOLISM (1019 ORFs)
ENERGY (236 ORFs)
CELL GROWTH, CELL DIVISION AND DNA SYNTHESIS (753 ORFs)
TRANSCRIPTION (722 ORFs)
PROTEIN SYNTHESIS (343 ORFs)
PROTEIN DESTINATION (512 ORFs)
TRANSPORT FACILITATION (300 ORFs)
INTRACELLULAR TRANSPORT (414 ORFs)
CELLULAR BIOGENESIS (170 ORFs)
SIGNAL TRANSDUCTION (119 ORFs)
CELL RESCUE, DEFENSE, CELL DEATH AND AGEING (331 ORFs)
IONIC HOMEOSTASIS (114 ORFs)
CELLULAR ORGANIZATION (2105 ORFs)
RETROTRANSPOSONS AND PLASMID PROTEINS (113 ORFs)
CLASSIFICATION NOT YET CLEAR-CUT (152 ORFs)
UNCLASSIFIED PROTEINS (2605 ORFs)

다. Functional Genomics를 위한 프로그램 및 데이터베이스

PROSITE

1. Use
   : a method of determining the function of unchracterized proteins translated from genomic or cDNA sequences
   : PROSITE.DAT (scan sequence(s) with patterns and/or matrices.
   : PROSITE.DOC (textual information that fully documents each pattern and profile.
2. Web site URL
   : http://www.expasy.ch/sprot/prosite.html
3. Reference
   : *Nucleic Acids Research, 1996, Vol.24.No.1 189-196*

TRANSFAC

1. Use

: a database on eukaryotic cis-acting regulatory DNA elements and trans-acting factors.

2. Web site URL
: http://transfac.gbf.de/TRANSFAC/

3. Reference
: *Nucleic Acids Research, 1996, Vol.24.238-241, 1996.*

PDB

1. Use
: an archive of experimentally determined three-dimensional structures of biological macromolecules, serving a global community of researchers, educators, and students.

2. Web site URL
: http://www.pdb.bnl.gov/

PIR (Protein Information Resource)

1. Use
: to maintain a comprehensive data set, covering all naturally occurring protein sequences, .

2. Web site URL
: http://www.gdb.org/Dan/proteins/pit.html

3. Reference
: *Nucleic Acids Research, 1996, Vol.24.17-20,1996*

SWISS-PROT

1. Use
: a curated protein sequence database which stives to provide a high level of annotation(such as the description of the function of a protein, its domain structure, post-translational modifiactions, variants, etc), a minimal level of redundancy and a high level of integration with other databases.

2. Web site URL
: http://www.expasy.ch/sprot/sprot-top.html

3. Reference
: Nucleic Acids Res. 25 : 31-36(1997)

REBASE

1. Use
: a collection of information about restriction enzymes and methylases.

2. Web site URL
   : http://www.neb.com/rebase
3. Reference
   : Nucleic Acids Res. 25 : 248-262(1997)

TREMBL

1. Use
   : a protein sequence database supplementing the SWISS-PROT Protein Sequence Data Bank. TEMBL contains the translations of all coding sequences(CDS) presents in the EMBL Nucleotide Sequence Database not yet integrated in SWISS-PROT.
2. Web site URL
   : http://www.ebi.ac.uk/ebi_docs/swissprot_db/swisshome.html
3. Reference
   : Nucleic Acids Res. 25 : 31-36(1997)

라. 앞으로 풀어야할 문제들

DNA / RNA
Sequence Analysis
Prediction of Gene Structure
(Exon, Intron, Splice Site, Promoter, Enhancer)
Hydration and Environment
Genome Mapping
Secondary Structure Prediction

Proteins
Classification according to Structure, Sequence or Function
De novo Design
Discovery of Structure / Function Relationships
Docking and Enzyme - Substrate
Evolutionary Relationships
Folding Process and Motion
Force Field and Energetics
Homology-based 3D Modelling
Inverse Folding Problem
Localizing, Targeting and Signal Sequence of Membrane Proteins
Packing, Accessibility and Hydrophobicity
Prediction of Structure / Function Motifs
Secondary Structure Prediction

Simulation of Metabolism
Solvent Interaction
Tertiary Structure Prediction and Refinement

3. 세계 3대 유전체정보 센터

가. National Center for Biotechnology Information (NCBI, NIH)

