Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

DNA Microarray 발현정보에 대한 생물학적 정보처리에 관한 연구

A Study of a Biological Information Processing for DNA Microarray Expression Data

조영임[1], 정형철[2*†]

[1] 수원대학교 IT대학 컴퓨터학과
E-mail: ycho@suwon.ac.kr
[2] 수원대학교 자연과학대학 통계정보학과
E-mail: jhc@suwon.ac.kr

요 약

본 논문은 바이오 인포메틱스의 분야를 간단히 소개하고, 기능유전체학에서 microarray 실험에 대한 통계적 방법론을 살펴보고자 한다. 또한 DNA chip 설계와 생물학적 특징에 대해 살펴보고 각 분야에서 적용되는 통계적 방법을 연구분석 해보고자 한다.

Key Words : 바이오 인포메틱스, DNA microarray, 통계적 정보처리방법

1. 서 론

바이오 인포메틱스(Bioinformatics)란 '생물정보학', '*computational biology*', '*computational molecular biology*'등의 용어로 불리우며, 컴퓨터를 이용해서 생물 정보의 해석을 목적으로 하는 학문의 분야를 말한다. 넓은 의미로는 컴퓨터를 이용하여 생물학을 연구하는 모든 분야를 뜻하며, 좁은 의미로는 DNA나 단백질(protein)의 서열정보를 해석하고자 하는 분야를 뜻한다[1,2].

본 논문에서는 바이오 인포메틱스의 분야를 간략히 소개하고, 기능유전체학에서 microarray 실험에 대한 통계적 방법론을 살펴보고자 한다. 최근 들어 국내외에서 microarray 기술을 사용하여 유전체 자료에 대한 연구가 집중적으로 이루어지고 있는데, 이는 기능유전체 연구에 있어서 상당히 중요하고 근본적인 원천기술 인식되고 있으며 유전자 발현 프로파일링에 널리 이용되고 있는 실험으로 앞으로 분자생물학 분야에서도 표준 연구방법의 하나로 사용되게 될 것이다.

DNA microarray 분석은 한번에 수천 개의 유전자를 동시에 혼성화 탐침(probe)으로 사용할 수 있게 해준다[3,4,5,6]. 즉, microarray 혹은 DNA chip이란 작은 유리판 혹은 membrane 위에 수천 개의 유전자를 얹어놓고 검사하고자 하는 샘플에서 추출해 낸 RNA의 발현정도를 한번의 실험으로 조사할 수 있는 방법이다. 1995년 미국 스탠포드 대학의 Pat Brown 등에 의해 개발된 이 DNA chip은 마이크로프로세서를 닮은 조그만 면에 여러 유전자의 cDNA를 붙여 놓은 것이다. 여기에 형광물질로 표시된 시료를 가한 후에 보합반응(hybridization)을 시킴으로써 DNA나 RNA의 특징적인 위치들 뿐 아니라 그 위치의 염기순서와 신호의 강도로부터 많은 보합 관련 정보를 얻을 수 있다. 이와 같은 DNA chip의 장점으로는 동일한 chip의 대량제작이 가능하며 수백 개의 유전자를 동시에 검사 가능하게 되어 기존의 Northern blot 방법보다 많은 시간과 비용을 절약할 수 있으며 극히 미량의 DNA로도 chip 제작이 가능한 점을 들 수 있다[7]. 또한 microarray 실험은 결과가 동일하고 표준화되어 나오기 때문에 컴퓨터로 분석 할 수 있다. DNA chip을 이용함에 따라 세포 및 조직의 생리학적 또는 병리학적 변화에 따라 유전자들의 패턴이 어떻게 변화하는지 종합적으로 파악할 수 있게 되었다[8,9]. 또한 생리학적 변화 뿐 아니라 외부적 처치 및 자극에 따른 반응을 동시에 관찰하면서 미지의 유전자들의 역할을 추정하고 궁극적으로 개개의 유전자들의 기능을 밝힐 수 있게 되었다. 특정 유전자의

* 교신저자

이상 또는 발현 변화는 단지 그 유전자의 변화만으로 끝나는 것은 아니며, 또 다른 여러 유전자의 발현 변화를 유도함으로써 최종적으로 특정 형질의 발현을 유도하게 된다. 즉 DNA chip은 특정 유전자의 이상이나 발현 변화로부터 나타나는 다차원적인 대사에 경로에 의한 유전체 활성의 총체적인 변화를 규명하는 방법으로 DNA chip이 21세기 생명공학의 시대를 주도할 것으로 예견하고 있다[7,10]. 본 논문에서는 DNA chip 설계와 생물학적 특징에 대해 살펴보고 각 분야에서 적용되는 통계적 방법을 살펴보는데 의의를 두고자 한다.

