총 설

생물정보학의 과제 (Issues in Bioinformatics)

서울대학교
장병탁, 강철주

1. 서 론

인간의 청사진이라 할 수 있는 게놈(genome)의 DNA 서열을 규명하고자 하는 Human Genome Project가 2001년 2월 인간의 유전체 서열을 약 99% 밝히면서 현재 생물학은 "Post-genome era"로 접어들게 되었다. 원자탄을 만들기 위한 맨하탄 프로젝트나 달 착륙을 위한 아폴로계획에 버금간다고 말하여지는 Human Genome Project는 생물학에 있어서 큰 파장을 불러오고 있다.

약 30억 개에 달하는 인간의 유전체 서열 이외에 각종 bacteria와 archaea, *C.elegans, S.cerevisiae*등 많은 종의 유전체 서열이 밝혀졌고 또한 계속 밝혀지고 있다. 이러한 결과로 GenBank의 데이터양은 기하급수적으로 증가하고 있다(그림 1).

이렇게 폭발적으로 증가하는 서열 데이터와 DNA microarray 등의 high-throughput 도구들은 대량의 데이터의 저장과 얻어진 데이터로부터 유용한 정보를 어떻게 얻을 것인가 하는 고민을 던져주고 있다.

이러한 상황에서 다량의 생물학적 데이터로부터 컴퓨터를 이용하여 유용한 자료를 얻어내고 정리하는 생물정보학(bioinformatics)은 21세기 생물학의 발전에 있어서 가장 중요한 요소가 되었다.

생물정보학(bioinformatics)은 단어 그대로 생물학(bio)에 정보학(informatics)이 결합된 학문이다. 생물학에 있어서 정보학의 결합은 과거 18세기 분류학이 발전하면서 수집된 종을 분류하기 위해 통계적인 방법이 동원되던 것이 시초라고 할 수 있다. 생물정보학은 생물학의 연구에 있어서 특히 분자수준의 연구에 있어서 컴퓨터 공학과 같은 정보학의 기술을 이용하는 것이라고 할 수 있다. 1950년대 Frederick Sanger에 의해 최초로 단백질(insulin)의 서열이 밝혀지고 아미노산 서열이 차츰 축적이 되면서 생물정보학이 태동했다고 본다. 이후 염기서열을 알아내는 방법이 개발이 되면서 차츰 염기서열의 양이 축적이 되어가는 과정에서 각종 자료를 축적하고 분석하는 초기의 생물정보학이 발전하였다. 1990년대 후반과 21세기에 이르러 Human Genome Project라는 원대한 작업의 결과와 DNA microarray과 같은 새로운 기술의 발전은 컴퓨터의 도움 없이 해석하기 힘든 방대한 양의 데이터를 생성해냈으며 이를 분석하고 이용하기 위한 생물정보학이 각광 받게 되는 환경을 만들어 냈다.

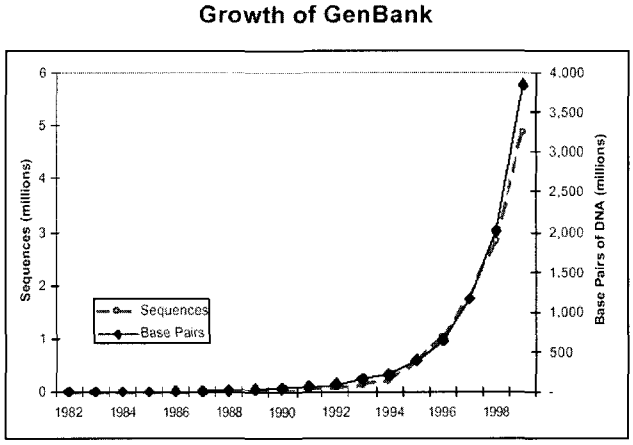


그림 1. DNA 서열 데이터베이스의 성장

초기의 컴퓨터 분자생물학(computational molecular biology)이라는 분자생물학의 한 분야로 간주되던 시대에서 생물정보학(bioinformatics)이라는 독자적인 분야로 성장했으며 이와 함께 새로이 유전체학(genomics), 단백질체학(proteomics)이라는 새로운 학문이 속속 등장하는데 큰 영향을 미쳤다고 볼 수 있다.

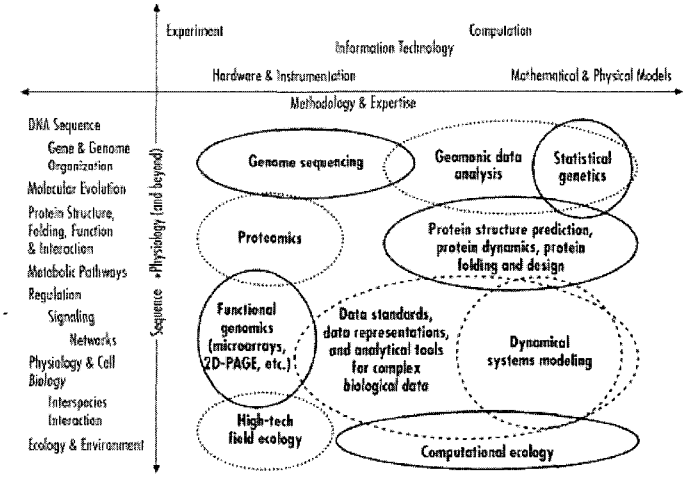


그림 2: 생물학과 생물정보학의 분야와 적용기술

그림2와 같이 생물정보학이 다루는 영역을 크게 본다면 DNA 서열, 아미노산 서열과 같은 서열 데이터의 영역을 제일 먼저 꼽을 수 있을 것이다. 서열에서 유전자나 promoter의 지역을 찾거나 혹은 다른 서열과의 상동성 탐색 등 생물정보학의 영역 중 가장 오래된 분야라고 할 수 있다. 이외에도 생물정보학의 적용범위는 염기서열, 단백질서열, 세포 내 네트워크 등 각종 생물학 데이터를 저장하는 데이터베이스의 구축, 서열데이터의 분석, DNA chip, 2D-PAGE gel 이미지 등의 발현데이터 분석, 단백질의 구조예측, 세포내의 유전자 네트워크 예측 등의 거의 모든 분자생물학의 연구영역에 미친다고 할 수 있다.

본 고에서는 이러한 생물정보학의 연구분야를 개괄하고 각 분야에서 현재까지 개발된 주요 데이터 분석 도구를 살펴보고자 한다. 1절의 서론에 이어 2절에서는 서열분석에 대하여 논하고 3장과 4절에서는 발현분석(expression analysis)과, 구조분석(structure analysis) 그리고 5절에서는 유전자 네트워크 분석(genetic network analysis)을 살펴본다. 마지막으로 6절에서는 요약과 결론을 제시한다.

