유전체 발현의 정보학적 분석과 응용

Genomic Applications of Biochip Informatics

김 주 한
서울대학교 의과대학 생명정보의학

Ju Han Kim, M.D., Ph.D., S.M.
Seoul National University Biomedical Informatics, Seoul National University College of Medicine,
28 Yongon-dong Chongno-gu, Seoul 110-799, Korea
juhan@snu.ac.kr,a

초 록

Bioinformatics is a rapidly emerging field of biomedical research. A flood of large-scale genomic expression data transforms the challenges in biomedical research into ones in bioinformatics. Clinical informatics has long developed technologies to improve biomedical research by integrating experimental and clinical information systems. Biomedical informatics, powered by high throughput techniques, genomic-scale databases and advanced clinical information system, is likely to transform our biomedical understanding forever much the same way that biochemistry did to biology a generation ago. The emergence of healthcare and biomedical informatics revolutionizing both bioinformatics and clinical informatics will eventually change the current practice of medicine, including diagnostics, therapeutics and prognostics.

Key word: Human Genome Project, bioinformatics, biomedical informatics, genomics, gene expression, DNA microarray

Ⅰ. 서 론

인간유전체사업(Human Genome Project)은 그간의 생명과학 기술발달 성과를 의학연구에 직접 적용하여 인류의 건강과 복지증진에 기여하고자 하는 목적으로 1980년대 말에 시작되었다. 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 보조수단의 형식으로 발달했다. 그러나 생명을 구성하는 유전체, 전사체, 단백체(genome, transcriptom and proteom) 등에 관한 방대한 정보가 체계적으로 축적되면서 생명정보학은 생명과학 연구의 중심 방법론으로 자리잡게 되었다.

생명정보학은 아직 그 걸음마기에 있다. 하지만 생명정보학이 PCR 기술처럼 생명공학 연구의 보조수단으로 남는 것이 아니라 의료정보학과 결합하며, 동시에 하나의 독립된 학제를 구성하는 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.

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

유전자 다형성(Genetic Polymorphism) 연구와, 유전자칩을 주축으로한 기능유전체학(Functional Genomics)으로 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 과녁이고, 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 시스템 생물학 연구의 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 넉아웃(knock out)하여 살아남은 200여 종의 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.

Ⅱ. 바이오칩 기술의 발달

유전자칩에 대한 상세한 리뷰는 관련논문들[3][4][5][6][7]에 잘 정리되어 있다. 유전자칩의 학문적, 상업적 성공에 힘입어 세포칩, 단백질칩 등 다양한 바이오칩 개발연구가 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 변증법적인 질적 변화,

1) Hughes TR, Marton MJ, Jones AR, et al. Functional discovery via a compendium of expression profiles. Cell 2000;102(1):109-26
2) Ideker T, Thorsson V, Ranish JA, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 2001;292(5518):929-34
3) Schena M, Shalon D, Davis RW et al. Quantitative monitoring of gene expression patterns with a cDNA microarray. Science 1995;270:467-470
4) 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
5) Pease AC, Solars D, Sullivan EJ et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis Proc Natl Acad Sci USA 1994;91:5022-5026
6) Lockhart DJ, Dong H, Byrne MC et al Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnol 1996;14(13):1675-1680
7) DeRisi JL, Iyer V, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680-686

즉 새로운 패러다임의 도래를 유발하는 것이다. 바이오칩의 도입은 유전자 연구를 유전체 전체로 확장하였고, 유전자 검출기법을(현존하는 몇 가지 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 어피메트릭스(Affymetrix) 사의 광식각(photolithography) 칩으로 나눌 수 있다. cDNA 칩은 이름 그대로 ORF's(Open Reading Frames, i.e. gene) 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 정도의 합성 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 수치 값으로 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 공개하였고 자사의 자료분석 소프트웨어 MAS(MicroArray Suite) 4.0을 5.0으로 개편 발표하였다.

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.

