FST를 이용한 마이크로어레이 유전자발현에 관한 해석

An analysis of microarray gene expression using FST

최경옥[1], 정환묵[2]

[1] 경북 경산시 하양읍 대구가톨릭대학교 컴퓨터정보통신공학부
E-mail: okajaa@cu.ac.kr
[2] 경북 경산시 하양읍 대구가톨릭대학교 컴퓨터정보통신공학부
E-mail: hmchung@cu.ac.kr

요 약

현재 생명공학은 급속도로 발전하고 있으며, 이를 통해 만들어지는 생물정보의 양은 기하급수적으로 늘어나고 있다. 이러한 것을 가능하게 하는 기술 중의 하나인 마이크로어레이 기법은 현재 질병의 진단 및 신약 개발 등을 위해 사용되고 있다. 마이크로어레이 유전자 발현에 관한 분석은 크게 유의한 유전자 추출, 클러스터링, 분류 및 유전자 네트워크 구축 등으로 볼 수 있다. 유의한 유전자 식별을 위한 통계학적 방법으로 T-test 및 Wilcoxon Rank Sum test 등이 있다. 최근에는 수정인자를 추가하거나 혹은 퍼지이론 등의 지능정보 이론을 추가하여 그 계산결과를 좀 더 상세화하고 세분화하는 연구들이 계속되고 있다.

본 논문에서는 두 개의 그룹에서 발현된 유전자들 중 유의하게 발현되는 유전자를 식별하기 위한 방법으로 퍼지이론을 도입하여 유의한 유전자를 규명하는 FST(Fuzzy Significance Test) 방법을 제안한다.

Key Words : 퍼지소속함수, 유의수준검정

1. 서 론

생물체내에서의 유전자의 역할을 규명하는 연구는 다양한 방법으로 행해지고 있다. 전통적인 방법은 "one gene in one experiment"에 기초하여 이루어졌으나, 현재에는 마이크로어레이 유전자 칩 등의 기술적 발전으로 인해 유전자 전체의 발현형태나 유전자 간의 상호작용 혹은 유전자 네트워크를 종합적으로 규명하는 연구가 활발히 진행되고 있다[10].

기존의 연구 방법과 달리 마이크로어레이 유전자 발현에 관한 연구방법은 수백에서 수천 혹은 수만에 이르는 유전자에 대한 발현정보를 한꺼번에 생산해낸다. 이러한 방대한 양의 유전자 발현 정보들 사이에서 서로 다른 그룹 혹은 다른 조건에서 유의하게 발현하는 유전자를 식별하기 위해 여러 가지 통계학적인 방법이 사용되고 있다[1,5].

본 논문에서는 기존의 통계적인 방법에 퍼지이론을 도입한 Fuzzy Significance Test(FST)를 제안도록 한다.

2. 관련연구

2.1 DNA 마이크로어레이 연구동향

1995년 Stanford 대학 Pat Brown 연구실에서 처음 개발된 DNA 마이크로어레이 기술은 세포내에서의 요구 혹은 외부적인 자극에 따라 다양하게 변화하는 DNA 내의 유전자의 전사(transcription)를 관찰, 연구하는 것이다[3]. 또한 마이크로어레이 유전자 발현에 관한 연구는 서로 다른 조건에서 유의하게 발현되는 유전자를 밝혀내고 이를 기초로 하여 새로운 유전자 발굴 및 유전자들의 상호작용을 규명하는 연구이다[4].

유의한 유전자 발굴을 위해 사용되는 유의수준 테스트 방법으로는 T-test, ANOVA, SAM, Wilcoxon rank sum test 등이 있으며, 퍼지이론에 바탕을 둔 유의수준 테스트 관한 연구도 활발히 진행 중에 있다[1,5].

2.2 퍼지이론 및 유의수준 검정

2개의 그룹에서 얻어진 유전자들의 발현양이 통계적으로 유의한가를 검정하기 위한 방법으로는 모수적 검정 방법인 T-test와 비모수적 접근방법인 Wilcoxon rank sum test 등이 있다[6].