2. 서열분석 (Sequence Analysis)

Sequencing 작업을 통해 얻은 DNA 서열이나 나 아미노산 서열을 분석하는 서열분석(sequence analysis) 작업은 1차원적인 A, C, G, T 혹은 20가지 아미노산의 배열에서 유용한 정보를 얻어내는 작업이다.

서열분석의 영역은 두 서열간의 homology를 찾기 위한 pair-wise alignment, 여러 서열을 비교하여 conserved region을 찾거나 서로간의 phylogenetic relationship을 보기위한 multiple alignment와 같은 기본적인 alignment와 gene finding, promoter finding, functional motif searching과 같은 pattern finding, 이외에 primer design, restriction enzyme site mapping, sequence assembling과 같은 기초적인 작업이 있다.

문 제	정 의	도 구
Pair-wise alignment	두 서열 사이의 homology가 어느 정도인가?	Smith Waterman Algorithm, BLAST, FASTA
Multiple alignment	여러 서열에서 conserved된 region은 어디인가?	Clustal W, MACAW
Phylogenetic analysis	여러 서열 사이의 진화학적 관계는 어떻게 되는가?	Clustal W, PHYLIP
Pattern Finding	서열 중 gene의 위치는?	GENSCAN, GeneMark, Glimmer, GRAIL
	서열에 알려진 특정 motif가 존재하는가?	PROSITE, *e*MOTIF
	아미노산 서열 중 특정 signal peptide가 있는가?	SignalP, ChloroP
Other sequence handling tools	실험 중 보다 편하게 sequence를 다루기 위한 도구	VectorNTI, Bioperl, Biopython, EMBOSS

표1 : 서열분석 문제 및 도구

유전자 발견(Gene finding)

인간과 다른 다양한 생물들의 전체 염기서열이 발표가 되면 필연적으로 유전자의 위치를 확인하고 유용한 유전자를 발굴하기 위한 작업이 뒤따르게 된다. 분자생물학적인 실험을 통한 유전자의 확인이 아닌 서열 데이터에서 유전자를 예측하려는 노력은 DNA sequencing 방법이 개발된 이후 곧 시도되었으며 현재 큰 발전을 이루고 있다.

Gene finding의 예측 기법은 크게 homology-based 방법과 *ab initio* 방법으로 나눌 수 있다. Homology-based 방법은 BLAST와 같은 alignment작업을 통하여 대상 서열과 일정 수준이상의 homology를 보이는 EST 또는 밝혀진 유전자를 찾아내고 비교함으로써 유전자를 예측한다. 이에 반해 *ab initio* 방법은 신경망(neural network), 은닉 마르코프 모델 (hidden Markov model) 등의 전산 통계적 방법을 기반으로 하여 유전자를 예측해낸다. Homology-based method의 장점은 error나 gap이 있는 서열에서도 비교적 수월하게 유전자를 찾아낼 수 있다는 장점이 있다. 하지만 이미 알려진 유전자가 데이터베이스에 존재하여야 함으로 타 종이나 EST clone이 확보가 되지않은 신규유전자의 경우에는 이 homology-based 방법으로는 결코 찾을 수가 없는 단점을 지닌다. 현재 homology-based 방법의 경우 약 60%정도의 유전자를 찾을 수 있다고 알려져 있다. 두 번째로 발현이 되지 않거나 그 기능이 없어진 non-functional pseudogene과 기능이 있는 유전자를 구별하기 힘든다는 점이 있다. *Ab initio*방식의 gene-finding 방식의 경우 일반적으로 homology-based 방식보다 sensitivity에 있어서 높은 효율을 보이고 있으나 다소 false-positive가 높은 경향을 보이고 있다. 이러한 *ab initio* 방법의 경우 진핵생물의 유전체에 적용될 경우와 원핵생물의 유전체에 적용될 경우에 다소의 차이가 있으며 진핵생물의 경우 특히 splicing site에 대한 고려가 필요하며[2] 원핵생물의 경우 전체적으로 유

전체에 유전자가 밀집된 구조인 것을 고려 하며 또한 상당수의 유전자가 다른 유전자와 겹치는 overlapping gene인 것 등에 대한 고려가 필요하다[3].

분석 도구	기 능
MZEF [4]	- Prediction of internal coding exon - Using the quadratic discriminant function for ultivariate statistical pattern recognition
GenScan [5]	- Gene prediction from eukaryotic genome - Using a number of different algorithms to predict introns, exons, donor and acceptor splice sites, and polyadenylation sites - Based on hidden Markov Model
Genie [6]	- Generalized Hidden Markov Model - Gene prediction from eukaryotic genome
GeneParser [7]	- Using standard content and site statistics weighted by a neural network. - Processing with a dynamic rogramming algorithm to find the maximum likelihood parsing of the sequence into functional domains.
GeneBuilder [8]	- Based on prediction of functional signals and coding regions by different approaches in combination with similarity searches in proteins and EST database
GeneMark [9]	- Based on hidden Markov Model - GeneMark.hmm algorithm which generates a maximum-likelihood parse of the DNA sequence into coding and non-coding regions
Grail [10]	- Multiple sensor-neural network approach - Combine a set of sensor algorithms
NetGene2 [11]	- Prediction of splice site in human, *C elegans* and *A. thaliana* genome - Based on Artificial neural networks combined with a rule based system
GLIMMER [3]	- Gene prediction from prokaryotic genome - Instead of using fixed Markov chain, using interpolated Markov chain
FGENESH [12]	- Prediction of multiple genes in genomic DNA sequences - Based on HMM similar with Genescan and Genie
SplicePredictor [13]	- Prediction of splice site from *A.thaliana* genome sequence by sequence inspection
TRNAscan-SE [14]	- Finding tRNA sequence from genomic DNA or RNA sequence - Finding polIII promoter sites - Combining several earlier programs

