종 설

예방의학회지 제35권 제2호(2002년 5월)
Korean J Prev Med 2002;35(2):83-91

# 생명정보학과 유전체의학

김주한

서울대학교 의과대학 예방의학교실

## Bioinformatics and Genomic Medicine

Ju Han Kim

Department of Preventive Medicine, Seoul National University College of Medicine

Bioinformatics is a rapidly emerging field of biomedical research. A flood of large-scale genomic and postgenomic data means that many of the challenges in biomedical research are now challenges in computational sciences. Clinical informatics has long developed methodologies to improve biomedical research and clinical care by integrating experimental and clinical information systems. The informatics revolutions both in bioinformatics and clinical informatics will eventually change the current practice of medicine, including diagnostics, therapeutics, and prognostics.

Postgenome informatics, powered by high throughput technologies and genomic-scale databases, is likely to transform our biomedical understanding forever much the same way that biochemistry did a generation ago. The paper describes how these technologies will impact biomedical research and clinical care, emphasizing recent advances in biochip-based functional genomics and proteomics. Basic data preprocessing with normalization, primary pattern analysis, and machine learning algorithms will be presented. Use of integrated biochip informatics technologies, text mining of factual and literature databases, and integrated management of biomolecular databases will be discussed. Each step will be given with real examples in the context of clinical relevance. Issues of linking molecular genotype and clinical phenotype information will be discussed.

*Korean J Prev Med 2002;35(2):83-91*

**Key Words:** Bioinformatics, Genomics, Proteomics, Medical informatics, Microarray

## 서 론

인간 유전체 사업(Human Genome Project)은 그간의 분자생물학 기술발달 성과를 의학 연구에 적용하여 인류 건강과 복지증진에 기여하고자 하는 목적으로 1980년대 말에 시작되었다. 인간 유전체 사업의 일차적인 목표는 인간 유전체를 구성하고 있는 약 30억 개의 염기서열을 밝혀내는 것이다. 하지만 인간 유전체의 전체 염기서열을 확인하는 것은 유전체학의 종결점이 아니라 새로운 후기 유전체 시대(Post-genomic era)의 시작점인 것이다. 인간 유전체 사업은 생명과학 전반에 수많은 변화를 유발하였다. 많은 변화 중에서도 가장 놀라운 것은 생명과학과 정보학의 결합을 통해 생명정보학(Bioinformatics)을 탄생시켰다는 점이다. 초기의 생명정보학은 방대한 유전체 사업을 지원하기 위한 방법론의 형식으로 발달하였다. 그러나 생명을 구성하는 유전체(genome), 전사체(transcriptom), 단백체(proteom) 등에 관한 방대한 정보가 체계적으로 축적되면서 생명정보학은 생명과학 연구의 중심 방법론으로 자리잡게 되었다. 후기 유전체 시대를 이끌 가장 핵심적인 생명과학 연구 방법론이 바로 생명정보학인 것이다.

생명정보학은 아직 그 걸음마기에 있다. 하지만 생명정보학이 PCR 기술처럼 생명공학 연구의 보조수단으로 남는 것이 아니라 하나의 독립된 학제를 구성하고 있는 것은 바로 생명현상이 "진정으로 정보학적 현상"이라는 근원적 사실에 기인하는 것이다. 생명정보학의 급속한 발전은 과거 분류학에 머물던 생물학이, 생명현상을 "물질계의 화학적 상호작용"으로 재조명하면서 유기화학 및 생화학으로 꽃피었고, 다시, 유전자, 단백질 등 생명체에 고유한 거대분자들의 결정론적 특성과 역할에 주목하며 세포생물학 및 분자생물학의 형식으로 발전을 거듭해온 현대 생명과학 발달의 현주소와 그 진화의 맥락을 같이하는 것이다. 생명정보학은 생명현상을 자연계를 구성하는 물질과 에너지간의 고도로 조직화된 정보교류체계의 발현으로 파악한다.

## 생명정보 대량획득기술과 생명정보학

생명정보학의 발달은 정보학 이론의 발달만으로는 불가능하다. 현대 생명정보학의 확립에는 매우 효율적인 생명정보 대량획득기술(High Throughput Technology)이 크게 기여했다. Figure 1의 도표는 유전자 서열정보 획득량과 유전자 관련 논문 발표량의 증가추이를 보여준다. 현재 대부분의 유전자 염기서열이 GenBank에 등록되므로 GenBank의 통계를 보면 현재까지 밝혀진 유전자 서열정보의 양을 알 수 있다. 그림에서 보는 바와 같이 서열이 밝혀진 유전자의 수는 90년대 초반을 기점으로 기하급수적으로 증가하고 있음을 알 수 있다. 이는 전자동

접수 : 2002년 4월 25일, 채택 : 2002년 5월 1일
책임저자 : 김주한 (서울시 종로구 연건동 28, 전화: 740-8320, 팩스: 747-4830, e-mail: juhan@snu.ac.kr)

염기서열분석기와 같은 자동화 정보획득 기기의 발달과 인간 유전체 사업과 같은 대규모 연구사업 지원이 기여한 바 크다. 반면에 메드라인에 등재된 논문중 유전자를 다루고 있는 문헌의 수는 여전히 산술급수적으로 증가하고 있으며 이미 90년대 중반에 유전자 서열정보와 양적인 역전현상이 일어났음을 알 수 있다. 대개의 유전자 관련 논문이 한 개 혹은 많아도 수 개의 유전자에 대한 일차분석을 담고 있다고 가정하면, 도표가 전달하고 있는 메시지는, 현재 일차분석조차 되지 못한 유전자 서열정보가 엄청난 속도로 축적되고 있다는 것이다. 생명정보 대량획득 기술 발달은 더욱 가속화되고 있으므로 이러한 상황은 더욱더 악화될 것임도 자명하다. 생명정보 대량획득기술의 급속한 발달은 하나 하나의 유전자를 연구해 온 기존의 연구방식으로부터의 획기적인 패러다임 전환을 요구하고 있는 것이다.

