지식축적기반 마이크로어레이 분석 통합개발환경 프로그램 설계

서민석*, 최지혜*, 오세종*
*단국대학교 나노바이오의과학과
e-mail:only_love_y@nate.com

IDE Design for Microarray Analysis Based on Accumulative Knowledge

Min-Seok Seok*, Ji-Hye Choi*, Se-Jong Oh*
*Dept of NanoBioMedical Science, Dankook University

요 약

최근, 마이크로어레이 실험 데이터의 품질과 재 생산성에 대한 신뢰도가 증가했기 때문에 마이크로어레이 데이터의 공유 및 활용에 대한 요구가 꾸준히 증가하고 있다.

하지만, 개별적으로 진행되는 이 실험에서, 연구자는 각각의 실험계획에 따른 실험을 위해 별도로 실험계획을 하고, 그에 따른 단편적인 결과를 얻을 뿐, 이를 다시 재활용 하는 방안에는 microarray databases를 이용하는 것만이 전부였다. 하지만, 이 방법은 일반 생물학자들이, 다시 데이터베이스를 이용해서 분석하는데 많은 어려움을 가져왔고, 또한 각각의 실험 과정을 이용하는 과정에서도, 통합개발환경을 구축하지 못 한 것에 대해 시간적 손해를 많이 입고 있다. 이에 본 논문에서는 실험계획부터 자료의 표준화 및 시각화, 유의한 유전자 탐색, 군집분석, 분류분석을 할 수 있는 통합개발환경 프로그램에 대해 제시하고, 결론적으로 이 데이터를 효과적으로 재활용 할 수 있는 방안에 대해서 제시하였다. 결론적으로, 이 프로그램은 개별적인 통계 프로그램으로 분석을 할 때에 비해, 편의성이 향상하며, 시간적인 소모를 줄임으로써, 상당히 많은 이득을 얻을 수 있으며, 한번 분석한 데이터를 효율적으로 저장해 놓음으로써, 추후에 제 2,3의 데이터 가공을 통해, 더 많은 정보를 얻어 낼 수 있다.

1. 서론

생명정보를 한 번의 실험으로 수많은 정보를 얻을 수 있는 실험 방법 중 하나인 마이크로 어레이 실험으로 생성이 되는, 다양하고 복잡한 구조의 마이크로레이 데이터에 효과적으로 접근하고 처리하는 방법에 대해서는 이미 많은 컴퓨터 학자들이 연구하고 이를 통합하려는 노력을 하였다. 실험방식과 분석하는 방식이 다양하기 때문에 데이터를 저장하는 형식이 다르기 때문에 . 따라서 상호간에 원활한 데이터 교환을 위한 기반으로 XML형식의 마이크로어레이 유전자 표현 데이터 (MGED: Microarray Gene Expression Data) 표준[1] 을 따른다. 마이크로 어레이 데이터에 대한 표준화 작업은 MIAME(Minimun Information About a Microarray Experiment)[2]는 check-list를 기초로 한 마이크로어레이 데이터 표현을 제안하였다. 최근에는 MIAME 기반의 MAGE-ML(Microarray GeneExpression Markup Langua-ge)[3], SOFT(Simple Omnibus Format in Text)[4], MINiML(MIAME Notation in Markup Language)[5],MAGE-TAB[6] 등의 데이터 교환 포맷을 이용하려고 노력중이다. 하지만 앞선 연구에서는 해당 실험에 대한 데이터들의 데이터 호환을 위해서 연구했을 뿐, 해당 관련된 데이터에 대해 직접적으로 관련짓는 것에 대해서는 연구된 바가 없다. 실제로, GO(Gene Ontology)[7]나 MEGE-ontology[8]과 같은 ontology는 생성되어 지고 있지만, 막상 이들을 손쉽게 이용 할 수 있는 프로그램에 대해서는 존재하고 있지 않다. 때문에 본 논문에서는 마이크로어레이에 대한 통합개발환경 분석 프로그램에 대해 제안하였다. 해당 프로그램은 실험계획, 품질관리, 데이터의 표준화, 유의한 유전자 탐색, 군집분석, 분류분석 등을 할 수 있는 프로그램과 더불어, 이들을 시각화해서 빠른 비교를 할 수 있는 툴을 제공하고, 마지막에는 해당 데이터들을 효과적으로 저장해

이 논문은 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(세계수준의 연구중심대학 육성사업, R31-2008-000-10069-0)