표2: 유전자 예측 도구

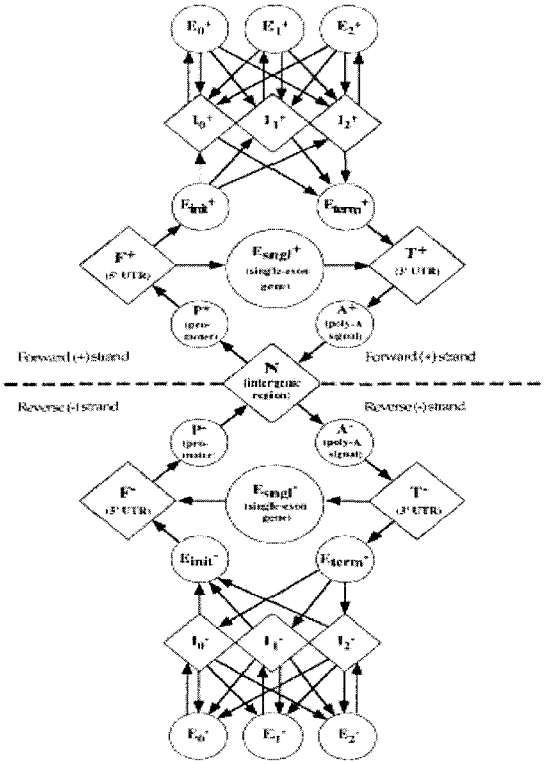


그림 3 : GenScan에서 설정한 유전체 구조

GeneMark, Genie 등의 현재 eukaryotic gene finding tool 중에서 GENSCAN의 성능이 뛰어나다고 알려져 있다. Carlin과 Burge에 의해서 개발된 GENSCAN은 graphical model인 은닉 마르코프 모델을 기반으로 한 유전자 예측 도구이다[5]. GENSCAN은 그림 3과 같은 모델을 설정하여 C+G content, initial state probability, transition probability, state length distribution 과 각종 시그널 (promoter, polyA signal) 모델 등을 계산하여 유전자를 예측하게 된다.

EBI, Sanger Centre, EMBL이 협력하여 인간의 유전자를 찾고 annotation 하기위한 협력 과제인 Ensembl project에서도 GENSCAN이 쓰이고 있다.

이러한 분석 방법과 함께 인간의 유전자를 예측하는 가장 좋은 수단은 차후 발표될 생쥐와 침팬지의 유전체 서열일 것이다. 인간과 함께 포유류에 속하는 이들 종의 유전체 서열이 얻어지게 되면 종간에 유전자는 highly conserved region 에 속하게 될 것이므로 이러한 부분들을 찾아내게 된다면 다른 유전자 예측 도구와 함께 보다 더 정확하게 유전자를 확인할 수 있을 것이다[15].

현재 유전자 예측 도구들의 효율은 정확성은 대상 생물 종과 경우에 따라 많은 차이가 있으나 대략 진핵생물의 유전체를 대상으로 할 경우 대략 전체 유전자 중 80%이상을 예측해내며 원핵생물의 유전체의 경우에는 90% 이상의 gene을 찾아내고 있다. 각 시험마다 다소의 차이를 보이고 있고 또한 대상 genome의 종류에 따라 차이를 보이고 있으나 현재 진핵생물 대상의 유전자 예측 도구들 중에서는 GENSCAN이 가장 성능이 뛰어나다고 인정되고 있으며 원핵생물 대상의 유전자 예측 도구의 경우는 Glimmer의 성능이 우수하다 인정되고 있다.

이러한 기대치 이상의 효율을 보이는 유전자 예측 도구이지만 다음 과제는 alternative splicing site의 예측이다[16]. Human genome project의 결과 발표에서 Celera는 인간의 유전자 수가 약 37000개라고 예측하였다. 이 수는 종래의 10만개 내외라 예측되었던 인간의 유전자 개수보다 크게 적은 것으로 *C.elegans*의 유전자가 1900여 개에 달한다는 것을 감안할 때 인간 세포의 복잡성을 설명하기에는 턱없이 적은 개수의 유전자가 존재하는 것이다. 바로 "c-value paradox"에서 "n(number)-value paradox"로 문제가 넘어가게 된 것이다. 이런 불합리성을 설명하기 위해 alternative splicing과 protein modification을 통해서 하나의 유전자가 몇 개의 다른 function을 보인다는 설명이 대두되고 있다. 실제로 척추동물에서는 alternative splicing의 예가 많이 보고 되고 있으며 EST 서열과 mRNA 서열을 유전체에 alignment 해보았을 때 적어도 35% 이상의 유전자가 alternative splicing을 할 것으로 예측되었다. 이러한 결과로 볼 때 인간의 유전체에서 유전자를 예측한 후의 작업은 바로 각 유전자의 alternative splicing의 형태를 예상하는 작업이라 할 수 있다. 다음 세대의 유전자 예측 도구는 이러한 alternative splicing까지 예측할 수 있는 툴이 되어야 하며 또한 alternative splicing은 각종 시그널에 의해 조절되므로 splicing regulatory pattern까지도 인식을 하는 것이 과제가 되고 있다.

3. 발현분석 (Expression Analysis)

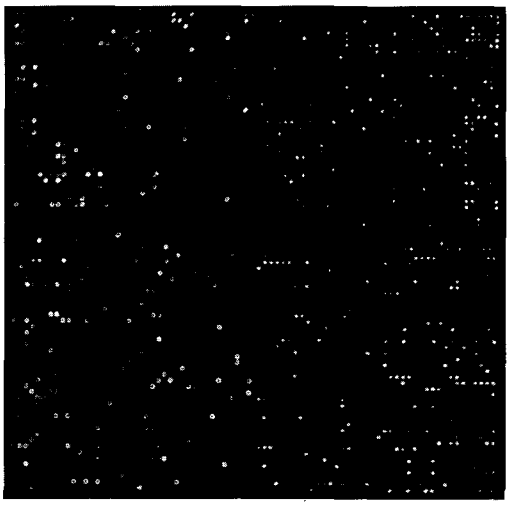


그림 4: DNA microarray의 hybridization 결과

1990년대와 21세기에 들어서 생물학계 전체에 걸친 두개의 사건을 꼽는다면 Human Genome Project와 DNA microarray 기술의 발명이 될 것이다.

그림 4와 같이 수천개의 single strand cDNA를 유리슬라이드 위에 배열하고 labeled cDNA를 hybridization시키는 일종의 reverse Northern blotting기술인 DNA microarray 기술은 생물학 연구에 일대 혁명적인 실험 기법이 라 할 수 있다. DNA microarray의 종류는 크게 cDNA microarray와 oligonucleotide chip으로 나눌 수 있다. cDNA microarray는 DNA를 슬라이드 글래스 위에 spotting을 하여 칩을 만들어낸다. 이에 반해 oligonuleotide chip은 DNA를 슬라이드 위에 spotting시키는 것이 아닌 nucleotide를 하나 하나 붙여가면서 chip위에서 직접 10mer에서 25mer정도의 oligonucleotide를 합성한다. 각 chip의 적용분야는 겹치는 경우도 있으나 대게 cDNA chip은 drug의 효과, metabolism, disease의 진단, gene pathway finding 등에 주로 사용되며 oligonucleotide chip은 SNP, 유전자 돌연변이의 진단, sequencing 등에 주로 사용되고 있다.

cDNA microarray	약물에 영향 받는 유전자 발굴 질병 진단 유전자 네트워크 질병유발 유전자 발굴
Oligonuleotide chip	SNP발굴 Sequencing by chip 특정 돌연변이 확인

표3: DNA 칩의 종류 및 응용분

DNA microarray실험은 방대한 양의 결과를 내는 high-throughput 실험이므로 생물정보학의 도움이 없다면 단순히 Northern blotting을 여러 번 하는 것과 별 차이가 없다. 효율적인 실험과 또한 좋은 데이터를 얻기 위해선 우선 DNA microarray의 디자인에서부터 생물정보학의 도움이 필요하며 hybridization 결과의 스캐닝시 spot의 화상처리와 같은 화상의 처리 작업이 요구된다. 얻은 화상 데이터를 분석하고 유용한 데이터를 얻어내는 과정과 다른 데이터와의 연동 등은 생물정보학의 좋은 예이다.