실제 유전자 정보 폭발의 문제는 도표에 보이는 것보다 더 심각하다. 도표는 단순히 구조유전체학(Structural Genomics)적인 유전자 서열정보 등록량만을 표시한 것이다. 하나 하나의 서열정보마다 그 서열의 기능이나, 전사인자, 모티프, Linkage map, 염기서열 다형성 (예, SNP), 단백질 삼차구조 정보 등 수많은 주석 정보가 생산되고 있으며, 유전체 정보 주석달기도 컴퓨터의 도움으로 자동화되어 가고 있다. 하지만 그보다 더욱 중요한 변화는 바로 기능유전체학(Functional Genomics)의 발달이다. 서열과 같은 구조정보는 한 개체에 한 벌의 정보가 있을 뿐이지만 (예를 들어, 유전자칩으로 포착되는) 기능유전체 정보는 문자 그대로 무한히 많은 상태(예를 들면, 특정 자극을 가한 후 시간의 흐름에 따라, 자극의 정도에 따라, 혹은 세포 조직의 종류에 따라)에 따른 방대한 양의 발현정보를 생산하고 있는 것이다.

현 시점에서 유전체의학을 선도하는 첨병은 스닙(SNP, Single Nucleotide Polymorphism)을 주축으로 한 유전자 다형성 (Genetic Polymorphism) 연구

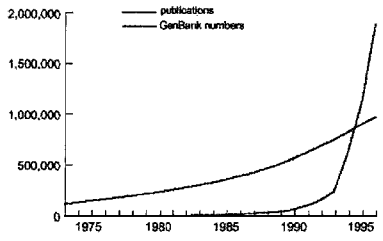


**Figure 1.** Growth of nucleotide sequences and related publications(From NCBI).

와, 유전자칩을 주축으로 한 기능유전체학(Functional Genomics) 연구로 볼 수 있다. 단백체학(Proteomics)의 경우 아직은 전체 단백체의 지도가 규명되지 않았고, 인산화와 같은 다양한 복제후변형(posttranslational modification)에 대한 이해가 부족하며, 세포내 위치에 따른 접힘 등의 복잡한 기전이 잘 알려져 있지 않고, 상보적 염기서열(complementary sequence)에 의한 직접적 정량화가 가능한 유전자에 비해 단백질은 그 정량화가 쉽지 않다는 단점이 있어, 유전체학만큼 발전하려면 다소간의 시간이 더 필요할 것으로 예상된다. 하지만 단백질은 많은 생명현상의 최종 발현의 매개체이며, 약물작용의 과녁 물질이고, 현재 단백질칩 연구 또한 활발하여, 빠른 시간 내에 큰 성과를 보게될 것으로 기대된다.

유전자칩과 함께, 2-D PAGE, 단백질칩, 혹은 대사플럭스 분석과 같은 소위 "대량병렬형" 생명정보 획득기술(massively-parallel data acquisition technology)들은 생명체를 완벽한 하나의 시스템으로 다루어 기능을 연구하는 것을 가능케 하는 주요 기술들이다. 유전자칩으로 상태의 변화에 따라 세포내 모든 유전자가 오케스트라 연주처럼 시시각각 발현되는 상황을 정량화하는 것이 가능해져, 10만 유전자의 활동이 10분 간격의 스냅사진처럼 낱낱이 드러나고 있다. 로제타 인파마틱스사는 효모의 모든 유전자를 넉아웃하여 살아남은 200여종의 유전자 발현 프로파일 디비를 구축하여 상용화하였고 다시 일반에 공개하였다 [1].

Ideker 등은 효모의 갈락토스 회로에 관여하는 모든 유전자를 하나하나 넉아웃하며 유전자발현 프로파일을 얻어 잘 알려진 갈락토스 사이클에서도 새로운 피드백 기전을 찾아낼 수 있음을 보고하였다 [2]. 현재 대표적인 유전자 발현정보의 공공 데이터베이스 구축은 NCBI의 GEO(Gene Expression Omnibus), NCGR의 GeneX, EBI의 ArrayExpress 등을 들 수 있다. 천문학은 바빌로니아인들이 하늘과 별의 지도를 완성한 데서 시작되었다 한다. 화학주기율표의 첫 완성에서와 같이, 전체 유전자의 지도 작성에 따른 막대한 정보폭발은 단순한 양적 변화를 넘어 생명현상 연구 정보처리 패러다임의 변증법적 전환을 예시하고 있는 것이다.

## 생명과 서열정보

생명의 핵심 성분인 단백질이 일반 텍스트 문자열과 같은 1차원 서열구조를 소유한다는 최초의 가설은 1883년 Curtius [3]가 제시했다. 좀더 진보된 펩타이드 가설은 Hofmeister와 Fisher가 제기했다. DNA를 최초로 발견한 Meischer [4]는 유전물질이 단순 텍스트와 같은 화학 기호들의 서열정보일 것임을 예견하였다. 1961년 최초의 tRNA 분절의 염기서열이 출판된 후 수많은 DNA 염기서열 발표가 이어졌다. 이론적으로는, 생명체 DNA 유전형에서 개체의 표현형 발현에 이르는 유전 정보의 흐름은 DNA를 구성하는 염기들의 순수한 서열정보 속에 코드화되어 있는 것이다. 또한 단백질의 2, 3차 구조 결정뿐 아니라, 프로모터, 인핸서 등 각종의 유전자 제어기제, 제한효소 작용부위, 스플라이스 부위, 나아가 돌연변이와 진화 전과정의 역사 등도 모두 서열정보 속에 코드화 되어있는 것이다. 생명의 신비를 밝혀온 현대 분자생물학의 눈부신 발전은 생명현상이 놀라울 정도로 '정보학적' 임을 밝혔다.

생명현상의 정보학적 연구의 가장 고전적인 예로는 서열상동성(sequence homology) 분석을 들 수 있다. Zuckerkandl과 Pauling [5]은 유사 단백질의 펩

타이드 서열 변이정도가 진화단계와 관련됨을 밝혀 분자진화학(Molecular Evolution)이라는 전혀 새로운 분야를 창시했다. 서열정보의 변이는 일정한 규칙을 가지고 일어나므로, 변이의 정도를 정량화하면 유전자 내에 기록된 분자시계(molecular clock)를 통해 진화의 역사를 추적할 수 있다. 그러므로 분자진화학의 도입을 통해 진화라는 고전적인 생명과학의 문제는 유전물질에 코드화된 정보 변화의 정량적 연구라는 정보학의 문제로 변환되는 것이다.