○ 미국 국립보건원(NIH) 산하기관으로 1988년 11월 미국 국회의 발의에 의해 설립되어 세계 최대의 유전정보데이터베이스인 GenBank의 운영 및 보급
○ 전세계 유전체정보연구의 주도적 역할 담당하며 약 200명 이상의 연구인력을 갖춤
○ 데이터 베이스 이외에 유전자 분석에 필요한 기초연구를 병행하여 세계 최고의 유전자 정보분석체계 확립
○ 기초연구로 유전자 및 단백질 구조기능연구, 유전체 해석 염기서열 분석에 필요한 수학 및 물리이론연구, 염기분석 기기 설계, 소프트웨어 개발과 데이터베이스 디자인 등을 수행
○ 유전체정보이용연구를 인체유전체연구종료 후 100년간 수행할 연구로 추진 중

나. European Bioinformatics Institute (EBI, EMBL)

○ EU를 대표하는 종합유전정보 연구소로, **20년의 역사를 갖는 European Molecular Biology Laboratory**의 데이터베이스운영으로부터 시작하여 1994년 EBI(영국)로 분리하여 약 100여명이상의 연구인력으로 운영
○ 세계 3대 유전자 데이터베이스인 EMBL Data Library와 아미노산 서열 데이터베이스인 SwissProt의 운영과 보급
○ 유전체연구와 관련된 분석프로그램 개발과 보급, 단백질 기능 해석에 가장 큰 비중을 두는 기관으로 BioCatalog라는 시스템 운영
○ **슈퍼컴퓨터 급의 병렬 초고속 컴퓨터인 MasPar-1과 16개의 CPU와 2GB의 주메모리를 갖는 컴퓨터를 서버**로 운영하고 SGI Supercomputer를 중심으로 연구용 전산시스템을 구성하여 EU의 유전체정보연구를 주도

다. Center for Information Biology (CIB, NIG)

○ 1995년 **일본의 국립유전학연구소** (National Institute of Genetics)내에 설립
○ 미국, 유럽과 함께 세계 최대 유전자 데이터베이스인 GenBank의 일본 Node를 운영하며 DNA Data Bank of Japan (DDBJ)라는 독자 데이터베이스 구축
○ 기초 유전정보해석연구를 위해, 유전정보분석연구실, 유전자기능연구실, 대량유전정보연구실, 분과분류연구실을 두고 있다 (약30여명의 박사급 연구원으로 구성).
○ CIB의 최대 강점은 NIG의 고성능, **고용량의 슈퍼 컴퓨터시스템 (Fujitsu VPP500/40, Cray CS6400/128)**의 자원을 받아 세계최대의 유전정보분석 기관으로 발전을 도모함

NCBI의 Functional Genomics 분야

1. Cancer Genome Anatomy Project(CGAP) : 세포 내에서 암을 유발하는 유전적 변이에 대한 연구를 통해 암의 예방과 초기 진단, 적절한 치료법 선택을 위한 프로젝트로 tumor에서 발현되는 모든 유전자의 index확립과 암 세포의 분자적 구조를 해석할 수 있는 새로운 기술을 개발하려는 목적으로 운영되고 있다. 암 세포로 만든 cDNA library와 암 관련 유전자들을 찾아 볼 수 있다.

2. Gene Map of the Human Genome : human transcript map의 web page : 각 chromosome에 해당하는 유전자 map과 대표적인 유전자의 자세한 정보들과 GenBank, Protein database로의 연결기능을 제공한다.

3. UniGene : Unique Gene Sequence Collection for Human and Mouse는 EST data 들을 유용하게 정리하기 위하여 만들어 진 곳으로 GenBank sequence를 반복되지 않는 고유한 유전자들의 묶음으로 정리한 곳이다. 그 유전자가 발현되는 조직과 유전자의 map 위치들의 정보도 함께 보여 준다. 인간(Homo sapiens)과 쥐(Mus musculus)의 유전자들이 정리되어 있다.