DNA microarray의 적용

초기 괄목할 만한 DNA microarray실험으로는 1996년 DeRisi의 budding Yeast(*S.cerevisiae*)에 대한 실험을 들 수 있다[17]. DeRisi등은 DNA microarray를 이용하여 yeast가 glucose의 anaerobic fermentation에서 ethanol의 aerobic respiration으로 전환하는 시기(diauxic shift)에 어떠한 유전자들이 발현되고 또 발현이 정지하는 것을 관찰하여 그 전까지 기능이 알려지지 않은 유전자들의 기능을 유추하였다.

다음 DNA microarray분석 기술은 군집화(clustering)이다. 1998년 Eisen등은 역시 budding Yeast의 cell cycle, diauxic shift, sporulation 등의 다양한 상태에서의 DNA microarray 데이터에서 계층적 군집화(hierarchical clustering) 방식을 이용하여 유전자들을 군집화 한 결과를 발표하였다. 기능의 유사한 유전자들은 같은 클러스터 내지는 인접한 클러스터에 위치하는 것이 확인되었다[18]. DNA microarray 데이터의 분석에 군집화 방법을 도입하게 되었고 이후 군집화는 DNA microarray 데이터 분석의 표준이 되었다.

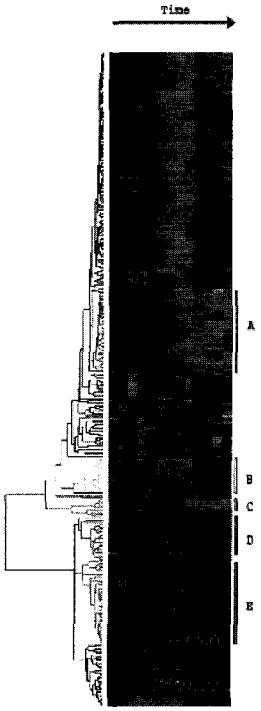


그림 5: DNA 칩 데이터의 계층적 군집화 (hierarchical clustering) [16]

군집화 방법 또한 hierarchical clustering, support vector machines[19], self-organizing map (SOM), k-means clustering 등의 여러 clustering 방법들이 적용이 되고 있다[20].

이러한 군집화와 같은 분석 방법은 DNA microarray 데이터에서 단순히 유전자의 기능만을 유추하는 것이 아닌 발현 패턴을 보고 세포의 상태를 유추하는 것을 가능하게 하였다.

Leukemia의 두 subtype인 acute myeloid leukemia (AML)과 acute lymphoblastic leukemia(ALL)의 DNA microarray 데이터를 군집화를 통해 분석함으로써 두 subtype간의 expression의 차이를 구별하고 새로운 샘플의 DNA microarray 데이터만을 가지고 샘플이 어느 종류에 속하는지 결정이 가능하게 하였다[21]. 이 실험 이외에도 다양한 상태의 질병을 DNA microarray를 이용하여 진단하는 것이 보고되고 있다.

DNA microarray의 화상분석, 데이터 분석 도구들은 다른 생물정보학 도구와 달리 상용 소프트웨어의 수가 많다. 무료 소프트웨어 와 상용 소프트웨어간의 성능의 차이는 그다지 크지 않으나 상용 소프트웨어의 경우 사용자의 편의성을 극대화하여 좀 더 쉬운 조작을 통해 데이터를 분석하도록 제공하는 등의 부가기능을 제공하고 있다.

화상분석, spot 확인, 정량 소프트웨어

Free software	ScanAlyze	M. Eisen, Stanford Univeristy
	TIGR spot finder	TIGR
	CrazyQuant	L.Hood, University of Washington
Commercial software	ImaGene	BioDiscovery
	ArrayPro	Media Cybernetics
	ArrayVision	Imaging Research, Inc
	Iplab	Scanlytics, Inc

데이터분석 소프트웨어

Free Software	Cluster	M.Eisen, Stanford Univerisy
	MAExplorer	Lab. Experimental and Computational Biology, National Cancer Institute
	Cyber T	T.Long and H.Mangalm, UC Irvine
	CLEAVER	Stanford Biomedical Informatics
	J-Express	B.Dysvik
	GeneCluster	Whitehead/MIT Centre for genome research
Commercial Software	arraySCOUT	LION Bioscience
	GeneSpring	Silicon Genetics
	GeneMaths	Applied Maths
	GeneSight	BioDiscovery
	Expressionist	GeneData
	AnalyzerDG	MolecularWare
	Resolver	Rosetta Inpharmatics, Inc
	ArrayVision	Imaging Research, Inc
	Spotfire Array Explore	Spotfire
	CHIPSpace	MiraiBio Inc.

표 4 : DNA micorarray 소프트웨어

DNA microarray가 mRNA level의 expression을 보는 high-throughput 도구라면 단백질수준의 high-throughput실험 도구는 2D PAGE가 있다. 세포내의 모든 단백질을 분자량과 isoelectric point(pI)에 따라 분리하여 각각의 단백질은 하나의 spot으로 나타나게 된다(그림 6). 각각의 spot은 gel에서 분리되어 matrix-assisted laser desorption ionization mass spectrometry(MALDI)로 분석이 이루어 진다. 결국 이러한 작업은 대량으로 이루어지게 되고 또한 데이터의 분석을 위해선 컴퓨터의 도움이 필수적이다.