서열 상동성 비교 문제는 순수한 두 문자열의 비교라는 순수 이산수학(Discrete Mathematics)의 문제이다. 아주 간단한 예를 들어 "BLUE"와 "BILE"이라는 두 서열이 있다고 하자. 진화론적으로는 몇 번의 염기서열 삽입(insertion), 삭제(deletion), 치환(mutation) 과정을 거쳐 "BLUE"가 "BILE"로 (혹은 "BILE"이 "BLUE"로) 진화한 것으로 생각할 수 있고, 그 변환 거리가 분자진화론적 진화거리인 것이다. 즉,

"BLUE" -> {series of insertion or deletion or mutation} -> "BILE"

그런데, 간단한 예를 들어, "BLUE"의 마지막 "E"가 삭제, 삽입, 삭제, 삽입을 계속 무한 반복하면서도 여전히 "BLUE"에 머무를 수 있다. 그러므로 "BLUE"와 "BILE"간의 변환경로의 가짓수는 무한히 많다. 즉 문제는 "BLUE"와 "BILE"간의 무한히 많은 변환경로 중에서 최소변환경로를 찾는 것이다. 이는 최소변환경로가 실제 자연계에서 일어난 사건에 가장 가까운 변환경로일 것이기 때문이다. 요약하면 서열 S1과 서열 S2 간의 진화거리는 {insertion, deletion, mutation}라는 편집작업을 통해서 서열 S1을 서열 S2로 변환시키는 최소편집거리 (Minimum Editing Distance, Levenstein Distance)를 구하는 문제이다.

결과부터 먼저 이야기하면, 아래와 같이 편집거리가 2인 등가의 두 해가 존재하며 "BLUE"와 "BILE"의 진화거리는 2이다.

| D(i,j) | | B | I | L | E |
|---|---|---|---|---|---|
| | 0 | ← 1 | ← 2 | ← 3 | ← 4 |
| B | ↑ 1 | ↖ 0 | ← 1 | ← 2 | ← 3 |
| L | ↑ 2 | ↑ 1 | ↖ 1 | ↖ 1 | ← 2 |
| U | ↑ 3 | ↑ 2 | ↖↑ 2 | ↖↑ 2 | ↖ 2 |
| E | ↑ 4 | ↑ 3 | ↖↑ 3 | ↑↖ 3 | ↖ 2 |

B I L - E
B - L U E

B I L E
B L U E

**Figure 2.** Measuring minimum editing distance by dynamic programming. Among the 3 possible pathways from upper (deletion), left (insertion), and upper-left (match or mutation) cells to each cell, D(i,j), only the optimum paths are marked. For example, although there are 3 possible paths to the cell defined by B and B, only the diagonal path for match or mutation was selected, the score of which is 0 because B and B are the same characters (i.e., they are matched). The score for the cell defined by B and I can be 1 with insertion from the left cell with score 0, 2 with mutation from the upper-left cell with score 1, or 3 with deletion from the upper cell with score 2. Just repeating the steps will fill up the whole table and reveal the two paths for global match with score 2.

한번의 삽입과 한번의 삭제가 일어난 경우 (편집거리 = 2)

B I L - E
B - L U E

두 번의 치환(I와 L, L과 U)이 일어난 경우 (편집거리 = 2)

B I L E
B L U E

물론, 실제에서는 단일 염기뿐 아니라, 긴 염기서열이 한 번에 삽입 또는 삭제되는 경우도 고려해야 하고, 삽입, 삭제, 치환이 일어나는 상대적 확률이 서로 다르다는 점도 고려하여 좀더 정교한 모델을 만들어야 한다. 수학적 일반 해를 구하기 매우 어려운 이 문제는 경로탐색(search) 기법의 하나인 동적프로그래밍(Dynamic Programming)으로 해결할 수 있다. Figure 2는 동적프로그래밍을 통해 가능한 모든 경로를 추출한 매트릭스이다. 수평, 수직 화살표는 삭제 혹은 삽입을, 그리고 대각선 화살표는 치환을 의미한다.

최소편집거리 탐색 문제는 문제의 크기가 커지면 계산량이 기하급수적으로 증가하여 정확한 계산이 불가능해지는 소위 NP(Non-deterministic Polynomial)-hard한 문제이므로, 약간의 해석학적 방법들을 적용하여 효율적으로 근사치를 찾는다. 좀더 진보된 알고리즘들은 유사한 아미노산간의 치환과 상이한 아미노산간의 치환 점수를 차등화한 점수표(scoring matrix)를 적용하거나, 해싱테이블을 이용한 성능 향상(FASTA), 혹은 은닉마르코프모델 등을 적용한 정교한 전이 (삽입, 삭제, 치환) 확율모델을 적용한다. 이러한 서열유사성 분석은 분자진화론적 계층구조를 밝히고, 모티프 찾기, 전사인자 결합부위 찾기, 유전자 조절네트웍 재구성, 스플라이싱 부위 찾기, 인트론 내 조절부위 찾기 등의 기본 원리이다. 스미스-워터만 알고리즘을 제공하는 Bic-sw Database(http://www2.ebi.ac.uk/Bic-sw/), 혹은 FASTA(http://www2.ebi.ac.uk/fasta3/)나 NCBI/GenBank의 BLAST(http://www.ncbi.nlm.nih.gov/BLAST/) 등을 사용하면 서열간 상관정도가 높은 수많은 유전자나 단백질 검색을 시험해 볼 수 있다.

샷건 방법의 개발로 사전지식 없이 기

계적인 서열결정이 가능해졌다. 즉, 먼저 유전물질을 분리해낸 후 그 유전자 서열을 결정하고 기능 규명을 추구하던 고전적 분자생물학 패러다임이, 반대로 유전체 서열(Genomic Sequence) 전체를 기계적으로 먼저 결정해 놓고, 그 서열구조 정보학적 분석을 통해 유전자, 유전자조절부위, RNA 등을 찾아내는 생명정보학 패러다임으로 전환된 것이다. 가장 대표적인 예로서 유전자 찾기(Gene finding)의 경우, 먼저 6개의 해독프레임(reading frame, 3(codon) x 2(strands) = 6)을 읽어 소위 충분히 긴 해독개시프래임(ORF, Open Reading Frame)을 찾는 간단한 알고리즘으로 시작된다. 물론 진핵생물의 경우 인트론과 액손 등의 복잡한 관계로 좀더 복잡한 모델링 문제가 된다.