1	현재의 GenBank의 EST division(dbEST)에 모인 EST(Expressed Sequence Tag)는 1,247,603개로 작년의 658,698개의 두 배로 뛰어 올랐다. 이 들 EST data는 100개 이상의 서로 다른 생물 종으로부터 나온 것으로써 가장 많은 EST를 얻어낸 5개의 생물 종은 human(65%), mouse(18%), nematode(5%), Arabidopsis(3%), rice(2%)의 순이다. GenBank는 새로운 유전자의 대부분을 차지하는 이 ESTs data를 유용한 형태로 정리하여 사용하기 위해 UniGene collection을 만들어 서비스하고 있다 UniGene은 현재 human과 mouse의 unique genes의 묶음들로 제공된다. NCBI에서 사용한 방법은 GenBank의 private division에 모인 human sequences들과 human ESTs들을 모아 유사성이 많은 3'UTRs(untranslated regions)을 공유하는 유전자들을 묶음으로 분류하였다. 이 방법을 사용하여 800,000개의 human ESTs를 41,000개의 묶음으로 분류하여 각각은 하나의 대표적인 human gene으로 간주하였다. UniGene cluster는 해당 유전자가 어느 정도 연구가 되었는가에 따라 cluster를 이루는 유전자의 숫자가 달라진다. 즉, hemoglobin subunit이나 serum albumin precursor같이 연구가 많이 된 유전자의 UniGene cluster는 가장 큰 cluster를 이루고 있다. 물론, 현재는 그 대상은 가장 많은 EST data를 얻은 human과 mouse의 유전자로 국한되어 있지만 NCBI에 의하면 다른 생물 종에 대한 UniGene collection을 계속 만들 계획이라고 한다. UniGene은 2달에 한 번 update되며 NCBI의 FTP site의 Data repository/unigene directory에서 file download가 가능하다. Web외의 다른 검색 도구는 제공되지 않는다.
2	UniGene database는 genome mapping을 위한 후보 유전자를 찾거나 새 EST등록시의 비교, 검색의 기준이 될 수 있다. 즉, 새로운 EST를 찾았을 때 UniGene cluster의 유전자 중 어느 것과도 유사성이 없다면 이 EST는 새로운 유전자이며, 또 하나의 UniGene cluster가 될 수 있는 것이다. large-scale expression analysis의 소재로 사용된다. genome mapping center들과의 연계하여 transcript map을 만들 수 있다. UniGene을 이용해 만들어진 transcript map을 사용하면 해당 유전자의 chromosome상 위치를 확인할 수 있다. 질병 유전자를 찾는 연구에 사용될 수 있다. 유전자 polymorphism의 연구에 사용될 수 있다.

4. Clusters of Orthologous Groups(COG) : 7개의 complete genome인 Escherichia coli, Hemophilus influenzae, Mycoplasma genitalium, Mycoplasma pneumoniae, Cyanobacteria-Synechocystis, Methanococcus jannaschii, Yeast-Saccharomyces cerevisiae의 protein sequence들을 비교하여 각 생물 종에서 서로 유사한 기능을 하는 유전자들을 크게 Information storage and processing, Cellular processes,

Metabolism, Poorly characterized로 나누어722개의 COG를 보여준다. 전체 6848개의 protein과 domain에 대해 분석하고 있다.

COGs: 서열이 완전히 밝혀진 유전체 Gene Family들간의 형태적 유사성에 의한 자연분류
NCBI에서 1차로 구축된 유전체 유전자 데이터베이스(이하 DB)를 이용하여 구축되는 2차 DB가 어떤 기능과 파급효과를 갖는지, NCBI의 2차 DB 중 하나인 "Clusters of Orthologous Groups (COGs)"로 예를 들어 분석해 보고자 한다. 급격히 축적되는 유전체 염기 서열로부터 최대한의 정보를 추출하려면, 진화상의 모든 보존적인 유전자를 그들의 상동 관계에 따라 분류해야한다. 서열이 완전히 밝혀진 7개의 유전체(이들은 진화상의 5대 주요 계통을 대표한다.), 즉 E. coli, H. influenzae, M. genitalium, M. pneumoniae, Cyanobacteria (Synechocystis), M. jannaschii, 효모(S. cerevisiae)의 단백질 서열들을 비교하여 각 생물 종에서 서로 유사한 기능을 하는 유전자들을 1)정보축적 및 처리 기능 관련 유전자군, 2)세포 생리 기능 관련 유전자군, 3)대사 기능 관련 유전자군, 4)기능 미확인유전자군 등으로 크게 구분하여 총 722개의 COG가 만들어졌다. 이들 전체 COG에는 총 6,995개의 단백질 도메인을 분석하고있다. 각각의 COG는 하나의 오르토로거스(orthologous; 2개의 유전자가 어떤 공통 조상에서 종분화하여 유래할 때 이들의 유전자는 orthologous라고 부른다.) 단백질 또는 최소 3 계통(phylogenetic lineage)으로부터 유래한 파라로거스(paralogous; 종분화가 아닌, 단 기능적으로 유사한 2개의 유전자가 유전자 중복에 의해 생길 때 '파라로거스'라고 부른다. 분자 계통수의 추정에 필요한 정보를 주는 것은 파라로거스 유전자가 아니고 오르토로거스 유전자이다.) 단백질들로 구성된다. 그러므로 각각의 COG는 공통 조상의 보존적인 도메인과 일치한다고 볼 수 있다. 이런 관계는 서열이 불완전하게 밝혀진 유전체에서 많은 기능 부위를 자동적으로 예견할 수 있는 기반이 된다. 그러므로 COG들은 기능적으로나 진화적으로 유전체 분석의 기본 틀이 된다.

