응용통계연구 (2012)
25(2), 269–277
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.2.269

유전자군 분석의 방법론과 응용

이태원[1] · Robert R. Delongchamp[2]

[1]고려대학교 정보수학과
[2]Department of Epidemiology, University of Arkansas for Medical Sciences

(2011년 2월 28일 접수, 2011년 4월 4일 수정, 2011년 4월 7일 채택)

요 약

마이크로어레이 분석은 특이 발현하는 개별적인 유전자보다 유전자 온톨로지(Gene Ontology)와 같이 기능적 분류나 생물학적 경로(pathway)와 관련된 유전자군을 찾아내는 것이 그 해석의 용이성 때문에 최근 더욱 많은 연구가 진행되고 있다. 약물 처리에 의한 생물학적 반응을 연구할 때, 한 유전자군에 속하는 유전자들 각각의 특이 발현 여부의 유의성을 나타내는 p-value들을 취합하여 그 유전자군의 유의성을 결정하는 통계 검증 방법을 본 논문에서 소개하였다. 본 논문에 제시된 유전자군 분석(Gene group analysis) 방법은 Fisher's exact test나 permutation test와 같은 기존의 대표적인 방법들보다 더 정확하고 적용범위가 넓음을 실재 생물학 실험 자료의 분석을 통해 보였다. 제시된 유전자군 분석 방법은 SAS 프로그램으로 구현되었고 저자의 홈페이지(http://cafe.daum.net/go.analysis)에서 내려 받아 사용할 수 있다.

주요용어: 마이크로어레이, 유전자 온톨로지, 유전자군 분석, p-value, SAS 프로그램.

1. 서론

생물 세포 내 수만 개의 유전자 발현 정도를 동시에 측정하는 마이크로어레이를 이용한 연구 방법이 여러 생물학 연구 분야에서 활발히 사용되고 있다. 그러나 마이크로어레이와 같은 유전자 데이터는 그 방대한 크기와 복잡함으로 인해, 데이터의 해석과 생물학적 의미도출이 많은 학자들에게 힘든 도전과제가 되어왔다. 독성학 연구와 같이 약물 처리군과 대조군의 유전자 발현의 차이가 연구의 주된 관심이 되는 경우, 마이크로어레이로 측정된 유전자들 중 특이 발현한 유전자(differentially expressed genes)를 개별적으로 찾아내어 그 유전자들의 세포 내 역활을 알아봄으로써 약물에 의한 생물학적 반응의 전체적인 기전을 이해하기는 쉽지 않은 일이었다. 그에 비해 유전자 온톨로지(Gene Ontology; GO)와 같이 기능적 분류나 생물학적 경로(pathway)와 관련된 유전자군이 약물 처리에 의해 대조군과 다르게 반응했는지 직접 알 수 있다면 마이크로어레이 자료의 해석이 쉬워진다.

기존의 많은 소프트웨어 (Cavalieri 등, 2007; Al-Shahrour 등, 2007; Backes 등, 2007)에 사용되는 Fisher's exact test는 한 유전자군에 속한 유전자들 중 특이 발현된 유전자들의 비율을 이용하여 그 유전자군의 유의성을 초기하분포에 따라 결정한다. 이 방법은 계산 속도는 빠르지만 유전자군에 대한 처리효과를 직접적으로 측정하기보다는 간접적으로 일정 수준(threshold) 이상 영향을 받는 유전자가 그

이 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2011-0015128).
[1]교신저자: (339-700) 충남 연기군 조치원읍, 고려대학교 세종캠퍼스 정보수학과, 조교수.
E-mail: taewon70@korea.ac.kr

유전자군에서 어느 정도 있는가를 측정하는 방법이다 (Draghci 등, 2003). 그리고, 각 유전자들의 유의성을 나타내는 p-value는 서로 독립이라는 가정을 전재로 한다. 그러나, 하나의 기능적 분류나 생물학적 경로와 관련된 유전자군에 속한 유전자들이 서로 독립적으로 약물에 반응한다는 것은 만족하기 힘든 가정이다 (이미성과 송혜향, 2009).