중요한 점은 유전자 찾기와 같은 유전체학의 주요 문제들이 서열구조의 패턴 분석 문제, 즉, 전통적인 정보학의 문제라는 점이다. 실제로 Grail(http://compbio.ornl.gov/grailexp/)과 GenScan(http://genes.mit.edu/GENSCAN.html)은 각각 인공신경망 및 은닉마르코프모델을 이용한 패턴분석 알고리즘이다. 흔히 국소 패턴분석과 전역패턴 분석을 수리적으로 (Ab Initio) 수행하고, 앞서 설명한 서열유사성 (sequence homology) 검색을 통해 정보학적으로 추론된 서열 부위가 유전자일 가능성을 확률론적으로 추론한다. 기존에 밝혀진 유전자와의 서열 유사성은 유전자 존재의 매우 강력한 증거이므로, 향후 유전체 분석이 증가할수록 서열 데이터베이스는 향상되고 이러한 방법론들의 유용성은 더욱 커질 전망이다.

Pipas와 McMahon은 RNA 염기서열이 밝혀지기 시작한 1970년대 초에 그 2차 구조를 1차 서열정보만으로 예측할 수 있음을 시사했다 [6]. Vienna RNA-Package(http://www.tbi.univie.ac.at/~ivo/RNA/)를 다운로드 받아 뉴클레오타이드 서열을 입력하면 최적의 2차 구조를 예측해준다. 주로 패턴 분석과 에너지 최적화 알고리즘 등이 활용된다. 단백질 구조 추론은 훨씬 더 어려운 문제이다. 아직까지 PDB나 SWISS-PROT 등은 X선 크리스탈로그래피 등 실험적 증거에 의한 구조만을 인정한다. 그러나 CASP(Critical Assessment of technology for protein Structure Prediction, http://predictioncenter.llnl.gov/)의 예에서와 같이 생명정보학적 방법에 의한 단백질 구조예측 분야도 매우 빠르게 발전하고 있고, 단백질 상호작용이나 유전자 조절 네트웍 같이 매우 복잡한 분야로 급속히 옮겨가고 있다.

## 바이오칩 기술의 발달

유전자칩에 대한 상세한 리뷰는 관련 논문들에 잘 정리되어 있다 [7-11]. 유전자칩의 학문적, 상업적 성공에 힘입어 세포칩, 단백질칩 등 다양한 바이오칩 개발 연구가 활발하다. 그러나 바이오칩의 발전은 기술적으로는 질적인 발전이라기보다는 양적인 변화이다. 유전자칩은 Reverse Dot Blot과 같은 기존의 유전자 검출기법을 대규모로 집적하여 병렬화한 대량획득기법(parallelized high throughput technology)에 불과하다. 하지만 중요한 것은 패러다임 전환이다. 생명체가 가진 유전자의 수가 유한하므로, 유전자 관찰능력의 양적인 팽창이 유전체 전체를 조망할 수 있는 단계에 이르르면, 이제 양적인 발전은 변증법적으로 질적인 변화를 유발하는 것이다. 바이오칩의 도입은 유전자 연구를 유전체 전체로 확장하였고, 유전자 검출기법을 (현존하는 몇가지 기술적인 제한점에도 불구하고) 정성적 분석에서 정량적 분석으로 바꾸었다. 이는 고전적으로 화학적, 정성적 방법론에 의존하는 생물학적 유전자 연구에, 수리적, 정량적 방법론이 도입됨을 의미한다. 즉 바이오칩의 도입으로 말미암아, 분자생물학적 유전자 발현연구는 발현량을 담은 (형광) 스캔이미지의 정량적 패턴 분석으로 변환된 것이다.

유전자칩은 사용하는 검출용 뉴클레오타이드에 따라 cDNA(200-500 bp) 칩과 올리고뉴클레에타이드(15-100 bp) 칩으로 나눌 수 있고, 제작 방법에 따라서는 핀마이크로어레이나 잉크젯 등의 로봇 프린팅 칩과 반도체 제작 공정을 이용한 어피메트릭스사의 광식각 칩으로 나눌 수 있다. cDNA 칩은 이름 그대로 ORFs(Open Reading Frames, i.e. genes) 혹은 EST(Expression Sequence Tags)의 전 염기 서열을 슬라이드에 부착시킴으로서 상보서열을 소유한 해당 유전자를 고유하게 식별한다. 어피메트릭스사의 올리고뉴클레에타이드 칩은 광식각 기술로 올리고뉴클레오타이드를 작은 유리판 위에 한 층 한 층 직접 합성한다. 이론적으로 25개의 염기서열로 이루어진 25mer의 경우 이론적으로 $4^{25}$개의 유전자를 고유하게 식별할 수 있다. 로봇 프린팅 방식의 칩이 보통 1-2만 개 정도의 염기서열을 집적할 수 있음에 비해 광식각 기술은 현재 약 50만개의 올리고뉴클레오타이드를 한 장의 슬라이드에 합성할 수 있다. 최근에는 cDNA 대신 60-100mer의 합성 올리고뉴클레오타이드를 로봇 프린팅 방식으로 집적하는 신기술도 주목받고 있다. 어피메트릭스사의 짧은 25mer의 단점과 비싼 가격, cDNA의 단점인 클론 라이브러리 관리상의 어려움을 극복하기 위해 70mer 정도의합성 올리고뉴클레오타이드 라이브러리를 매우 저렴한 가격에 공급한다. 유전자칩은 이제 더 이상 고가의 연구 장비가 아닌 것이다.

로봇 프린팅 방식의 칩은 흔히 두 검체를 (i.e., 연구검체와 대조검체) 각각 다른 색으로 염색하여 경쟁적 결합(competitive binding)을 일으키는 이중염색 기법(Two-dye technique)을 많이 사용한다. 연구검체에는 적색 형광물질을 부착하고 대조검체에는 녹색 형광물질을 부착하여 두 형광물질의 발색강도의 적/녹 비율에 따라 mRNA의 발현량을 정량화한다. 발색강도는 이미지로 촬영된 후 이미지 분석기에 의해 수치값으로 변환된다.

어피메트릭스사의 칩은 레이저 판독기로 직접 발현 강도를 측정한다. 각 올리고뉴클레오타이드(Perfect Match)와 나란히 25개 염기서열의 중앙점인 제13번 염기서열을 변형시킨 (Mis-Match) 올리고뉴클레오타이드를 나란히 배열, 결합량을

비교함으로서, 교차결합에 의한 잡음을 제어한다. 그러므로 어피메트릭스사의 칩은 로봇 프린팅 칩과 달리 두 샘플의 경쟁적 결합의 상대 비율이 아닌 단일 샘플의 절대 발현량을 측정한다. 2001년 말에 어피메트릭스사는 자사의 웹페이지를 통해 25mer 올리고뉴클레오타이드의 염기서열을 공개했다.