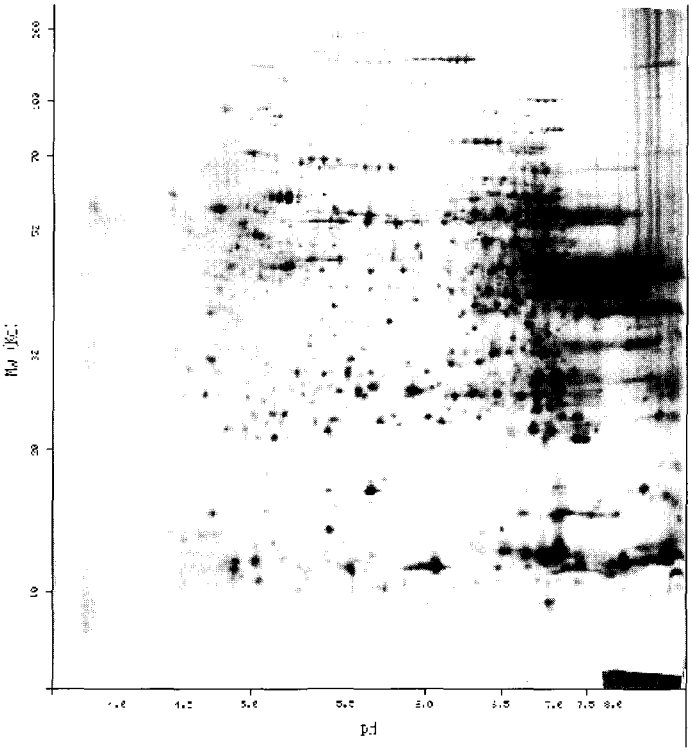


그림 6: 사람 간 단백질의 2D-gel 사진(ExPASy)

DNA microarray기술과 응용은 급격한 발전을 이루고 있으며 미래의 microarray기술이 어느 정도 발전할지는 예측하기 힘들다. Hardware적으로는 보다 재현성이 높도록 DNA microarray자체의 제조 기술과 실험방법이 개선 될 것이며 또한 microarray자체가 소형화 되고 전자공학과의 결합으로 scanning을 통하지 않고 직접 데이터를 얻게 될 것이다. 이러한 발전은 DNA microarray기술이 더욱더 실생활에 적용되기 쉽다는 것을 의미한다. 실험 데이터 분석에 있어서도 기존의 군집화 방법보다 높은 효율의 군집화 기법의 고안과 함께 군집화 이외의 다른 데이터 분석기법이 고안될 것이다. 단백질 발현 분석의 경우 DNA microarray보다 재현성이 떨어지는 것을 방지한 protein chip과 같은 실험기법이 개발될 것이다.

4. 구조분석 (Structure Analysis)

생체내의 많은 macromolecule 특히 단백질의 경우에는 구조가 그 기능에 중요한 역할을 하게 된다. 단백질의 구조를 밝히는 일은 protein-protein interaction, protein-DNA interaction, antibody-antigen binding, drug-receptor binding 등을 이해하는데 있어서 필수적인 요소라고 할 수 있다.

단백질의 구조를 밝히고자 하는 노력은 생물학의 주요한 한 부분이었다. 실험적으로는 원하는 단백질을 결정화 시켜 X-ray회절을 통해 단백질의 구조를 예측하는 X-ray crystallography는 많은 수의 단백질의 구조를 밝혀냈고 최근에는 nuclear magnetic resonance(NMR) 방법을 통해 단백질의 구조를 밝히기도 한다. 이렇게 밝혀진 단백질의 구조들은 Protein Databank에 고유한 형식으로 저장된다.

이러한 실험적인 방법은 시간적인 문제와 또한 in vivo가 아닌 in vitro상태의 구조를 예측한다는 문제, 그리고 단백질의 dynamic structure를 예측하지 못한다는 문제가 존재하게 된다. 이러한 문제를 해결하고 보다 쉽게 구조를 예측하려는 것은 생물정보학의 주요한 연구분야이다.

단백질 구조 예측

단백질은 1차 구조인 아미노산의 서열이 local하게 α-helix, β-sheet, coil등의 2차 구조를 이루며 결국 3차원적인 3차 구조를 결정하게 된다. 결국 1차 구조인 아미노산 서열에 단백질 구조에 대한 모든 정보가 저장되어 있다고 볼 수 있는 것이다(그림7).

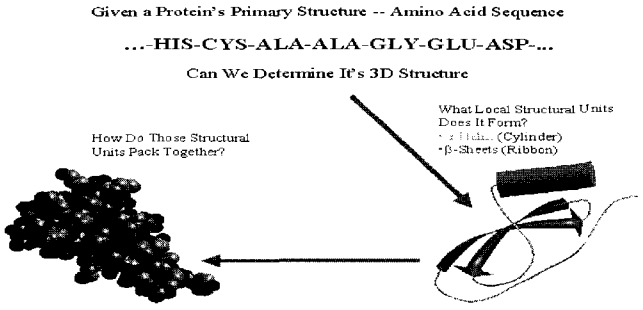


그림 7: 단백질 구조 예측

컴퓨터를 이용한 단백질의 구조 예측 방법은 크게 두 가지로 나눌 수 있다. 하나는 지식기반 방법으로 이 방법은 이미 알려진 단백질 구조를 참고하여 새로운 단백질의 구조를 예측한다. 다른 하나는 단백질의 물리적 힘을 계산하여 단백질의 구조를 밝히려는 시도이다.

2차 구조의 예측은 alignment-based 방식과 single-sequence based 방식으로 구별할 수 있다. Alignment-based 방식은 두 서열사이의 homology가 일정 threshold이상일 경우 2차 구조가 거의 유사함을 이용한다. 이에 반해 single-sequence based method는 alignment를 이용하지 않고 아미노산의 특성이나 확률 등을 계산하여 2차 구조를 예상하게 된다. Chou-Fasman 방법은 아미노산의 side-chain 등에 대한 생화학적 성질 등을 고려하여 2차 구조를 예측하며[22], GOR 방식은 20종류의 아미노산이 각각의 2차 구조에 나타나는 확률값을 기반으로 2차 구조를 예측하게 된다[23].

2차 구조의 예측의 정확도는 현재 크게 증가하여 α-helix의 예측의 경우에는 80% 정도의 정확성을 보이고 있다. 이는 3차 구조의 예측에 비해 크게 높은 수치 이나 β-sheet의 예측의 경우에는 정확도가 떨어지는 경향이 있다.

NnPredict [25]	- Using two-layer, feed-forward neural network
BTPRED [26]	- Predict beta-turns - using combination of neural networks
Jpred [27]	- Combining a number of predition methods
PREDATOR [28]	- Based on recognition of potentially hydrogen-bonded residues in the amino acid sequence
GOR [23]	- Based on information theory
SOPMA [29]	- Improving self-optimized prediction method

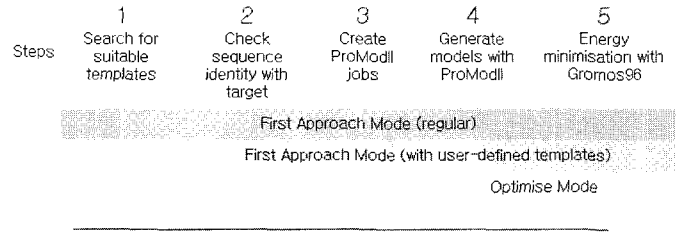
표5: 단백질 2차 구조 예측

3차원 구조의 예측은 앞에서 언급한 바와 같이 homology modeling, threading과 같은knowledge-based method와 단백질의 물리적 힘을 계산하여 구조를 예측하는 *ab initio* method로 나눌 수 있다.