본 논문의 2장에서는 바이오 인포메틱스에 대해 소개하고, 3장에서는cDNA microarray 자료에 대한 통계적 방법을 소개하고, 4장에서 통계적 분석방법론을 소개하고, 마지막으로 5장에서 결론을 맺고자 한다.

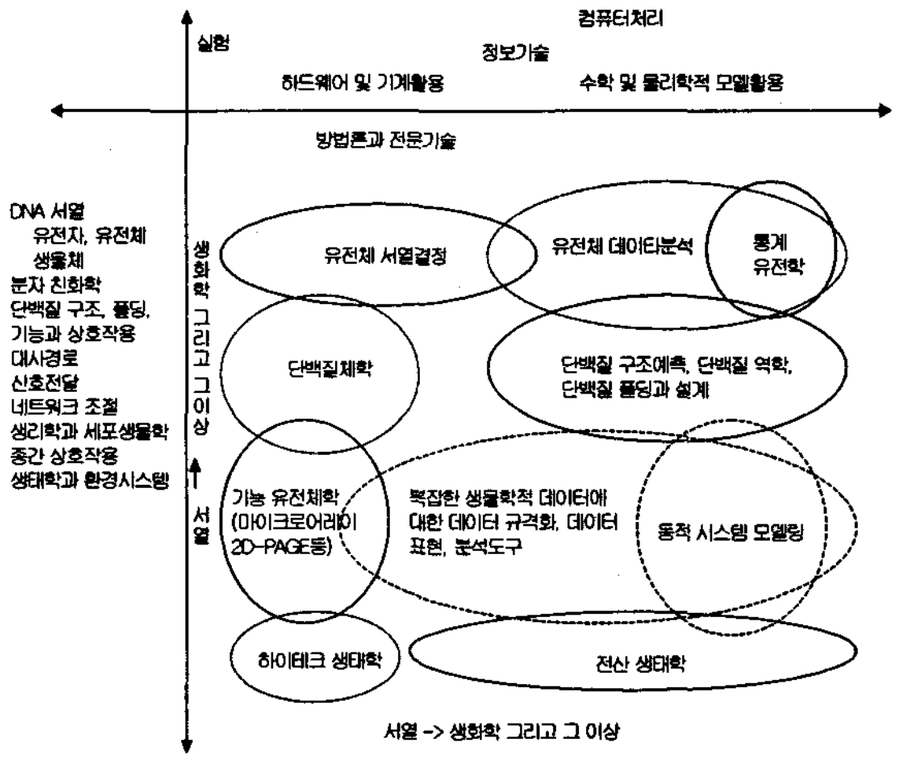


그림 1. 바이오 인포메틱스에 있어서 생물학과 다른 분야와의 관계

2. 바이오 인포메틱스

바이오 인포메틱스는 유전학(genomics)과 프로테오믹스(proteomics)로 구분할 수 있고 유전학은 구조 유전학(structural genomics)과 기능 유전학(functional genomics)로 나뉠 수 있다. 프로테오믹스는 단백질 데이터 베이스 구축과 단백질 기능 예측으로 구분할 수 있다[1,2].

바이오 인포메틱스를 연구하려면 그림 1에서와 같이 컴퓨터와 생물학은 물론 수학, 통계학, 언어학 등 다양한 분야의 보편성과 타당성을 동시에 갖고 있어야 한다. 즉, 생물학적 자료의 근본적인 성격을 이해하지 못한 채 패턴인식이나 정보학 분야에만 치중하다보면 잘못된 오류를 범할 수 있기 때문이다. 컴퓨터의 데이터베이스 기술이나 자료구조, 알고리즘, 통계적 분석 방법 등과 DNA 서열 분석 및 비교예측 방법을 잘 응용해야 하는 것이 매우 중요한 기술이다.

그러나 바이오 인포메틱스는 위에서 설명한 단순 서열분석 및 비교, 서열 데이터베이스 구축보다 훨씬 복잡한데, 단백질 서열 데이터 분석과 단백질의 구조 분석에서부터 대사과정 모델링, 집단과 생태 시스템에서의 양적 분석에 이르기까지 다양한 과학 분야가 이용되고 있다. 바이오 인포메틱스는 컴퓨터를 도구로써 이용하여 생물 시스템의 성격을 수학이나 물리학적 모델로 바꾸고 그 데이터를 분석하기 위해 새로운 알고리즘으로 데이터베이스를 개발하며 그 데이터베이스를 이용할 수 있는 접근 도구까지 만드는 총체적인 기술을 의미한다.