유전자칩 패러다임의 성공에 힘입은 바이오칩 신기술의 발전은 매우 급속히 이루어지고 있다. 단백칩과 미세거울을 이용한 올리고칩 뿐 아니라, 최근에는 이중나선 DNA칩으로 DNA-단백질 상호작용을 측정하는 기술이나, 세포칩 등 수많은 신기술이 하루가 다르게 출현하고 있다.

## 기능 유전체학

1994년 4월에 효모 유전체 전체 서열이 분석되고 나서 과학자들을 가장 당혹하게 한 것은, 염기서열을 밝혀낸 6200개 전체 유전자중에서 그 기능을 추측이나마 해볼 수 있는 것이 전체의 4분의 1도 되지 않는다는 사실이었다 한다 [12]. 특히 효모는 각종 유전 연구의 모델 생물로서 거의 모든 유전학적 기법을 동원하여 연구되어 당시까지 유전학자들은 적어도 효모의 유전자에 관한 한은 알만큼은 안다고 자부했었기에 이러한 결과는 더욱 충격적인 것이다. 이는 또한 유전체 연구는 개개 유전자를 연구하여 전체를 재구성하고자 하는 분석적 분자생물학의 패러다임만으로 불충분하며 전체 유전체 수준의 연구가 필요함을 강하게 정당화하는 중요한 발견인 것이다.

현재까지 약 60종 정도의 생명체의 유전체 염기서열이 밝혀져 있고, 유전체 사업이 진행중인 것이 약 200종 이상이다. Table 1에서와 같이 그 기능이 알려져 있는 유전자의 비율은 약 40% 전후라 한다. (물론 알려진 기능들조차도 매우 단편적인 것들이다.) 기능유전체학은 인간 유전체 사업 관련 연구들이 주로 서열분석과 같은 구조적 측면에 주력한 구조유전체학이었음과 대비되어 명명된 새로운 분야이다. 가장 간결한 의미에서의 기능유전체학은 이러한 각각의 밝혀진 염기서열(유전자, ORF's)의 기능을 밝히는 분야이다. 달리 말하면, 각각의 밝혀진 염기서열들이 어떤 단백질을 코딩하고 있으며, 어떠한 유전자 조절기능의 제어를 받고 있으며, 어떠한 약물이나 환경자극하에서 어떻게 활성화되며, 다른 유전자들과 어떻게 상호작용하며, 개체변이와는 어떻게 관련되며, 질병 치유 등 다양한 생명현상에 어떻게 관여하는가와 같은 것들을 직접 다루는 분야이다. 유전자의 기능을 밝힐 경우 신약개발, 질병 진단 등 그 응용 범위가 매우 넓을 뿐 아니라, 그에 대한 특허권을 청구할 수 있기 때문에 현재 기능유전체학 연구 경쟁은 유래 없이 치열하다. 이것이 한 민간 기업이 지나치게 많은 특허를 독점해가고 있는 것을 방지하기 위해서 미국연방정부가 대규모의 연구비를 주요 대학과 연구소에 지원하며 민간기업과 경쟁을 벌이고 있는 웃지 못할 사연이기도 하다.

**Table 1.** Genomic sequences with the genes of known and unknown functions (From Church et al., 1997 Science 277:1433)

| Organism | #Genes | % Unknown Function |
|---|---|---|
| S cerevisiae | 6034 | 49% |
| E coli | 4288 | 38% |
| B subtilus | 4000 | 42% |
| Synechocystis sp. | 3168 | 56% |
| A fulgidus | 2471 | 52% |
| H influenzae | 1740 | 42% |
| M thermoautotrophicum | 1855 | 56% |
| H pylori | 1590 | 43% |
| M jannaschii | 1692 | 54% |
| B burgdorgeri | 863 | 42% |
| M pneumoniae | 677 | 51% |
| M genitalium | 470 | 31% |
| Total | 28848 | 47% |

기능유전체학의 연구가 수년 내에 획기적으로 발전할 것임은 명약관화한 일이다. 현재 가장 많이 응용되고 있는 연구방법론은 유전자칩을 이용한 유전자 발현 연구이다. 즉 미지의 유전자를 포함한 수많은 유전자의 활성도(mRNA expression level)를 동시에 측정하여, 특정 조건 전후의 발현 상태를 비교하거나, 서로 다른 조직 혹은 이웃한 조직에서의 발현 프로파일을 비교하거나, 다른 종에서 관련 유전자의 발현을 비교하거나, 시계열 발현 프로파일을 분석하는 등의 다양한 방법에 의해서 유전자 정보를 대량 처리하여 새로운 지식을 찾아내는 것이다.

단순하게 세포에 특정한 처치를 한 후, 조건 전후의 유전자 발현비율(intervention fold difference)을 분석하여 실험조건간의 발현량의 차이가 유의한 유전자들(differentially expressed genes)을 찾는 것만으로도 매우 흥미로운 결과를 얻을 수 있다 [13, 14]. 좀더 체계적인 방법은 그림 5에 도시된 바와 같이 기능성 클러스터 분석을 수행하는 것이다. 클러스터 분석은 마치 밤하늘의 별들을 관찰해서 "우주가 은하계들로 이루어져있음"을 밝혀내어 그 은하계들을 가려내고 다시 "은하계는 수많은 태양계로 이루어져 있음"을 밝혀 가는 작업에 비유될 수 있는 탐색적 자료분석(Explorartory Data Analysis)의 하나이다. 현재 클러스터 분석의 주된 전략은 다양한 실험조건에서의 발현패턴에 따른 유전자 클러스터들을 찾아 조건변화와 무관히 '강하게 동반조절되고 있는 유전자군(tightly co-regulated genes)을 찾는 것이다. 클러스터 분석이 동반조절 유전자군을 찾아주지만 유전자 기능을 다 밝혀주는 것은 아니다. 오히려 클러스터 분석은 미지의 유전자가 어떤 유전자들과 상관되어 있는가 하는 단초를 제시하여, 복잡한 관찰로부터 새로운 가설을 제시함으로서 (data-driven hypothesis generation), Figure 3에 표시된 것과 같이 다음 실험을 설계하는 시간과 노력을 줄여주는 것이다.