독립된 두 그룹(S_1,S_2) 간의 평균의 차이가 통계적으로 유의한가를 평가하는 방법인 T-test에서 t값은 다음과 같이 계산될 수 있다[1].

$$t(S_1,S_2)=\frac{|\mu_{S_1}-\mu_{S_2}|}{\sqrt{\frac{\sigma_{S_1}^2}{|S_1|}+\frac{\sigma_{S_2}^2}{|S_2|}}} \quad (1)$$

여기서, μ_S는 평균, σ_S는 표준편차를 나타낸다.

그리고 퍼지이론을 이용한 유의수준 검정 방법도 다양하게 제안되고 있다. 퍼지 소속함수의 소속정도를 기준으로 유의성을 검정하는 방법인 fuzzy-set-theory-based method test(FM-test) 등이 있다[1]. 이 방법은 퍼지 소속함수를 이용하여 각각의 요소들이 자신이 속한 그룹이 아닌 다른 그룹에 속할 정도를 계산하고 (FM c-value : convergence degree), 이를 이용하여 divergence value인 FM d-value를 계산하였으며, 이에 대한 empirical p-value를 계산하고 있다[1].

$$c(S_1,S_2)=\frac{\sum_{i=1}^{n_1+n_2}(I_{S_1}(x_i)f_{F(S_2)}(x_i)+I_{S_2}(x_i)f_{F(S_1)}(x_i))}{n_1+n_2} \quad (2)$$

$$d(S_1,S_2)=1-c(S_1,S_2) \quad (3)$$

여기서$f_{F(S_1)}$과 $f_{F(S_2)}$는 각 요소의 각각의 그룹에 대한 소속함수를 나타내며,

$S=S_1\cup S_2=\{x_i,i=1,...,n_1+n_2\}$,

$n_1=|S_1|$, $n_2=|S_2|$, 그리고 $I_{S_i}(x)=1$

if $x\in S_i$ and 0 otherwise for $i=1,2$

이다[1].

3. FST를 이용한 유전자 식별

3.1 퍼지 소속도

기존의 퍼지 소속함수를 이용하여 각각의 구성요소들에 대한 Fuzzy 소속도는 식(4), 식(5)를 이용하여 구할 수 있다.

$$f_{FS_1}(x)=e^{-(x-\mu_1)^2/2\sigma_1^2} \quad (4)$$

$$f_{FS_2}(x)=e^{-(x-\mu_2)^2/2\sigma_2^2} \quad (5)$$

여기서, μ_S는 각 그룹의 평균을, σ_S는 각 그룹의 표준편차를 나타낸다.

퍼지 소속함수인 $f_{FS_1}(x)$와 $f_{FS_2}(x)$를 이용하여 각각의 구성요소들이 그룹 S_1과 그룹 S_2에 속할 정도인 퍼지 소속도를 계산한다[1].

3.2 FST 유전자 식별

유전자 식별을 위한 마이크로어레이 연구에서는 일반적으로 수천에서 수만개에 이르는 유전자에 대해 유의검정을 실시하게 된다. 이 경우 유의확률 0.05 수준의 낮은 확률값에도 불구하고 0.05수준은 확률적으로 상당히 많은 수의 유전자가 유의한 유전자로 판정된다. 또한 많은 경우의 마이크로어레이 실험에서 사용되는 모표본의 수는 그다지 크지 않은 경우가 많다. 이 같은 경우 T-test는 극단값들(extreme value)에 의해 영향을 받게 된다.

여기서는 전통적인 T-test에 퍼지 소속도값을 이용한 수정인자(fudge factor)인 FST fc-value를 이용한 방법인 FST(Fuzzy Significance Test)를 제안하였다. FST 방법은 각각의 요소들에 대해 자신이 소속된 그룹에 대한 소속정도를 구하고 이를 이용하여 유의확률을 구하는 방법이다.

$$t_{FST}(S_1,S_2)=\frac{|\mu_{S_1}-\mu_{S_2}|}{\sqrt{\frac{\sigma_{S_1}^2}{|S_1|}+\frac{\sigma_{S_2}^2}{|S_2|}}}\times FC \quad (6)$$

$$FC(S_1,S_2)=\frac{\sum_{e\in S_1}f_{F(S_1)}(e)}{|S_1|}+\frac{\sum_{f\in S_2}f_{F(S_2)}(f)}{|S_2|} \quad (7)$$

여기서 μ_1과 μ_2는 각각의 그룹에 대한 평균이며, $\sigma_{S_1}^2$와 $\sigma_{S_2}^2$ 은 각각의 그룹에 대한 분산이다. 즉, 두 그룹의 평균의 차에 대한 정도를 구하고 여기에 퍼지 소속도를 계산하여 소속 정도에 따라 t 값을 수정하므로써 퍼지 소속도에 따른 통계값을 계산하였다.