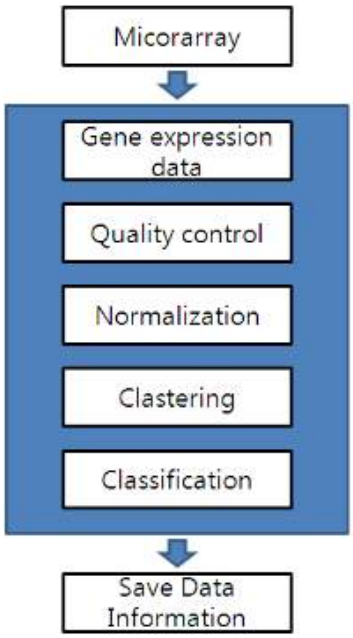
**교신저자

놓고, 추후에, 다른 사람이나 본인이 비슷한 실험을 했을 때 이를 참조함 으로써, 제3의 데이터 분석이 가능한 환경을 조성하는 것이 목표이다.

2. 본론

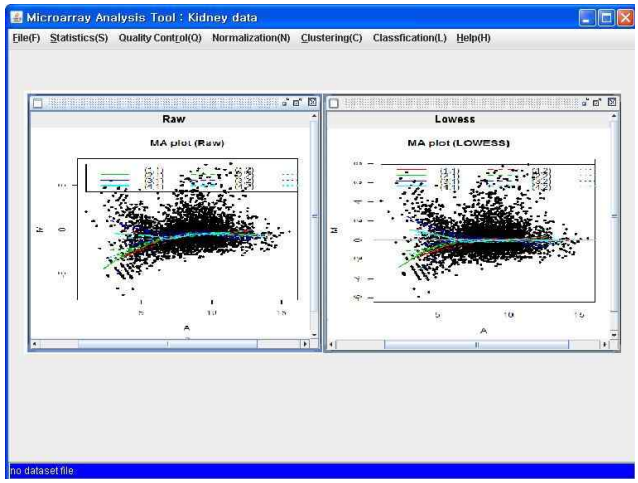
2.1 마이크로 어레이 데이터 분석 실험

마이크로 어레이 데이터는 해당 생물학자의 생물학적 의문에 의해, 실험이 계획 되어 진다. 때문에 각기 다른 생물학적 의문은 매우 다양한 마이크로어레이 데이터의 형태를 띄게 되고, 이는 분석법도 매우 다양하게 나타날 수 밖에 없다는 것을 의미한다.



[그림1] 마이크로 어레이 분석과정

[그림 1]은 마이크로어레이의 전체적인 분석과정을 나타낸다. 실험계획에 의해 얻어진 마이크로 어레이 데이터는 스캐너에 의해 gene expression 데이터로 얻어지고, 이는 품질관리 (불량Spot 찾기, Spot의 품질관리, 슬라이드 품질관리) [9]와 표준화(슬라이드 간 표준화, 슬라이드 내 표준화) [10]를 통해 수치화 된다. 또한 이를 통해 변하는 과정은 모두 저장되어지며, 이는 반드시 그래프를 통해 시각화 되



[그림 2] 마이크로어레이 분석 통합개발환경

어져야만 한다. 후에 발현값을 통해 군집분석 (hierarchical clustering, K-means clustering, K-medoid, SOM, PCA, MDS) [11]이나 분류분석 (k-NN, SVM, 인공신경망) [12]등을 행함에 있어서도, 실험자는 어떤 자료에 어떤 방법이 가장 최적이라는 것을 실험 전에는 알 수 없다. 때문에 이러한 알고리즘을 선택했을 때 결과는 일일이 비교할 수밖에 없는데, 제안하는 방법에서는 [그림 2]와 같이 이를 통합한 프로그램 내에서 결과를 서로 비교하는 방법을 통해 저장하고, 추후에 저장 할 때에도 해당 알고리즘의 효율성 등을 같이 저장함으로써, 나중 실험에 인용 가능하게 만들 수 있다.