클러스터 분석은 크게 계층적 클러스터링과 분할형 클러스터링으로 구분된다. Spellman 등 [15]은 계층적 클러스터 분석을 사용하여 유전자 발현의 주기성을 분석하여 효모에서 세포 분열 주기에 관여하는 유전자 800여 개를 새로 찾아내었다. 이는 짝지은 유전자 발현 프로파일(이 경우는 시계열 자료) 사이의 피어슨 상관계수(pair-wise correlation coefficient)를 구해서, 가까운 것끼리 묶어서 작은 클러스터를 만들고, 결합 역치를 조금씩 낮추어 가면서, 점점 더 큰 클러스터들로 묶어 올라가, 결국 전체를 하나의 수형도 구조로 만드는 (Bottom up) 매우 간단한 알고리즘이다 [16](Figure 4). 공개 소프트웨어는 http://rana.stan-ford.edu/software/에서 구할 수 있다. Butte와 Kohane [17]의 Relevance networks는 정 반대의 알고리즘을 이용한다. 이는 역치기반 클러스터링(Threshold-based Clustering)이라 불리우는 방법으로 먼저 완벽한 N by N correlation matrix를 만든 후에 특정 역치 이하의 링크를 모두 삭제하면 소위 'naturally emerging cluster'들을 찾을 수 있다. Figure 5는 relevance networks를 이용해 상관된 유전자 네트웍을 찾아낸 예이다. Butte 등은 피어슨 상관계수가 선형상관관계만을 찾아주는 한계점을 극복하기 위해 정보이론에서 유래한 상호엔트로피(mutual information)를 활용하여 복잡한 상관관계도 찾고자 하였다.

계층적 클러스터링이 자료구조를 계층적 수형도로 조직화함에 반해 분할형 클러스터링은 자료를 몇 개의 클러스터로 직접 분할한다 (Top down). Church 등은 *K*-means 클러스터 분석을 적용했고 [18], Tamayo 등은 인공 신경망 기법의 하나인 SOM(Self-Organizing Maps) [19]으로 의미 있는 유전자 클러스터들을 찾을 수 있음을 제시했다 [20](Figure 6). SOM의 적용을 위한 공개 소프트웨어는 http://waldo.wi.mit.edu/ MPR/ software.html에서 다운로드받을 수 있다. 알고리즘적으로는 순차적 *K*-means 알고리즘에 몇 가지 보완을 가한 것으로도 해석할 수 있다. 이러한 알고리즘들이 한번에 하나씩의 데이터씩만 고려한다는 한계점을 극복하기 위해 고안 된 Kim 등의 매트릭스 분할법도 소개되었다 [21, 22].

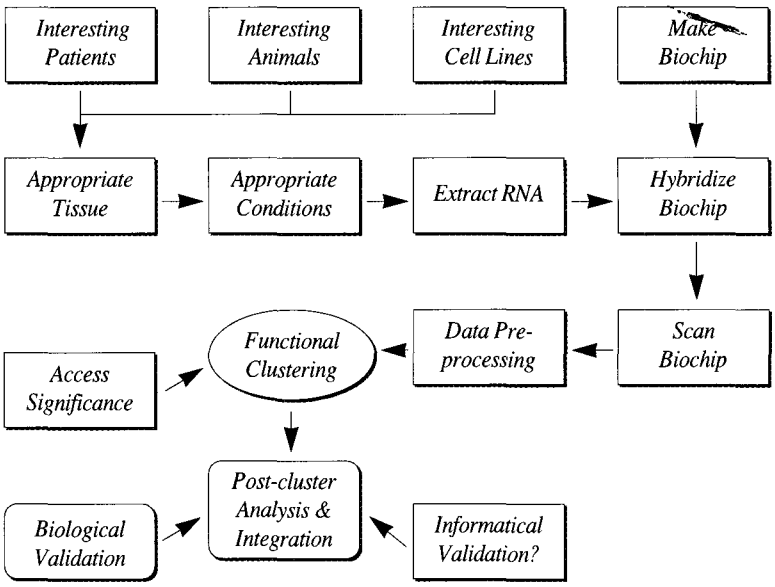


**Figure 3.** A strategy for functional genomics with biochip informatics.

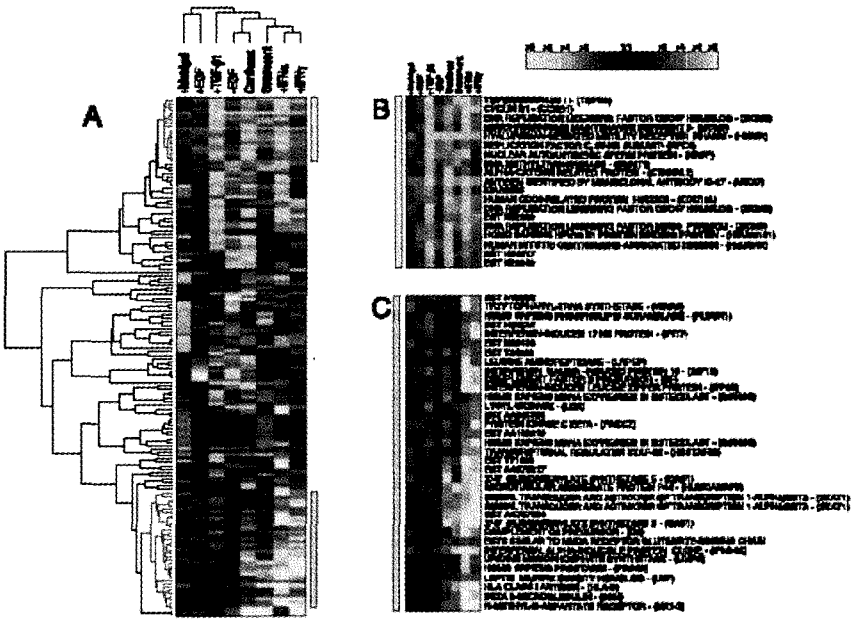


**Figure 4.** Hierarchical tree clustering.

클러스터 분석과 같은 탐색적 자료분석 기법은 복잡한 관찰결과에 대한 선지식(a priori knowledge)이 없는 경우에 적합한 방법이다. 이러한 기법은 인공지능 분야에서는 비감독 기계학습 (Unsupervised Machine Learning)으로 분류된다. 복잡한 생명 현상에 대한 지식이 증가하면서, 감독 기계학습 (Supervised Machine Learning) 기법도 빠르고 성공적으로 기능유전체학 분야에 도입되고 있다. 유전체 발현의 감독-비감독 기계학습 기법에 의한 분석뿐 아니라 다양한 정보학적 기법의 체계적인 통합과 자동화도 기능유전체학 연구에 빠르게 도입