3. cDNA Microarray

유전자는 DNA로 구성되어 있고, 이 DNA는 뉴클레오티드 또는 염기(base)라고 불리는 각 단위로 만들어진 선형의 중합체(polymer)이다. 모든 생명체의 DNA 서열을 구성하는 네 개의 염기는 아데닌(Adenine), 구아닌(Guanine), 시토신(Cytosine), 티민(Thymine)이고 각각 A, G, C, T라는 글자로 표현한다. 선형 DNA의 염기배열 순서는 개체를 만드는데 필요한 지령을 담고 있는데, 이 지령은 복제, 전사, 번역이라 불리는 과정을 통해 읽히게 된다. 또한 DNA에 저장되어 있는 유전정보는 생물학의 중심도그마에 따라 두 단계의 과정을 거쳐서 발현(expression)한다[3,4].

첫 단계가 DNA의 염기서열이 mRNA(messenger)로 전사되는 것이며, 두 번째 단계는 mRNA가 단백질 합성공장인 리보솜에서 단백질을 생산하도록 번역되는 것이다. 특히 RNA는 DNA의 염기와 비교하여 티민이 우라실로 변형되는 염기를 취하고 있으며 DNA의 복사본인 mRNA, 리보솜 RNA인 rRNA(ribosome RNA), 아미노산을 전달해주는 tRNA(transfer RAN)의 3가지 종류가 있다. 또한 RNA로부터 만들어지는 단백질은 A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y로 표기되는 20종류의 아미노산으로 구성되어 있으며, 선형구조를 넘어서는 복잡한 3차원 구조에 의해 결정된다.

cDNA란 complementary DNA를 말하며 mRNA를 주형으로 하여 얻어진 DNA를 의미한다. 즉 mRNA를 역전사하여 복사한 DNA이다. 일반적으로 동물, 식물, 박테리아 등은 세

포 내에서 DNA, RNA, protein 순으로 물질이 만들어지는데 반해 바이러스의 경우 reverse transcriptase라는 효소가 존재하여 RNA에서 DNA로의 역으로 합성이 가능하다. 이러한 reverse transcriptase를 이용하여 실험에 사용하는 방법을 RT-PCR(Reverse Transcriptase Polymerase Chain Reaction)이라 한다.

Chen[11]에 의하면 cDNA Microarray 실험의 목적은 유전자 probe를 고정시키고 조직에서 얻은 mRNA를 가하여 발현양상을 조사하는 것이다(그림 2 참조).

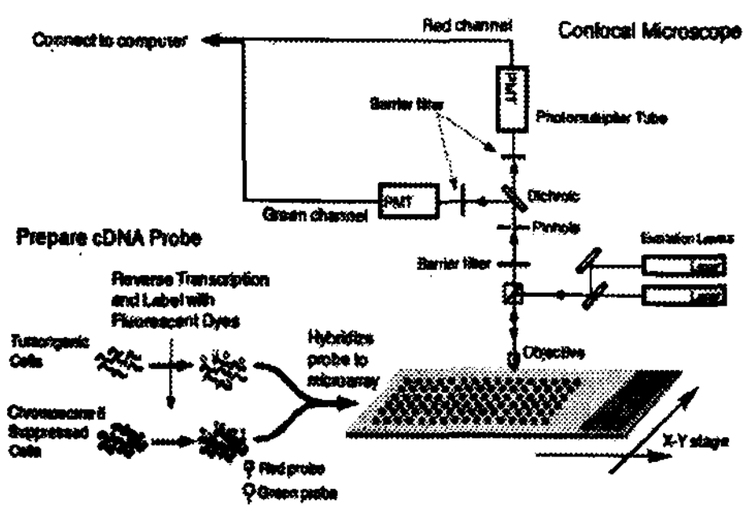


그림 2. Microarray 실험과정

cDNA microarray 실험은 그림 2에서 보는 바와 같이 실험군과 대조군 사이에서 유전자 발현의 변화를 상대적으로 비교하는데 있다. cDNA microarray 실험이 끝나면 스캐너를 통하여 2개의 이미지 파일을 얻고 이를 합하여 최종 결과를 얻는다[11]..