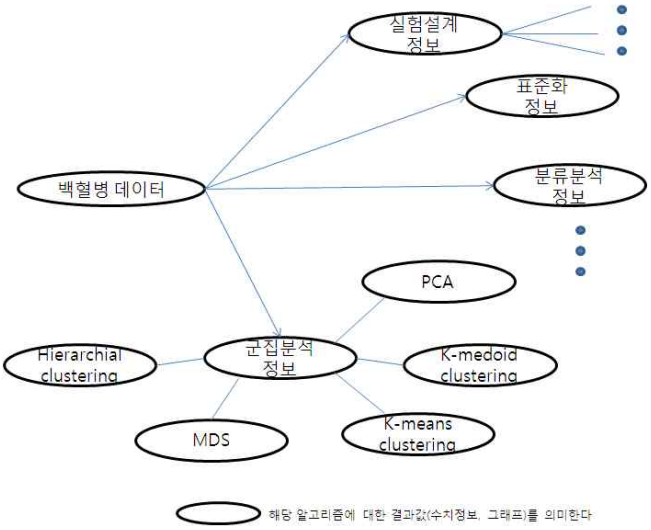
2.2 마이크로 어레이 분석 통합개발환경 툴

마이크로 어레이 실험은 생물학자의 생물학적 의문을 해결하기 위해 진행한다. 때문에 분석 할 때 발현 값 데이터는, 이를 바탕으로 해결 되어야만 한다. 하지만 이를 분석하는 것은 주로 통계학자나, 컴퓨터과학자가 해결하게 된다. 본 논문에서 제안하는 통합개발 환경은, 생물학자, 통계학자, 컴퓨터과학자 모두가 공통적으로 사용 할 수 있는 환경을 만드는 것에 그 의의가 있다. 여러 분석과정을 하나의 툴로 만듦으로써, 분석 속도를 높임과 동시에, 각 실험 단계를 지날 때마다, 데이터를 남겨, 실험의 정확도를 높이자 함에 있다. 각 단계에서 의미 있는 주석을 남기고, 해당 주석은 마지막 정보저장 단계에서 의미 있는 정보로 남아야 한다. 이는 통계학자나, 컴퓨터과학자가 소홀한 생물학적인 부분을 생물학자가 보충하는데 의미를 두며, 이와 반대로, 생물학자는 툴을 이용하는데 있어서 알고리즘이나 그 분석 방법에 대한 지식이 상대적으로 미비 하다. 때문에 이

툴은 해당 알고리즘의 결과를 시각화 하고, 직접적인 비교를 할 수 있기 때문에, 생물학자에 부족한 부분을 보충해 줄 수 있기를 기대해 볼 수 있다.

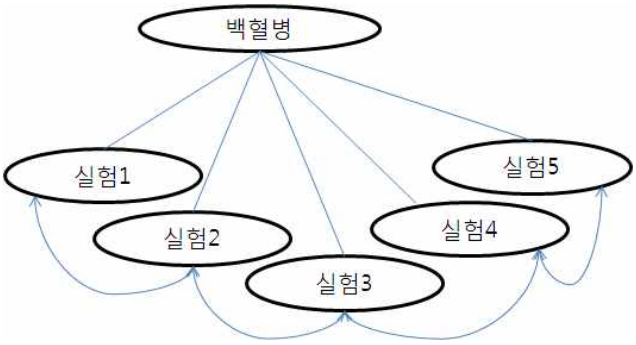
2.3 최종 분석데이터 저장 및 공유

분석된 데이터는 해당 분석 파트에 따라 각기 다른 결과 값을 가지고 있다. 표준화를 통한 표준화 수치를 비롯해, 해당 슬라이드의 발현 값 데이터, 어떠한 군집 분석 알고리즘을 사용했으며, 해당 알고리즘의 결과 그래프, 어떠한 분류분석 알고리즘을 사용했으며 해당 정확도 이 모든 정보를 데이터베이스에 정리함은 물론 이거니와, 이는 MGED 표준을 따르며, 그곳에 추가적인 부분이 되어야만 할 것이다. 본 논문에서는, Ontology를 이용한 데이터 저장에 대해서 서술해본다. Ontology는 지식과 지식을 관계 짓는 분야로, 검색엔진 분야에서 많이 응용되어지고 있다. 본 논문에서는 각 종 수치에 대한 정보를 트리형식으로 저장해놓고, 이에 대한 지식축적을 바탕으로, 비슷한 실험이 있을시, 이를 연관 짓는 Ontology 방식을 제안한다.