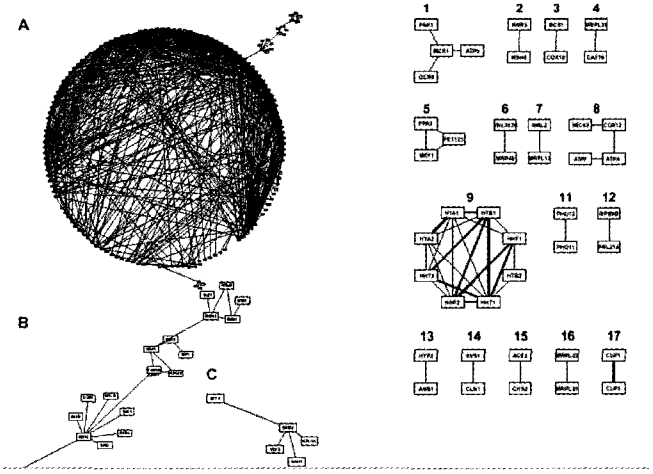

**Figure 5.** Clustering expression profiles by Relevance Networks (from Butt & Kohane, 2000)

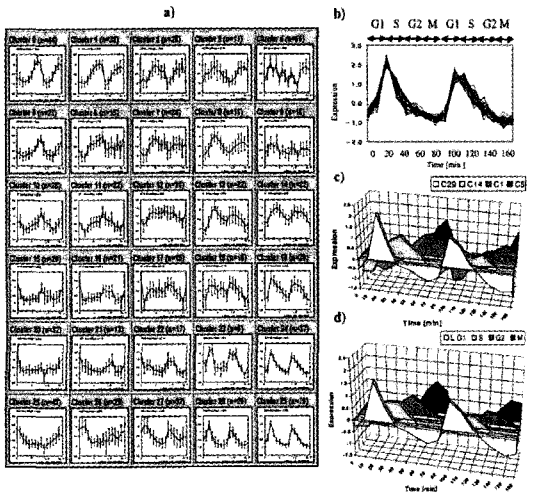

**Figure 6.** Partitional clustering of yeast cell division cycle data by Self-Organizing Maps(SOM). (From Tamayo et al., 2000)

되고 있다. 예를 들어 퍼브진(PubGene, http://www.pubgene.org/)은 텍스트 마이닝 기법을 활용하여 유전체 발현 정보를 메드라인과 같은 문헌정보와 통합해 낸다 [23].

다양한 메타 데이터베이스가 개발되고 있으며 자연어처리 기법도 생명-의료 분야의 문헌 및 사실 데이터베이스로부터 유전자 조절 상호작용 네트웍을 재구성하는데 유용하게 이용된다 [24].

KEGG(http://www.kegg.org/)와 같은 대사 네트웍의 데이터베이스의 빠른 객체화와 조직화도 눈여겨 볼만하다. 구조유전체학적 서열정보도 유전자 기능연구에 유용한 정보를 제공한다.

## 생명 · 의료정보학 (Biomedical Informatics: Emergence of New Medicine)

생명정보학의 태동기에서 최신 동향까지를 주마간산 격으로 살펴보았다. 생명과학과 컴퓨터과학의 이질적 분야를 오가며 설명해야 하는 난점을 피하기 위해, 가능한 한 많은 실례를 들어 이해를 돕고자 하였다. 실례의 선택이 소개자의 주관이 많이 개입되어, 이 글이 생명정보학 전반에 대한 포괄적인 종설이 되기에 부족함을 밝힌다. 이 글은 매우 방대하고 빠르게 변화하는 생명정보학을 현시점에서 순간 포착한 스냅사진 정도에 해당된다. 특히 스닙, 단백체학, 대사체학 및 유전자 조절 네트웍 재구성 및 신호전달체계 분석등의 등의 매우 중요한 분야들과, 실제 기계학습기법들을 자세히 소개해 드리지 못한 아쉬움이 남는다. 신규 논문도 이미 출판시점이면 낡은 것이 되어 버리는 생명정보학 분야의 연구 동향을 파악하기에 제일 좋은 방법은 관련 학회를 돌아보는 것이다. 3대 학회라 할 수 있는 것은 ISMB (Intelligent Systems for Molecular Biology), PSB (Pacific Symposium on Biocomputing), 그리고 RECOMB (Currents in Computational Molecular Biology) 등(http://www.iscb.org/)이다.

의학에 있어 유전체 발현분석은 임상의학과 의학연구에 바로 손쉽게 적용할 수 있는 매우 강력한 도구이다. 의학의 제 영역을 크게 진단, 치료, 예후판정, 그리고 의학 지식체계 관리로 나눌 수 있을 것이다. 유전체 발현 분석이 암의 진단에 적용될 수 있고, 나아가 새로운 아형 발견 및 예후판정에 활용될 수 있음이 보고되었으며 신약개발에도 본격적으로 활용되고 있다. Golub 등은 6,817 개의 유전자 발현 패턴 분석만으로 백혈병의 아형인 AML와 ALL을 판별할 수 있음을 보고하였다(Molecular Classification) [25]. Alizadeh 등은 Diffuse Large B-cell Lymphoma (DLBCL)의 유전자 발현을 클러스터 분석하여 두 개의 새로운 아형이 존재함을 발견하였다 (Class Discovery by Pattern Analysis) [26].

특히 생존률 분석을 통해 새로 규명한 아형중 germinal center B-cell 과 유사한 유전자 패턴을 보이는 DLBCL이 activated B-cell과 유사한 패턴을 보이는 DLBCL군보다 현저하게 높은 생존율을 보임을 보고하여 (Prognostic Prediction) [26] 미래의학의 패러다임이 영구히 변환될 것임을 예견하였다. 이종 암세포가 상이한 병태생리학적 특성

(behavior)을 보이는 것과, 상이한 형태학적 특징(morphology)을 보이는 것도 결국은 그 세포들이 서로 다른 유전자 발현 패턴을 보이는 것에 기인함은 명백한 것이다. 전통적 형태학 연구에서 면역조직병리학이나 세포표식자 연구 등으로 발전하고 있는 병리학의 발달과정과 비교한다면, 유전체 발현 패턴 연구는 결과적으로 유전자라는 수만 개의 매우 중요한 세포표식자를 동시 검출하는 것에 비유될 수 있다. 그러므로 유전체 발현연구가 진단 분류와 같은 병리학 분야에 적용되는 것은 전혀 놀라운 일이 아니다. 이러한 유전자 발현패턴에 따른 세포 분류학은 매우 빠르게 발전할 것으로 기대된다. 기존의 신약 개발들이 주로 과녁 단백질을 대상으로 하여 그에 작용하는 후보 물질을 연구한 것이라면, 기능유전체학을 통한 연구는 대상 대사 경로에 관여하는 유전자와 그 유전자의 발현 경로를 과녁으로 하여 그에 작용하는 후보 물질들을 추적하는 것이다. 즉 약 500개 정도인 기존의 약물작용 대상 물질이 수만개의 유전자로 확대되는 것이다. 더욱이 이러한 연구들은 이미 축적된 방대한 데이터베이스를 기반으로 데이터마이닝 등의 생명정보학적 패러다임을 사용한다.