Homology modeling은 아미노산의 서열이 유사한 단백질의 경우에는 그 구조가 거의 일치한다는 점에 근거를 두고 있다. 구조를 밝히고자 하는 단백질의 아미노산 서열과 이미 구조가 밝혀진 단백질의 아미노산 서열이 서로 유사할 경우 밝혀진 단백질의 구조를 기반으로 하여 원하는 단백질의 구조를 예측한다. Homology modeling은 대략 다음과 같은 과정을 거친다(그림 8).

1. query 단백질과 서열이 유사한 구조가 알려진 template 단백질을 찾아낸다.
2. 두 단백질의 alignment
3. template 단백질의 구조를 기초로 하여 단백질의 backbone을 모델링한다.
4. sidechain을 덧 붙이고 최적화한다.

5. 에너지 최적화 등의 방법을 통하여 전체적인 구조를 최적화한다.



Step	Program/Method	Database	Action
1	BLASTP2	ExNRL-3D	Will find all similarities of target sequence with sequences of known structure.
2	SIM	-	Will select all templates with sequence identities above 25% and projected model size larger than 20 residues. Furthermore, this step will detect domains which can be *modelled* based on unrelated templates
3	-	-	Generate ProModII input files
4	ProModII	ExPDB	Generate all models
5	Gromos96	-	Energy minimsation of all models

그림 8: SWISS-MODEL의 순서도

Homology modeling은 위와 같이 alignment를 통하여 이미 알려진 단백질의 구조를 참고하게 된다. CPHmodels[29], Swiss-Model [30], 3D-Jigsaw[31], Modeller[32] 등의 분석 도구들이 이러한 homology modeling을 기반으로 하고 있다.

다른 단백질의 밝혀진 구조를 참고로 하는 homology-modeling과는 달리 *ab initio*방법은 단백질 분자를 구성하는 하나 하나의 물리적인 힘을 고려하여 단백질의 구조를 예측하게 된다. *Ab initio* 방법은 특정 생화학적 조건 하에서 전체계의 자유에너지가 가장 낮도록 단백질의 3차원 구조가 결정된다는 Anfinsen의 이론을 기본으로 하고 있다[33].

현재 어느 정도의 성과를 보이고 있는 homology modeling과는 달리 *ab initio* 방법은 아직 뚜렷할 만한 성과를 보이지 못하고 있는 실정이다. 이는 더 나은 컴퓨터와 각 에너지 함수의 최적화를 필요로 하고 있다.

컴퓨터를 이용한 단백질의 구조예측은 단백질의 기능을 예측하는데 큰 도움이 될 뿐 아니라 실험을 통한 구조 예측과 달리 그 시간을 크게 단축할 수 있는 장점이 있다. 또한 지용성 단백질의 일부와 같은 크리스털이 생성되기 힘든 단백질의 구조를 예측할 수 있다.

현재 가장 빠른 컴퓨터로 개발이 되는 IBM의 Blue Gene도 이러한 단백질의 구조에 대한 문제를 풀도록 계획이 되어 있다(그림9). Blue Gene은 백만 개의 기가FLOP급의 floating processor를 병렬처리 하여 1 petaFLOP/s이상의 속도를 내도록 계획하고 있는 슈퍼컴퓨터이다.

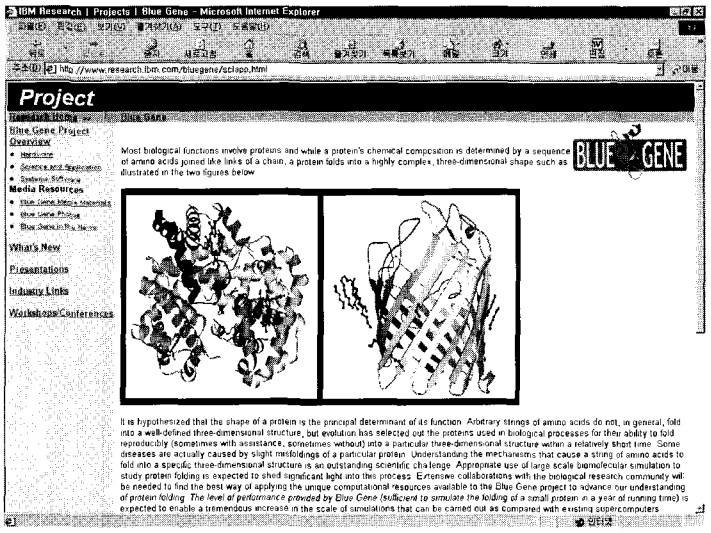


그림 9: IBM Blue gene homepage
(http://www.research.ibm.com/bluegene)

5. 네트워크 분석 (Network Analysis)

분자생물학의 전통적인 접근방법은 유전자(gene)와 그 산물인 단백질(protein)의 기능을 알아냄으로써 전체적인 생물체를 이해하고자 하는 것이었다. 이러한 접근법은 전통적인 생물학적 실험이 가지고 있는 한계에 기인한다. 과거의 생물학적 실험은 생성할 수 있는 데이터의 양이 한정되어 있었기 때문에 그 결과를 해석하는 방법이나 밝혀낼 수 있는 지식에 한계를 가지고 있었다. 이러한 실험에서는 주로 어떤 특정한 형질(phenotype)에 영향을 미치는 소수의 유전자를 실험을 통해 밝혀내고 그 유전자의 발현과 형질간의 관계를 결정 트리(decision tree)와 같은 간단한 통계적 기법을 통해 밝혀 왔다. 이렇게 부분을 이해함으로써 전체를 이해하고자 하는 환원주의적(reductionism)인 접근 방법은 지금까지의 분자생물학의 발전에 있어서 큰 도움이 되어 왔던 것은 부정할 수 없다. 하지만 생명의 청사진인 인간의 유전체 서열이 밝혀진 지금 이러한 유전자 결정론은 오히려 큰 반론의 대상이 되고 있다. 사실, 생명은 몇 개의 유전자의 단순한 pathway로 구성된 것이 아니고 수 많은 유전자들이 복잡한 network을 이루고 있는 구조이다. 이는 일종의 복잡계(complex system)로 이를 이해하기 위해서는 각각의 원소인 유전자의 기능에 대한 이해 뿐만이 아닌 각 원소간의 상호작용을 이해해야 한다[34]. 최근, cDNA microarray, oligonucleotide chip, SAGE(serial analysis of gene expression)와 같은 기술의 발전은 이전에는 불가능했던 대량의 생물학 실험 데이터 생성을 가능하게 했으며 이러한 대량의 데이터를 이용해서 수 많은 유전자들간의 복잡한 관계를 파악하려는 시

도가 행해지고 있다. 유전자 네트워크 분석은 이러한 시도에 속한다고 볼 수 있다.