4. 시뮬레이션 및 결과 고찰

4.1 사용된 데이터 및 시뮬레이션 방법

본 논문에서 insulin resistant 샘플 5개와 insulin sensitive 샘플 5개로 이루어진 마이크로어레이 유전자 발현 실험결과 데이터인 GSE121-GPL100_series_matrix.txt를 NCBI의 GEO 데이터베이스로부터 다운로드하여 사용하였으며, 통계분석 툴인 R을 사용하여 분석하였다[7,8]. T-test와 Wilcoxon rank sum test 및 여기서 제안한 FST-test 는 통계분석 툴인 R을 이용하여 계산하였으며, 퍼지이론을 기반으로 하는 마이크로어레이 유전자 발현에 관한 분석기법인 FM-test는 FM-test 전용 툴인 GeneDiff를 이용하였다[1,9].

4.2 분석결과 및 고찰

그림 1은 Null 값이 아닌 데이터 6086개 데이터에 대한 T-test에서의 p-value 와 Wilcoxon rank sum test에서의 p-value 및 본 논문에서 제안안 FST-test에서의 p-value의 분포를 나타낸 것이다. 일반적으로 유의수준 5%를 고려한다고 할 때 그림 1에서와 같이 유의수준 낮은 값일수록 T-test에 비해 FST-test의 값이 좀 더 엄격한 결과를 나타내는 것을 알 수 있다.

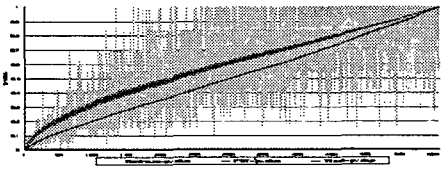

그림 1. T-test의 p-value, Ranksum의 p-value 및 FST-test에서의 p-value의 분포

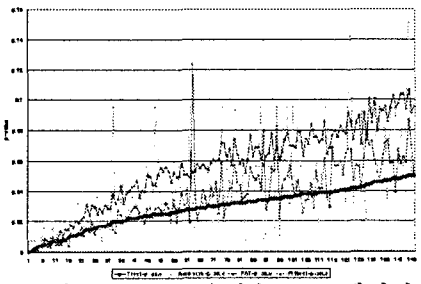

그림 2. T-test 유의수준 5%에서의 T-test p-value와 Ranksum p-value, FM-test p-value, FST-test p-value의 비교

그림 2와 그림 3은 T-test의 유의수준 5%와 10%수준에서의 다른 검정들의 유의확률분포를 나타내고 있다. 그림 2와 그림 3에서와 같이 본 논문에서 제안한 FST-test의 결과는 Wilcoxon Ranksum test에 비해 좀더 유연하고, T-test보다는 엄격한 결과를 나타내고 있다. 뿐만 아니라 T-test와 Wilcoxon rank sum test 모두에서 강하게 유의하다고 판정되는 것들에 대해서 FST-test에서도 유의하다고 판정되고 있다.

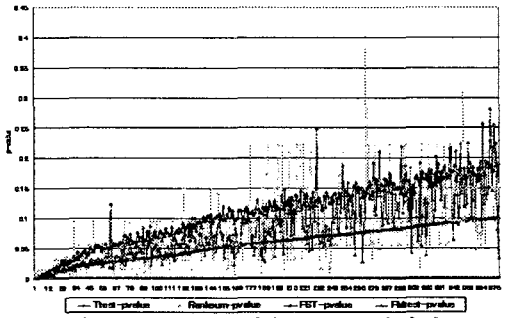

그림 3. T-test 유의수준 10%에서의 T-test p-value와 Ranksum p-value, FM-test p-value, FST-test p-value의 비교

또한 표 1에서와 같이 T-test와 Wilcoxon rank sum test에서 유의확률값이 크게 차이가 나는 경우인, 등분산 검정통계량이 5%이하인 데이터들에 대해서도 FST는 신뢰할 만한 결과를 나타내고 있다.