Ⅲ. 기능 유전체학

기능유전체학에서가장 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 비교하거

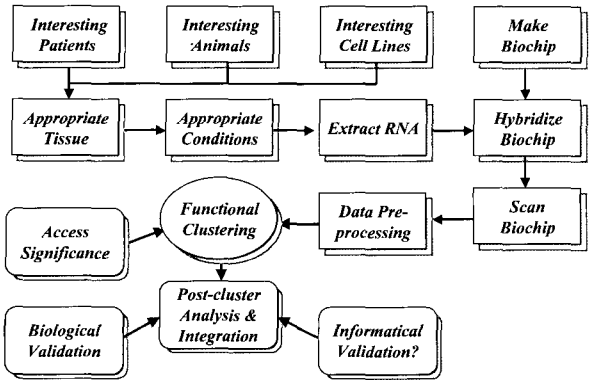

Figure 1. A strategy for functional genomics with biochip informatics

나, 다른 종에서 관련 유전자의 발현을 비교하거나, 시계열 발현 프로파일을 분석하는 등의 다양한 방법에 의해서 유전자 정보를 대량 처리하여 새로운 지식을 찾아내는 것이다(그림 1).

단순하게 세포에 특정한 처치를 한 후, 조건 전후의 유전자 발현비율(intervention fold difference)을 분석하여 실험 조건간의 발현량의 차이가 유의한 유전자들(differentially expressed genes)을 찾는 것만으로도 매우 흥미로운 결과를 얻을 수 있다.[8][9] 좀더 체계적인 방법은 그림 2에 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 밝혀가는 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 동반 조절되고 있는 유전자군(tightly co-regulated genes)을 찾는 것이다. 클러스터 분석이 동반조절 유전자군을 찾아주지만 유전자 기능을 다

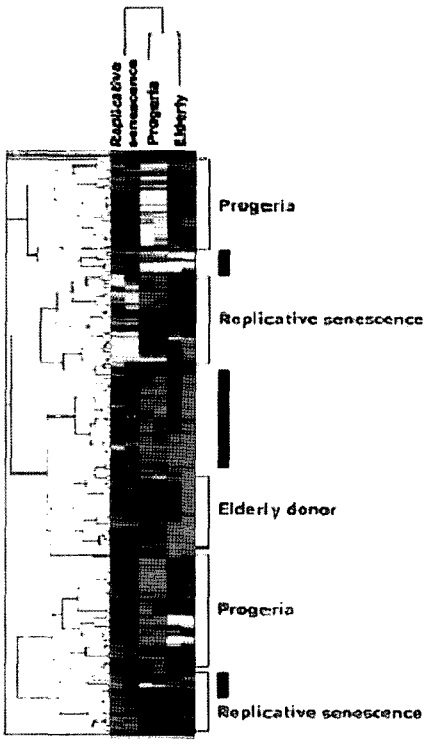

Figure 2. Clustering gene expression profiles by hierarchical tree clustering

밝혀주는 것은 아니다. 오히려 클러스터 분석은 미지의 유전자가 어떤 유전자들과 상관되어 있는가 하는 단초를 제시하여, 복잡한 관찰로 부터 새로운 가설을 제시함으로서(data-driven hypothesis generation), 그림 1에 표시된 것과 같이 다음 실험을 설계하는 시간과 노력을 줄여주는 것이다.

클러스터 분석은 크게 계층적 클러스터링과 분할형 클러스터링으로 구분된다. Spellman 등(1996)[10]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

8) DeRisi JL, Penland L, Brown PO et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14(4):457-600)

9) Heller RA, Shena A, Chai A et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94(6):2150-2155

10) Spellman PT, Sherlock G, Zhang MQ et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces Cerevisiae by microarray hybridization. Molecular Biology of the Cell 1998;9:3273-3297

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(그림 2).[11] 공개 소프트웨어는 http://rana.stanford.edu/ software/에서 구할 수 있다.