5. OMIM : Online Mendelian Inheritance in Man는 인간의 유전자와 유전 질병에 관한 database이다.

6. dbEST : Database of Expressed Sequence Tags는 GenBank에 등록된 EST들을 새 division으로 나누어 만든 ESTdatabase이다.

7. dbGSS : Database of Genome Survey Sequences는 EST가 cDNA partial sequence인 것에 비해 genomic DNA의 partial sequence들을 모아 database화 함.

8. dbSTS : Database of Sequence Tagged Sites는 genome map의 표지로서 사용되는 짧은 sequence들을 database화한 것이다.

9. Electronic PCR : 웹 도구로서 원하는 유전자 상에서 dbSTS의 data와 맞는 부위를 찾아 주어 유전자의 map위치와 유전자 상의 primer binding site를 알려준다.

10. MMDB: Molecular Modelling Database는 X-ray crystallography 와 NMR spectroscopy로 밝혀진 3차 구조database이다. 3차 구조를 볼 수 있는 프로그램인 Cn3D와 3차 구조간의 상동성을 비교하는 VAST (Vector Alignment Search Tool)을 제공한다.

Research: Comparative analysis of macromolecular 3D-structure

Cn3D: Visualization of 3D structure and structural alignments

MMDB (Molecular Modelling Database) : A validated "computer-friendly" structure database

Threading: Algorithms for protein sequence-structure comparison

VAST (Vector Alignment Search Tool) : Protein structure-structure comparison algorithm and database of structural neighbors

Entrez3D: WWW-retrieval tools for all of the above

11. NCBI Taxonomy : 30,000가지의 서로 다른 종에 대한 계통 database이다. 한 달에 600개씩의 새로운 종이 첨

가된다. EMBL, DDBJ협의체와 외부의 이 분야 전문가들에 의해 만들어진다.

12. ORF Finder : 웹 도구로서 사용자의 유전자나 database의 유전자에서 Open Reading Frame을 찾아서 가로 막대 그림으로 보여준다. 원하는 ORF부위를 click하면 translation된 protein이 보여지고 그 protein에 대한 BLAST search를 할 수 있도록 연결되어 있다.

13. Human/Mouse Homology Maps : 인간과 쥐의 chromosome상 유전자 상동성을 table로 보여준다.

This is the Davis Human/Mouse Homology Map, a table comparing genes in homologous segments of DNA from human and mouse sources, sorted by position in each genome. A total of 1793 loci are presented, most of which are genes. The authors did not include pseudogenes, members of multigene families where specific homology relationships could not be determined, nor any other genes for which homology was in doubt. In addition, for 568 of the loci there are provisional assignments of markers that link the homology map with that of the Gene Map of the Human Genome. These links also provide a rough approximation of the position of markers in the Genethon linkage map. In constructing this table, the authors first ordered genes so as to best maintain order according to both human cytogenetic position and mouse genetic map position. Within these homologous regions, genes were ordered according to the mouse genetic mapping data. For approximately half of the genes in this database, no more detailed information is available; thus, much of this map should be interpreted as a reflection of probable, not confirmed, homology relationships. Where more detailed information was to be found, the authors adjusted the map to correspond to available human physical mapping data, which they consider to be definitive. The present rendition contains 201 homology groupings.