표 1. 등분산검정에 따른 각 검정별 유의확률

	p-value				
ID_REF	var.test	Ttest	Ranksum	FST	FM
M13058_s_at	0.004	0.02	0.095	0.038	0.015
Z29481_at	0.048	0.03	0.016	0.061	0.028
S62539_at	0.017	0.031	0.016	0.072	0.022
HG1071-HT1071_at	0.05	0.032	0.093	0.074	0.052
U36221_at	0.027	0.034	0.056	0.065	0.033
M80482_at	0.041	0.039	0.095	0.065	0.038
D31884_at	0.057	0.042	0.151	0.086	0.071
S71824_at	0.022	0.054	0.032	0.113	0.064
X78031_at	0.005	0.056	0.095	0.113	0.092
D80004_at	0.025	0.057	0.095	0.115	0.112
HG4018-HT4288_at	0.052	0.059	0.032	0.115	0.103
M86699_at	0.037	0.059	0.075	0.11	0.059
S90469_at	0.044	0.065	0.032	0.115	0.074
HG3395-HT3573_s_at	0.055	0.071	0.016	0.128	0.060
S69265_s_at	0.04	0.071	0.016	0.115	0.061
U61734_s_at	0.008	0.071	0.116	0.137	0.054
M28983_at	0.053	0.074	0.056	0.128	0.086
U49835_s_at	0.011	0.077	0.151	0.134	0.028
HG2280-HT2376_at	0.032	0.079	0.222	0.167	0.114
HG2809-HT2920_s_at	0.004	0.099	0.151	0.192	0.141
U48251_at	0.002	0.1	0.151	0.191	0.034

현재 마이크로어레이 유전자 발현에 관한 연구는 통계학적 유의수준 테스트에 의한 유의한 유전자 발굴뿐만 아니라 유전자 클러스터링, 유전자 네트워트 구축 등의 연구가 활발히 진

행 중에 있다. 여기서는 퍼지이론을 바탕으로 기존의 유의수준 테스트를 보완한 새로운 기법의 유의수준 검정 방법인 FST-test를 제안하였다. 본 논문에서 제안한 방법은 기존의 T-test와 Wilcoxon rank sum test 등의 결과를 크게 벗어나지 않고 있으면서 T-test 결과보다는 엄격하고, Wilcoxon rank sum test에 비해서는 유연한 결과를 보여주고 있다. 또한 등분산 5%정도의 비모수 데이터들에 대해서도 전통적인 방법의 결과를 크게 벗어나지 않음으로써 두 방법의 문제점을 보완하였다. 이러한 결과를 바탕으로 향후 유의한 유전자 발굴뿐만 아니라 유전자 네트워크 구축에도 활용하고자 한다.

참 고 문 헌

[1] L.R.Liang, S.Lu, X.Wang, Y.Lu, V. Mandal, D.Patacsil, D. Kumar, "FM-test: a fuzzy-set-theory-based approach to differential gene expression data analysis," BMC Bioinformatics, Vol. 7(Suppl 4):S7, 2006.

[2] M.Schena, D.Shalon, R.W.Davis, P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", Science, Vol. 270 (#5235), pp.467-470, 1995.

[3] http://www.ncbi.nlm.nih.govAbout/primer/microarrays.html.

[4] Brazhnik P, de la Fuente A, Mendes P., "Gene networks: how to put the function in genomics", Trends in biotechnology, vol. 20(11), 467-472, 2002.

[5] V.G.Tusher, R.Tibshirani, G.Chu, "Significance analysis of microarrays applied to the ionizing radiation response" PNAS vol. 98(9), pp 5116 - 5121, 2001.

[6] L.Deng, J.Pei, J.Ma, D.Lee, "A Rank Sum Test Method for Informative Gene Discovery", In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), 22-25, 2004.

[7]ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SeriesMatrix/GSE121/

[8]http://www.r-project.org/

[9]http://database.cs.wayne.edu:8080/bioinformatics/index.jsp

[10]http://www.gene chips.com/