Butte와 Kohane(2000)[12]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. 그림 3은 relevance networks를 이용해 상관된 유전자 네트웍을 찾아낸 예이다. Butte 등은 피어슨 상관계수가 선형 상관관계만을 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. Church 등은 K-means 클러스터 분석을 적용했고[13], Tamayo 등(1999)은 인공 신경망 기법의 하나인 SOM(Self-Organizing Maps)[14]으로 의미 있는 유전자 클러스터들을 찾을 수 있음을 제시했다(그림 4).[15] SOM의 적용을 위한 공개 소프트웨어는 http://waldo.wi.mit.edu/MPR/software.html에서 다운로드 받을 수 있다. 알고리즘적으로는 순차적 K-means 알고리즘에 몇 가지 보완을 가한 것으로도 해석할 수 있다. 이러한 알고리즘들이 한번에 하나씩의 데이터씩만 고려한다는 한계점을 극복하기 위해

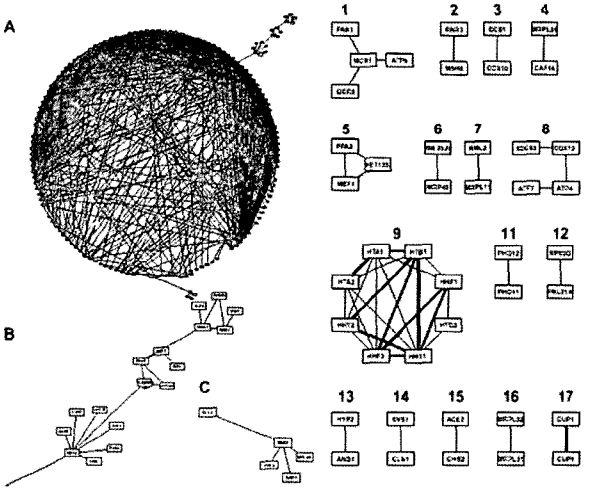

Figure 3. Clustering expression profiles by Relevance Networks
(from Butt & Kohane, 2000)

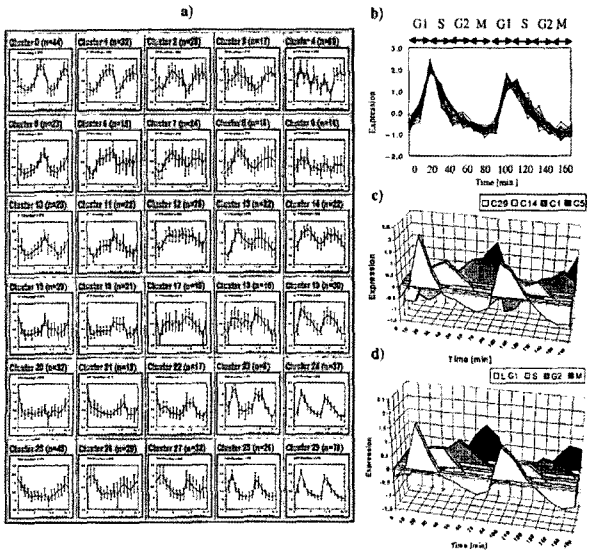

Figure 4. Partitional clustering of yeast cell division cycle data by Self-Organizing Maps(SOM).
(From Tamayo et al., 2000)

11) Eisen MB, Spellman PT, Brown PO et al Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863-14868
12) Butte AJ and Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000;418-429
13) Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nature Genetics 22: 281-285
14) Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics 1982;43:59-69
15) Tamayo P, Slonim D, Mesirov J et al. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96:2907-2919