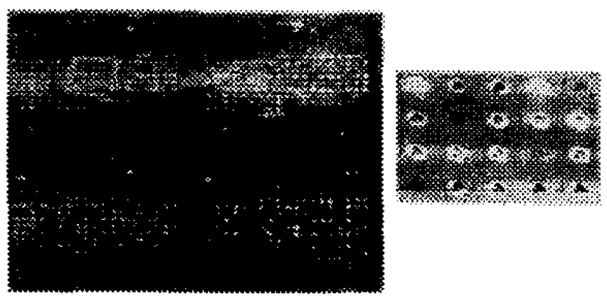


그림 3. Microarray 실험에 의한 이미지와 부분 확대 사진

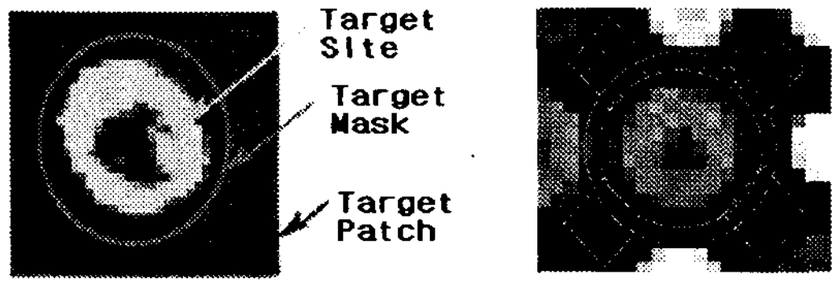
다음으로는 그림 3과 같은 Tiff 파일은 특별히 고안된 프로그램에 의하여 수치로 분석되는데 기본 원리는 각 spot을 형성하는 pixel들의 밝기 강도를 측정하고 이들의 평균치를 구하여 spot의 강도로 나타낸다.

4. cDNA Microarray 실험과 통계방법

Microarray 실험은 생물학적 질문, 실험설계, 이미지 분석, 표준화(normalization), 유의한 유전자를 탐색하기 위한 각종 통계적 분석, 생물학적 확인 및 해석의 단계를 거치는데, 실험설계 단계부터 고려되는 통계적 방법들을 소개하면 다음과 같다.

Microarray 실험을 하기 전에 슬라이드의 개수와 각 슬라이드에 어떠한 mRNA 샘플 들이 사용되어야 할 것인가를 결정해야 한다.

이미지 처리의 목적은 각 표본 내의 특정 DNA 양을 측정한다는 의미이다. 그림 4A에서는 하나의 spot에 대한 영상이 제시되었다. 그림 4B는 segmentation 방법을 통해 spot mask를 결정하는 과정을 보여준다. 스캐닝 된 슬라이드에는 수많은 spot에서 후경(background) 강도, 전경(foreground) 강도, 샘플 1과 2에 의한 target의 보합반응에 의한 빛, 유리표면에서 발생하는 빛 등 여러 종류의 형광성 빛이 혼합되어 있다. 그런데, 이와 같은 슬라이드에서 결국 이미지 처리는 각 spot에서 red, green channel의 전경 및 후경의 밀도(intensity)를 얻어 target의 평균 형광 강도를 추출하는 것이다.



(그림 A) (그림 B)

그림 4. 이미지처리 그림

표준화는 microarray 실험 후 통계분석을 행함에 앞서 계통적 변동을 제거하는 작업이다. Microarray 분석의 핵심적인 목표 중 하나는 어떤 유전자들이 다르게 발현된다는 강한 증거를 보이는가 하는 것이다. 유전자 발현의 추정 문제는 microarray 분석의 기본이 된다.

Microarray 분석의 핵심적인 목표 중 하나는 어떤 유전자들이 다르게 발현된다는 강한 증거를 보이는가 하는 것이다. 즉 여기에는 두 부분이 있다. 첫째는 가장 강한 증거부터 약한 증거까지 발현에 차이가 나는 유전자들의 순위를 매길 수 있는 통계량을 찾는 것이다. 둘째는 기각치를 찾는 것이다. 이를 통계적으로 추정과 검정의 문제로 볼 수 있다[12].

유전적 특징은 한 두 유전자에 의해서 나타나는 것이 아니고 여러 유전자간의 관계 하에서 이해할 수 있으므로 특정 유전자의 특징을 살피기 위해서는 유사하게 발현되는 유전자를 하나의 집단으로 모아 볼 필요가 있다. 이러한 접근 방법이 곧 군집 및 분류분석이다[12]. 이 중 분류분석은 계급(class)이 이미 존재할 때 효율적인 분석방법이다. 군집분석은 보편적으로 많이 사용되고 있는 분석방법으로, hierarchical clustering, k-means clustering,

MDS, PCA, self-organizing map, fussy method 등의 다양한 분석기법들이 이미 통계학에서 사용되어 왔으며 언급된 많은 군집분석 방법이 유전자 분석에 활용되고 있는 실정이다.