생명 · 의료정보학(Biomedical Informatics)은 분자생물학과, 컴퓨터과학과, 임상의학의 제 지식이 삼위일체를 이루어가며 미래 생명과학을 주도할 핵심 학문이다. Alizadeh 등 [25]이나 Golub 등 [26]의 종양의 병태생리 연구에 새로운 지평을 여는 연구도 관련 의료기관들의 훌륭한 임상 정보 시스템이 없었다면 불가능했을 것임은 자명한 일이다. 생명과학 연구의 최종산물의 최대 수요처도 다름 아닌 임상의학분야이다. 유전체 수준의 체계적인 생명정보와 방대한 의료정보 시스템의 통합에 관한 논의가 활발하다 [27].

최근 스탠포드와 컬럼비아 대학교에서 기존의 의료정보학과정에 생명정보학 프로그램을 통합했다. 선진국에서도 생명 · 의료정보학 연구자를 구하기가 매우 힘들다. 컴퓨터나 수학, 물리학이라면 기피증부터 보이는 생명과학도와, 한 번 관심을 갖고 문을 두드렸다가도 생명과학의 방대한 지식량에 질려서 달아나는 컴퓨터 공학도간의 이질적인 학제와 교육전통의 문제가 심각하다. 두 분야 모두에 대한 깊이 있는 지식을 갖춘 연구인력을 양성할 수 있는 여건을 갖춘 연구기관이 매우 드물다는 것도 심각한 문제이다. 대량병렬형 생명정보 획득기술과 생명정보학의 발달은 생명현상의 시스템적 통합을 가능케 하여 의학과 생명공학 연구의 패러다임을 영구히 변환시킬 것이다. 소위 "Omic Revolution"(i.e., genomics, transcriptomics, proteomics, metabolomics, physiomics, and biomics)이 모든 생명체 구성단위의 수평적 통합을 이끈다면, 생명의료정보학(i.e., bio-molecular informatics, computational cell biology [28],computational physio logy [29], digital anatomy [30],chemoin formatics [31, 32], clinical informatics [33],and public health informatics [34])은 생명과질병현상의 분자론적 미시수준에서 인간과 건강사회의 거시수준에 이르는 수직적 통합을 이끌고 있다. 생명현상의 분자론적인 이해와 정보학적인 통합이라는 씨줄과 날줄의 섬세한 조직화를 통해 다가올 미래의학의 모습을 그려볼 수 있다.

## 참고문헌


1. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakrabortty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. *Cell* 2000; 102(1): 109-26
2. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 2001; 292(5518): 929-934
3. Curtius T. Ueber das Glycocoll. *Chem Ber* 1883; 16: 753-757
4. Meischer JF. *Med-Chem Unter* 1871; 441
5. Zukerkandl E, Pauling L. Molecular disease, evolution, and genic hetero geniety. In Kasha, M. and Pullman B Eds. Horizons in Biochemistry Academic Press, New York; 1962. p. 189-225
6. Pipas JM, McMahon JE. Method for predicting RNA secondary structure. *Proc Natl Acad Sci U S A* 1975; 72(6): 2017-2021
7. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a cDNA microarray. *Science* 1995; 270: 467-470
8. Shalon D, Smioth SJ, Brown PO. A DNA micro-array system for analysing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996; 6: 639-645
9. Pease AC, Solars D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A* 1994; 91: 5022-5026
10. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol* 1996; 14(13): 1675-1680
11. DeRisi JL, Iyer V, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997; 278: 680-686
12. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat Genetics Suppl* 1999; 21: 33-37
13. Risi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996; 14(4): 457-60
14. Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci U S A* 1997; 94(6): 2150-2155
15. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization. *Mol Biol Cell* 1998; 9: 3273-3297
16. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns.

*Proc Natl Acad Sci U S A* 1998; 95: 14863-14868
17. Butte AJ and Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput* 2000; 418-429
18. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determi nation of genetic network architecture. *Nature Genetics* 1999; 22: 281-285
19. Kohonen T. Self-organized formati on of topologically correct feature maps. *Biol Cybern* 1982 ;43: 59-69
20. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoi etic differentiation. *Proc Natl Acad Sci U S A* 1999; 96: 2907-2919
21. Kim JH, Ohno-Machado L., Kohane IS Unsupervised Learning from complex data: the Matrix Incision Tree Algorithm. *Pac Symp Biocomput* 2001; 30-41
22. Kim JH, Ohno-Machado L, Kohane IS. Visualization and Evaluation of Clus tering Structures for Gene Expression Data Analysis. *J Biomed Inform* 2002 (accepted and in press)
23. Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001; 28(1): 21-28
24. Park JC, Kim HS, Kim JJ. Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar. *Pac Symp Biocom put* 2001; 6: 396-407
25. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286: 531-537
26. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000; 403: 503-511
27. Altman RB. The Interactions Between Clinical Informatics and Bioinformatics: A Case Study. *J Am Med Inform Assoc* 2000; 7(5): 439-443
28. Tomita M. Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol* 2001; 19(6): 205-210
29. Chicurel M. Databasing the brain. *Nature* 2000; 406: 822-825
30. Brinkley JF. Structural informatics and its applications in medicine and biology. *Academic Medicine* 1991; 66: 589-591
31. Brown FK. Chemoinformatics: What is it and How does it Impact Drug Discovery. *Annual Reports in Medicinal Chemistry* 1998; 33: 375-384
32. Hann M, Green R. Chemoinformatics - A new name for an old problem. *Curr Opin Chem Biol* 1999; 379-383
33. Degoulet P, Fischi M. Introduction to clinical inforamatics. 1997, Springer, New York
34. Friede A, Blum HL, McDonald M. Public health informatics: how informa tion-age technology can strengthen public health. *Annu Rev Public Health* 1995; 16: 239-252