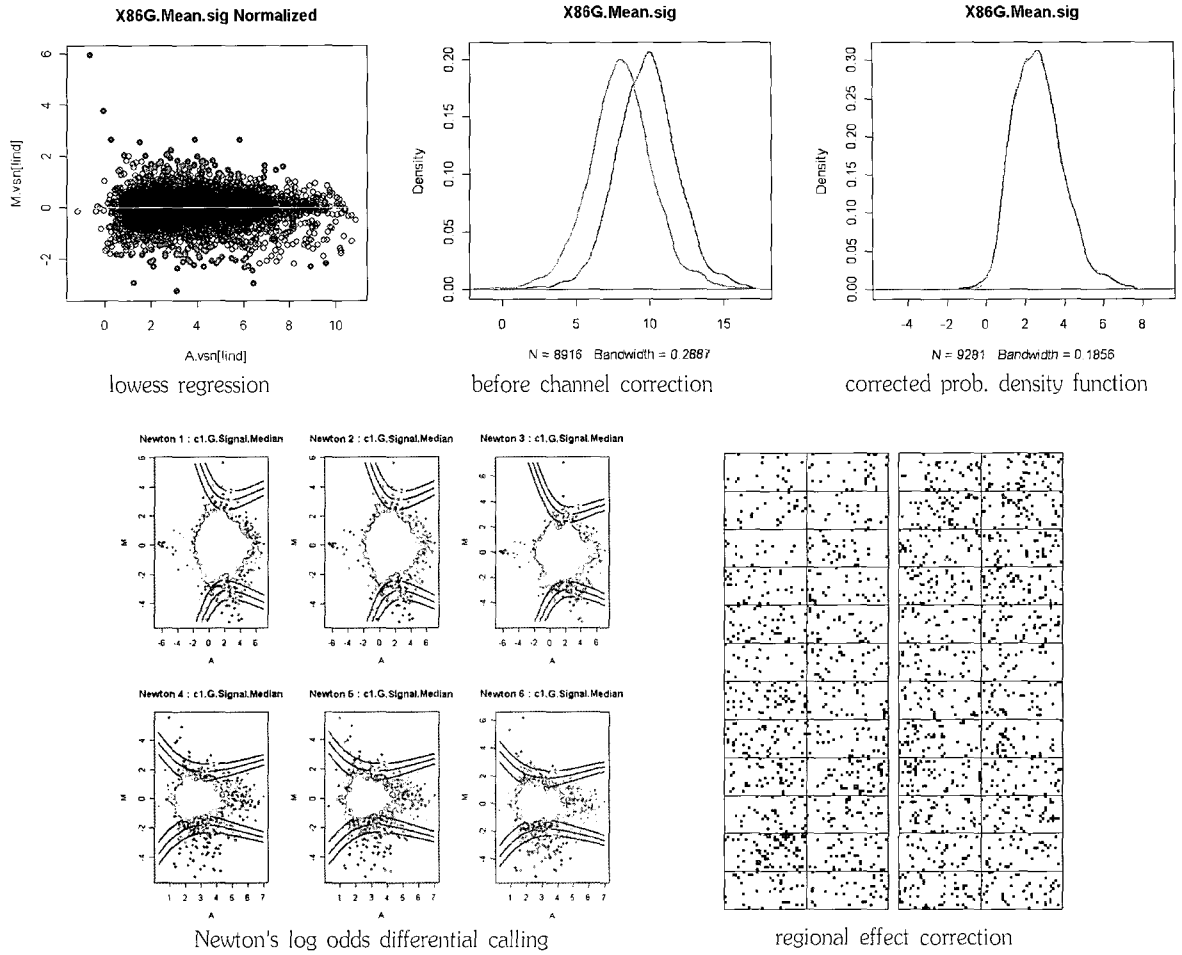

Figure 5. *Steps in DNA microarray analysis*

고안 된 Kim 등(2001)의 매트릭스 분할법도 소개되었다.[16][17] 그림 5은 유전자칩 분석의 제 단계를 간략히 도시한 것이다.

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 도입되고 있다. 예를 들어 퍼브진 (PubGene, http://www.pubgene.org/)은 텍스트 마이닝 기법을 활용하여 유전체 발현 정보를 메드라인과 같은 문헌정보와 통합해낸다.[18] 다양한 메타 데이터베이스가 개발되고 있으며 자연어처리 기법도 생명-의

16) Kim JH, Ohno-Machado L., Kohane IS Unsupervised Learning from complex data: the Matrix Incision Tree Algorithm. Pac Symp Biocomput 2001;30-41

17) Kim JH, Ohno-Machado L, Kohane IS. Visualization and Evaluation of Clustering Structures for Gene Expression Data Analysis. J Biomed Inform 2002(accepted and in press)

18) Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 2001 May;28(1):21-8.