이외에도 Kerr 의 분산분석 기법[13], Holster의 singular values decomposition (SVD)에 의한 유전자 발현 분석[14] 등 상당히 다양한 방법들이 시도되고 있다.

5. 결론

바이오 인포메틱스분야에서 유전체연구는 기능유전체(functional genomics)의 시대로 접어들고 있다. 유전체 연구에 있어서 선도적 역할을 하고 있는 미국의 경우에도 생물과 통계적 지식을 모두 갖춘 바이오 인포메틱스 연구자의 수가 많지는 않다. 그러나 대학 및 여러 연구기간에 생물통계학자들이 다수 포진해 있고 각 기관에서 풍부한 지원과 물적, 인적 교류를 바탕으로 유전체 연구분야에 관한 통계적 방법의 연구가 활발히 진행되고 있다. 유전체 자료에 대한 통계적 분석 방법이 계속해서 쏟아져 나오고 있지만 이들 방법이 완전히 검증되어 있다고는 볼 수 없다. 따라서 현재 사용되고 있는 유전체 자료의 분석도구들에 대한 보다 정밀한 연구가 필요하다. 또한 기존의 존재하는 방법들에 대한 비교분석을 통해 microarray 자료분석의 각 단계에서 발생되는 통계적 쟁점에 대한 이해가 요구된다.

최근 우리나라에서도 많은 물적, 인적 투자를 통해 유전체 연구사업에 박차를 가하고 있다. 그런데 유전체 연구는 생물학자만의 것이 아닌 학제간 연구가 활성화 되었을 때 21세기 post genome 시대에 우리나라도 우수한 원천기술을 보유하게 될 것이다. 이를 위해서는 각 분야의 지식이 총체적으로 사용되는 체계가 이루어져야 하며 통계적방법과 같은 측면지원에 많은 통계학자의 관심이 필요하다.

참고문헌

[1] 이정근, 오석준, 김종민 역, *바이오인포메틱스*, 한빛미디어, 2001

[2] 조영임, 인공지능시스템, 홍릉과학출판사, 2003

[3] Dudoit, S., Fridlyand, J., and Speed, T., Comparison of methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, vol.97, pp.77-87, 2002

[4] Dudoit, S., Tanh Y.H., Speed T.P., and Callow M.J., Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica,* vol.12, pp.111-140, 2002

[5] Nguyen, D.V., Arpat, A.B., Wang, N., and Carroll, R.J., DNA microarray experiments: Biological and technological aspects, *Biometrics*, vol.58, pp.701-717, 2002

[6] Nguyen D.V., Wang N., and Carroll, R.J., Evaluation of missing value estimation for microarray data. *Journal of Data Science*, 2004

[7] 이재원, DNA chip 자료분석에서의 통계적 방법의 응용, *Technical Report, Department of Statistics, Korea University,* 2002

[8] Hilsenbeck S.G., Friedrichs W.E., Schiff R., O'Connel P., Hansen R.K., Osborne C.K., Fuqua S.A.W., Statistical analysis of array expression data as applied to the problem of tamoxifen resistance, *J. Nat. Cancer Institute,* vol.91, pp.453-459. 199.

[9] Perou C.M., Sorlie T., Eisen M.B., Van de Rijn M., Jeffrey S.S., et al., Molecular portraits of human breast tumors, *Nature,* vol.406, no.6797, pp.747-752, 2000

[10] Marton M.J., DeRisi J.L., Bennett H.A., Iyer V.R., et al., Drug target validation and identification of secondary drug target effects using DNA microarrays, *Nat. Med,* vol.4, no.11, pp.123-1301, 1998

[11] Chen Y., Dougherty E. R., and Bittner L. Ratio based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics,* vol.2, pp.364-374, 1997

[12] Smyth G.K., Yang Y.H., and Speed T., Statistical issues cDNA microarray data analysis. *Research Report, Walter and Eliza Hall Institute of Medical Research,* Australia, 2002

[13] Kerr M.K., and Churchill G.A., Experimental design for gene expression microarrays. *Biostatistics,* vol.2, pp.183-201, 2001

[14] Holster, N.S., Maritan, A., Cieplak, M., Fedoroff, N.V., Banavar, J.R., and Fedoroff, N.V., Fundamental patterns underlying gene expression profiles: Simplicity from complexity, *Proceedings of the National Academy of Sciences*, vol.98, pp.1693-1698, 2000