유전자 네트워크 분석은 유전자 및 생명체에 존재하는 여러 가지 효소나 단백질들 사이의 복잡한 상호작용을 표현하는 네트워크를 구성하고 이를 통해서 생명체 안에서 일어나는 여러 현상들을 이해하고 예측하려는 시도이다. 기존의 유전자 네트워크 구성은 주로 protein-protein interaction, DNA-protein interaction, RNA-protein interaction, substrate cascade 등의 자료를 이용해 왔다[35]. 특히, 최근에는 cDNA microarray 기술과 같은 대규모의 데이터를 얻을 수 있는 기법이 발전함에 따라 이러한 대량의 데이터를 이용해서 유전자 네트워크를 구성하는 방법에 대한 연구가 활발히 이루어지고 있다[36]. 이러한 시도는 개체의 행동을 관찰해서 그 내부 기작을 밝히려는 것으로 일종의 역공학(reverse engineering)이라고 할 수 있다. 대량의 데이터를 이용한 유전자 네트워크의 구성에는 통계학과 기계학습(machine learning)의 기법들이 이용된다. 이러한 기법에는 Boolean modeling, differential equation modeling, stochastic modeling 등이 있다. Boolean modeling은 간단한 Boolean equation을 통해 유전자들 간의 발현 관계를 표현하며 genetic network의 가장 간단한 형태이다. Boolean model에서는 유전자들의 발현을 단순한 이진 상태로만 표현할 수 있다. 미분 방정식(differential equation)을 통한 modeling은 Boolean modeling보다 조금 복잡한 형태로 유전자의 발현을 실수 형태로 표현 할 수 있다. stochastic modeling은 각 유전자들간의 발현관계를 확률적으로 표현하는 기법으로 대량의 데이터를 이용해 구성되는 genetic network이 궁극적으로 지향해야 할 모델이라 할 수 있다. stochastic model은 생물체 내에 존재하는 확률적인 특성과 데이터의 생성에서 기인하는 확률적인 면들을 고려하게 되므로 다른 모델 보다 표현력이 강하다고 볼 수 있다. 물론 표현력이 강한 유전자 네트워크를 구성하려면 훨씬 많은 데이터를 필요로 한다.

이러한 유전자 네트워크의 구성은 기존의 생물학, 약학, 의학 연구에 큰 변화를 가져올 것이다. 예를 들어, 특정한 암의 발병과 관련된 유전자에 대한 기존의 연구는 정교한 생물학적 실험을 통해 발병과 직접 관계가 있는 소수의 유전자만을 밝혀내는 것이었다. 하지만 실제 암의 발병과 관련이 있는 유전자들 사이의 발현 관계는 매우 복잡할 수 있으며 이러한 관계는 기존의 방법으로는 밝혀낼 수 없었다. 유전자 네트워크를 구성하게 되면 암의 발병과 관련된 수십, 수백 개의 유전자들의 상호 작용에 대한 이해를 얻을 수 있으며 이를 통한 질병 진단 기법 개발, 신약 개발 기간의 단축, 독성 검사 등이 가능하다. 또한 궁극적으로 한 세포 내의 모든 유전자 및 화학 물질로 구성된 genetic network이 구성된다면 이를 통해 세포 내에서 일어나는 여러 가지 현상들을 시뮬레이션 해 볼 수 있게 될 것이다. 이러한 virtual cell은 독성물질의 예측과 각종 약제의 효능, 유전자 조작 등 각종 실험을 실제 실험이 아닌 컴퓨터 시뮬레이션을 이용하여 그 결과를 예측하게 함으로서 시간과 비용을 크게 단축할 수 있다. 현재 게이오 대학의 E-cell은 그림 10과 같은 구조로 glycolysis, lipid biosynthesis, transcription, translation등에 관여하는 일부 유전자(127개)로 세포를 시뮬레이션하고 있다[37].

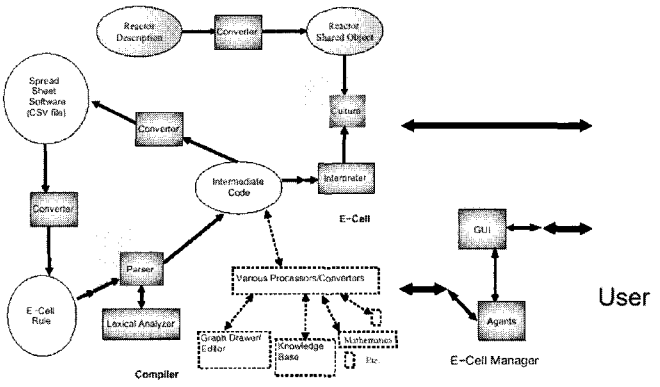


그림 10: E-cell의 시스템 구조도

Pathway와 유전자 네트워크에 관한 데이터베이스로는 KEGG(Kyoto Encyclopedia of Genes and Genomes)가 있다. KEGG는 *E.coli*의 전체적인 metabolic pathway와 regulatory pathway를 구성하려는 시도로서 각 pathway와 enzyme, substrate등의 정보와 서열 정보를 통합한 데이터base이다[38].

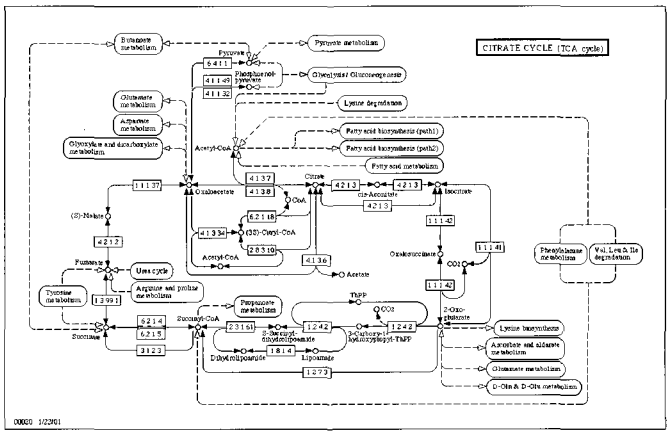


그림 11: TCA cycle의 네트워크 (KEGG)

6. 결 론

과거에는 생물학의 대부분이 지금까지 컴퓨터나 통계와 같은 학문과는 다소 거리가 있는 것이 사실이었다. 그러나 현재의 생물학의 과제들은 컴퓨터의 도움이 없이는 결과를 얻어내기 힘든 상황이 되었다. 이러한 이유로 생물정보학은 최근 급격히 발전하고 관심의 대상이 되고 있다. 생물정보학은 대규모의 생물학 데이터를 수집,관리,분석하여 이를 생물학 뿐만 아니라 의약학, 농학, 환경 등의 문제를 해결하기 위한 유용한 지식을 얻어내기 위한 연구분야로서 현대 생물학과 생명과학의 발전에 필수적인 요소이다.