료 분야의 문헌 및 사실 데이터베이스로부터 유전자 조절 상호작용 네트웍을 재구성하는데 유용하게 이용된다.[19] KEGG(http://www.kegg.org/)와 같은 대사 네트웍의 데이터베이스의 빠른 객체화와 조직화도 눈여겨 볼만하다. 구조유전체학적 서열정보도 유전자 기능연구에 유용한 정보를 제공한다.

Ⅳ. 생명의료정보학: 새로운 의학의 태동

인간유전체사업과 생명정보학의 태동기에서 최신 동향까지를 주마간산 격으로 살펴보았다. 소개된 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 등(1999)은 6817 개의 유전자 발현 패턴 분석만으로 백혈병의 아형인 AML와 ALL을 판별할 수 있음을 보고하였다.[20] Alizadeh 등(2000)은 Diffuse Large B-cell Lymphoma(DLBCL)의 유전자 발현을 클러스터 분석하여 두 개의 새로운 아형이 존재함을 발견하였다.[21] 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[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 기반으

19) Park JC, Kim HS, Kim JJ. Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar. Pac Symp Biocomput 2001;6:396-407.

20) Golub TR, Slonim DK, Tamayo P et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7.

21) Alizadeh AA, Eisen MB, Davis RE et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503-511

로 데이터마이닝 등의 생명정보학적 패러다임을 사용한다. 특기할 점으로 유전체 발현정보의 데이터 모델이 최근 OM-MAGE/ML(Object Model-Microarray Gene Expression Markup Language)로 통합된 것이다.[22] 이는 MGED[23] 그룹에 의해 주도되었으며 객체지향형 모델링과 XML기반의 자료교환 체계를 정의하고 있어서 그 구조와 기능, 목적에 있어 임상의료정보의 대표적 모델링링 HL-7의 CDA(Clinical Documet Architecture[24])와 그 맥락을 같이한다. 향후 임상의료정보체계와 생명정보체계가 대통합을 이루기 위한 초석을 쌓는 과정으로 이해되어 저자의 연구실에서 이에 대한 연구를 진행하고 있기도 하다.

생명-의료정보학(Biomedical Informatics, 혹은 정보의학)은 분자생물학과, 정보과학과, 임상의학의 제 지식이 삼위일체를 이루어가며 미래 생명과학을 주도할 핵심 학문이다. Alizadeh 등(2000)[28]이나 Golub 등(2000)[27]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.[25] 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, 정보의학(i.e., bio-molecular informatics, computational cell biology[26], computational physiology[27], digital anatomy[28], chemoinformatics[29][30], clinical informatics[31], and public health informatics[32])은 생명과 질병현상의 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.

---

22) Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ Jr, Brazma A. Design and implementation of microarray gene expression markup language(MAGE-ML). Genome Biol. 2002 Aug 23;3(9):RESEARCH0046.

23) MGED - Microarray Gene Expression Data Society [http://www.mged.org/]

24) Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, Kimber E, Lincoln T, Mattison JE. The HL7 Clinical Document Architecture. J Am Med Inform Assoc. 2001 Nov-Dec;8(6):552-69.

25) Altman RB. The Interactions Between Clinical Informatics and Bioinformatics: A Case Study. J Am Med Inform Assoc 2000;7(5):439-443.

26) Tomita M. Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol 2001 Jun;19(6):205-10

27) Chicurel M. Databasing the brain. Nature 2000;406:822-825.

28) Brinkley JF. Structural informatics and its applications in medicine and biology. Academic Medicine 1991;66:589-591

29) Brown FK. Chemoinformatics: What is it and How does it Impact Drug Discovery. Annual Reports in Medicinal Chemistry 1998;33:375-384

30) Hann M, Green R, Chemoinformatics - A new name for an old problem. Current Opinion in Chemical Biology, 1999;379-383

31) Degoulet P, Fischi M. Introduction to clinical inforamatics. 1997, Springer, New York.

32) Friede A, Blum HL, McDonald M. Public health informatics: how information-age technology can strengthen public health. Annu Rev Public Health. 1995;16:239-52.