유전자군에 대한 처리효과를 직접적으로 측정하는 방법으로는 표본(sample)들의 실험군과 대조군의 구분(label)을 서로 치환(permutation)하는 랜덤검정법(randomization test)들이 많이 개발되었다 (Lee 등, 2005; Tian 등, 2005). 랜덤검정법들은 우선 한 유전자군에 속하는 유전자 전체의 약물 처리에 의한 효과를 대표하는 요약 통계량(summary statistic), (예: Subramanian 등 (2005)에서 enrichment score(ES), Efron과 Tibshirani (2007)에서 average z-score)을 계산하고, 표본들의 실험군과 대조군의 구분을 서로 치환하여(sample permutation) 귀무가설 (Tian 등 (2005)에서 Q2) 아래에서의 요약 통계량의 분포를 구해서, 그 유의성을 계산한다. 이 방법은 한 유전자군에 속하는 유전자들 사이의 상관관계를 반영하는 좋은 성질이 있다. Fisher's exact test나 Kolmogorov-Smirnov test와 같이 유전자들 사이의 구분을 서로 치환하여(gene permutation) 귀무가설 (Tian 등 (2005)에서 Q1) 아래에서의 그 요약 통계량의 유의성을 계산하는 방법도 많이 연구되고 있지만 이런 방법들은 유전자들 사이의 상관관계를 감안하여 유의성을 계산하지 않는다.

동물 실험을 통한 표본을 사용해서 마이크로어레이 실험을 하는 많은 경우, 표본 갯수의 제한과 정밀하고 구조적인 실험 계획으로 인해 서로 치환할 수 있는 동등한 표본의 수가 적을 수 있다. 이럴 경우 sample permutation 방법은 귀무가설(Q2) 아래에서의 유의성을 충분히 정밀하게 계산할 수 없다. Delongchamp 등 (2006)은 이런 경우에도 적용할 수 있으면서 유전자들 사이의 상관관계를 감안한 유전자군 분석 방법을 제안하였다. 귀무가설 아래에서 계산된 p-value들은 균등분포(uniform distribution)를 따르는데, 여러 개의 p-value들을 정규분포(normal distribution)를 따르도록 치환한 다음 그것들을 합하면 하나의 정규분포를 따르는 하나의 확률 변수가 된다는 사실을 이용하여 유전자군의 유의성을 구하는 방법이다. Lee 등 (2008)은 일표본 t 검정뿐만 아니라 여러개의 처리군을 가지는 고정효과 선형모형 아래에서 처리효과의 유의성을 나타내는 각 유전자의 p-value들을 취합하여 유전자군의 유의성을 구할 수 있도록 그 방법을 확장하였다.

본 논문에서는 앞으로 실재 독성학 실험 자료의 분석에 사용된 예를 이용하여 Delongchamp의 방법을 설명하고, 다른 방법들이 갖지 못한 그 장점을 보이며, 다른 연구자들이 마이크로어레이 자료를 분석할 때 사용하도록 SAS로 구현된 프로그램을 제시할 것이다.

2. 통계적 방법

n의 표본을 이용하여 여러 종류의 약물 처리에 의한 유전자의 세포내 변화를 관찰하는 실험을 나타내는 고정효과 선형모형에서, 각 유전자의 발현을 나타내는 반응변수 $\boldsymbol{y}$는 $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$으로 나타낼 수 있다. 여기에서 $\boldsymbol{\epsilon}$은 $\mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I_n})$를 따르는 오차들의 $n \times 1$ 벡터이고 $\boldsymbol{X}$는 계수(rank)가 r인 계획 행렬(design matrix)이다. 계수 $\boldsymbol{\beta}$와 오차 분산 σ^2의 불편추정량은 각각

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$
$$\widehat{\sigma}^2 = \frac{1}{n-r}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$$

이고, 처리효과의 유의성을 나타내는 p-value는 두 처리군들 사이의 대비(contrast)가 0이라는 귀무가설

($\boldsymbol{H}_0 : \boldsymbol{c\beta} = 0$) 아래에서 통계량

$$T = \frac{\boldsymbol{c}\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}\sqrt{\boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}'}}$$

가 자유도 $n-r$의 t 분포를 따름을 이용하여 구한다.