[그림 3] 결과정보 저장 자료구조

[그림 3]은 한 실험에 대한 결과정보를 저장한 자료구조를 표현한 그림이다. 위와 같이 모든 그래프 정보와 수치 데이터는 트리구조로 저장된다. 이는 프로그램에서 읽어 질 수 있는 파일 구조 형태이며, 이런 트리구조를 통해, 추후에 백혈병 데이터에 의한 다른 실험법등을 통해 분석 할 때에, 이를 즉시 참조 할 수 있으며, 발전된 방법이나 알고리즘을 즉시 비교 할 수 있는 reference 가 될 수 있다. 마이크로어레이 실험은 수많은 정보를 포함하고



[그림 4] 크로스 레퍼런스 관계

있기 때문에, 실제로 원하는 실험결과를 제외하고, 의미 있는 데이터를 살리기란 실제로 너무나 힘이 든다. 하지만, 원하는 결과에 쓰인 데이터가 아니더라도 이는 분명 가치 있는 데이터기 때문에, 이를 서로 참조하는 것은 필수적인 요소라 할 수 있겠다. 때문에 본 논문 에서는 [그림4] 와 같이 비슷한 실험 군에 대해서는 서로 크로스 레퍼런스를 통해 해당 실험의 정확도를 높일 뿐만 아니라, 단일 실험의 목표달성을 제외하고 또 다른 제3의 정보의 획득을 목표로 하였다.

3. 결론

본 논문에서는, 마이크로 어레이 분석 통합개발환경에 대해서 제안해 보았다. 실제로 분석을 함에 있어서 R등의 통계프로그램을 사용한 후, 패키지를 이용하고 있지만, 이는 실제로 명령어를 치는 과정에서 프로그래밍 적인 지식을 가지고 있어야 하며, 이는 실제로 생물학자 에게 부담으로 다가오는 것이 사실이다. 해당 프로그램 에서는 모든 메뉴를 명령어 방식이 아닌, GUI방식으로 처리했기 때문에 간편하며 직관적이다. 또한 데이터를 저장함에 있어서 이용자가 별도로 데이터를 백업할 필요 없이, 프로그램 자체에서 해당 데이터셋에 대한 모든 데이터를 효과적으로 저장해 놓고 사용할 수 있다. 이런 통합개발환경을 구축하고 효과적으로 데이터를 축적함에 의해, 분석속도를 높이고, 비슷한 실험의 연관관계를 추리해 별도의 분석 결과를 얻을 수 있다고 기대 해 본다.

4. 향후계획

본 논문에서 제안한 프로그램은, 현재 나온 기술들을 기반으로 만들었다. 즉, 알고리즘의 개선이나,

추가적인 정보를 수용해 데이터를 재분석하는 능력에서 뒤떨어 질 수 밖에 없다. 해당 프로그램은 인터넷 마이크로 어레이 데이터 베이스 로부터 검색을 토대로 본 프로그램 형식에 맞는 데이터 파일 생성을 하는 부분도 추가적으로 개발 되어야 할 것이다.

참고문헌

[1] C. A. Ball and A. Brazma, "MGED Standards: Work in Progress," OMICS, Vol.10, No.2, pp.138-144, 2006.

[2] Brazma, "Minimum information about a micro array experiment (MIAME)-toward standards for microarray data," Nature Genet., Vol.29, No.4, pp.365-371, 2001.

[3] Spellman, "Design and implementation of microarray gene expression markup language (MAGE-ML)," Genome Biology, Vol.3, No.9, RESEARCH0046.1-0046.9, 2002.

[4] Y. Yi, C. Li, C. Miller, and A. L. George Jr., "Strategy for encoding and comparison of gene expression signatures," Genome Biology, Vol.8, 2007.

[5] D. Field and S. Sansone, "A Special Issue on Data Standards," OMICS: A Journal of Integrative Biology, Vol.10, No.2, pp.84-93, 2006.

[6] Rayner, "A simple spreadsheet- based, MIAME -supportive format for microarray data: MAGE -TAB," BMC Bioinformatics, Vol.7, pp.489-507, 2006.

[7] The Gene Ontology (GO) database and infor matics resource. Harris MA, Nucleic Acids Res. 2004 Jan 1;32(Database issue):D258-61.

[8] Whetzel, H Parkinson, HC Causton, L Fan, J The MGED Ontology: a resource for semantics based description of microarray experiments oxf ordjournals.org 2006 - Oxford Univ Press

[9] Chen, Y Dougherty, E.R Ratio-based decisions and the Quantitative analysis of cDNA microarray images. Biomedical Optics ,2,313 -324

[10] Edward,D. Non-linear normalization and bac kground correction in one-channel cDNA mic roarray studies. Bioinformatics, 19(7),825-833.

[11] Eisen, M.B. Cluster Analysis and Display of Genome-Wide Expression Patterns. PNAS,95,1486 3-14868,

[12] Golub, T.R. Molecular classificatioin of cancer : class discovery and class prediction by gene expression monitoring. Science,286,531-537