생물정보학 분야는 아직 외국에 비해 국내의 수준은 크게 떨어지고 있는 것이 사실이다. 하지만 연구 인력의 신규확보와 함께 기존의 연구인력 들이 생물정보학에 관심을 가지고 국내의 특성에 맞는 과제를 선정하고 매진한다면 선진국과의 격차를 줄일 수 있을 것이다.

참고문헌

1. C. Gibas and P. Jambeck, *Developing Bioinformatics Computer Skills*, O'Reilly & Associates, Inc., 2001.
2. G.D. Stormo, Gene-finding approaches for eukaryotes. *Genome Res.* (2000) **10**, 394-397
3. S.L.Salzberg, A.L. Delcher, S. Kasif, and O. White, Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* (1998) **26**, 544-8.
4. M.Q. Zhang, Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* (1997) **94**, 565-8.
5. C. Burge and S. Karlin, Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* (1997) **268**,79-94.
6. D. Kulp, D. Haussler, M.G. Reese and F.H. Eeckman, A generalized hidden Markov model for the recognition of human genes in DNA. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* (1996) **4**, 134-42
7. E.E. Snyder and G.D. Stromo, Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res.* (1993) **21**, 607-13.
8. L.Milanesi, D. D'Angelo, and I.B. Rogozin, GeneBuilder: interactive in silico prediction of gene structure. *Bioinformatics* (1999) **15**, 612-21.
9. M. Borodovsky and J. McIninch, Recognition of genes in DNA sequence with ambiguities. *Biosystems* (1993) **30**, 161-71.
10. E.C. Uberbacher and R.J. Mural, Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* (1991) **88**, 11261-5.
11. S.M. Hebsgaard, P.G. Korning, N. Tolstrup, J. Engelbercht, P. Rouze and S. Brunak, Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* (1996) **24**, 3439-52.
12. A.A Salamov and W. Solovyev, The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* (1997) **5**, 294-302.
13. V. Brendel and J. Kleffe, Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA. *Nucleic Acids Res.* (1998) **26**, 4748-57.
14. T.M. Lowe and S.R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* (1997) **25**, 955-64.
15. D.J. Galas, Sequence interpretation: Making sense of the sequence. *Science* (2001), **291**, 1257-1260.
16. D.L. Black, Protein Diversity from alternative splicing: A challenge for bioinformatics and post-genome biology. *Cell* (2000), **103**,367-370.
17. J.L. DeRisi, R.I. Vishwanath and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* (1997) **278**, 680-686.
18. M.B. Eisen, P.T Spellman, P.O Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* (1998) **95**,14683-14868.
19. M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr and D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* (2000) **97**, 262-267.
20. G.Sherlock, Analysis of large-scale gene expression data. *Curr. Opin. Immunal.* (2000) **12**, 201-205.
21. T.R. Golub, D.K. Slonim, P.Tamayo,C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield and E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* (1999) **286**, 531-537.
22. P. Chou and G Fasman, Prediction of the secondary structure of proteins from their amino acid sequence, *Advanced Enzymology* (1978) **47**, 45-148.
23. J. Garnier, J-F. Gibrat and B. Robson, GOR secondary structure prediction method version IV. *Methods in Enzymology* (1996) **266**, 540-553
24. D.G. Kneller, F.E. Cohen and R. Langridge, Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network. *J. Mol. Biol.* (1990) **214**, 171-182.
25. A.J. Shepherd, D. Gorse and J.M. Thornton, Prediction of the location and type of beta-turns in proteins using neural networks. *Protein Sci.* (1999) **8**, 1045-55.
26. J.A. Cuff, M.E. Clamp, A.S Siddiqui, M. Finlay and G.J.Barton, Jpred: A Consensus Secondary Structure Prediction Server, *Bioinformatics* (1998) **14**, 892-893.
27. D. Frishman and P. Argos, Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* (1997) **27**, 329-35.

28. C. Geourjon and G. Deleage, SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.* (1995) **11**, 681-684.
29. O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen, and S. Brunak, Protein distance constraints predicted by neural networks and probability density functions. *Protein Engineering* (1997) **10**, 1241-1248.
30. N. Guex and M.C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* (1997) **18**, 2714-2723.
31. P.A. Bates and M.J. Sternberg, Model building by comparison at CASP3: Using expert knowledge and computer automation. *Proteins* (1999) **37(S3)**, 47-54.
32. A. Sali and T.L. Blundell, Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* (1993) **234**, 779-815
33. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* (1973) **181**, 223-230.
34. J.M. Bower and H. Bolouri (eds.), *Computational Modeling of Genetic and Biochemical Networks*, MIT Press, 2001.
35. A. Wagner, A computational genomics approach to the identification of gene networks. *Nucleic Acids Res.* (1997) **25**, 3594-3604.
36. R. Somogyi and C.A. Sniegoski, Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Complexity* (1996) **1**,45-63.
37. M. Tomita, K. Hashimoto, K. Takahashi, T.S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J.C. Venter and C.A. Hutchison 3rd., E-CELL: software environment for whole-cell simulation. *Bioinformatics* (1999) **15**, 72-80.
38. M. Kanehisa and S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* (2000) **28**, 29-34
39. *Science* (2001) **291**, Whole issue.
40. *Nature* (2001) **409**, Whole issue.
41. R.Durbin, S.R. Eddy, A.Krogh, G.Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998

장병탁
1986: 서울대 컴퓨터공학과 학사
1988: 서울대 컴퓨터공학과 석사
1992: 독일 Bonn 대학교 컴퓨터과학과 박사
1992-1995: 독일국립정보기술연구원(GMD) 연구원
1995-1997: 건국대학교 컴퓨터공학과 조교수
1997-현재: 서울대학교 컴퓨터공학부 조교수, 부교수
2001-현재: 서울대 바이오정보기술연구센터(CBIT) 소장
URL: http://bi.snu.ac.kr/~btzhang/

강철주
1998: 고려대 생물학과 학사
2000: 고려대 생물학과 석사
2000 - 현재: 서울대 바이오정보기술연구센터(CBIT) 연구원