m개의 유전자를 가진 유전자군에 대한 처리효과의 유의성은 위에서 구한 각 유전자들의 p-value들 $(p_k, k = 1, \ldots, m)$을 치환하고 합하여 만들어진 정규분포를 따르는 하나의 확률 변수로부터 구할 수 있다. 즉 Φ가 누적정규분포함수일때, 귀무가설 아래에서 $(z_k = \Phi^{-1}(1-p_k),\ k = 1, \ldots, m)$들은 각각 표준정규분포 $\mathcal{N}(0,1)$를 따르고 그들의 합 $\sum_{k=1}^{m} z_k/\sqrt{m}$도 p_k들이 서로 독립일 경우 표준정규분포를 따른다. 그러나 같은 세포내 기능으로 묶여진 하나의 유전자군에 속하는 유전자들이 서로 독립적으로 약물에 반응하지는 않으므로 p_k들 사이의 상관관계를 감안하여 처리효과를 검증하여야 할 것이다.

각 유전자들에 대한 처리효과의 유의성을 나타내는 p_k들이 대립가설 ($\boldsymbol{H}_1 : \boldsymbol{c\beta} > 0$) 아래에서 단측검정으로 구해졌고 $(z_1, \ldots, z_m)$의 공분산 $\mathbf{R}$을 알고 있을 경우,

$$\mathbf{1}'\boldsymbol{z} = \sum_{k=1}^{m} z_k = \sum_{k=1}^{m} \Phi^{-1}(1-p_k) = \sum_{k=1}^{m} T_k$$

의 분산이

$$\mathrm{Var}(\mathbf{1}'\boldsymbol{z}) = \sum_{k=1}^{m} \mathrm{Var}(T_k) + 2\sum_{s>t} \mathrm{Cov}(T_s, T_t) = \mathbf{1}'\mathbf{R}\mathbf{1}$$

임을 알 수 있다. 이것은 $r_{s,t}$가 $\mathbf{R}$의 s행 t열 원소일 때

$$\begin{aligned}
\mathrm{Cov}(T_s, T_t) &= \mathrm{Cov}\left(\frac{\boldsymbol{c}\widehat{\boldsymbol{\beta}}_{\boldsymbol{s}}}{\sigma_s\sqrt{\boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}'}}, \frac{\boldsymbol{c}\widehat{\boldsymbol{\beta}}_{\boldsymbol{t}}}{\sigma_t\sqrt{\boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}'}} \right) \\
&= \frac{1}{\sigma_s\sigma_t\boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}'} \mathrm{Cov}\left(\boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_{\boldsymbol{s}}, \boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_{\boldsymbol{t}} \right) \\
&= \frac{1}{\sigma_s\sigma_t\boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}'} \boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \mathrm{Cov}\left(\boldsymbol{y}_{\boldsymbol{s}}, \boldsymbol{y}_{\boldsymbol{t}} \right) \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}' \\
&= \frac{\sigma_{s,t}}{\sigma_s\sigma_t\boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}'} \boldsymbol{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}' \\
&= \frac{\sigma_{s,t}}{\sigma_s\sigma_t} = r_{s,t}
\end{aligned}$$

임을 보면 알 수 있다. 따라서 유전자군에 대한 처리효과의 유의성을 나타내는 p-value는

$$p = 1 - \Phi\left(\frac{\mathbf{1}'\boldsymbol{z}}{\mathbf{1}'\mathbf{R}\mathbf{1}} \right)$$

로 구한다. 실재로는 $\mathbf{R}$을 알지 못하므로 그것의 추정량 $\widehat{\mathbf{R}}$을 사용한다. 여기에서 $m \times m$행렬 $\widehat{\mathbf{R}}$의 s행 t열 원소는

$$\widehat{\sigma}_{s,t} = \frac{1}{n-r}\left(\boldsymbol{y}_s - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_s \right)' \left(\boldsymbol{y}_t - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_t \right)$$

로 추정한다.

또, p_k들이 대립가설 ($\boldsymbol{H}_1 : \boldsymbol{c\beta} \neq 0$) 아래에서 양측검정으로 구해졌을 경우, $\mathrm{MVN}(\mathbf{0}, \mathrm{Cov}(\boldsymbol{z}))$를 따르도록 발생시킨 $(z_1, \ldots, z_m)$의 랜덤표본을 사용하여 구한 경험적 분포로 요약 통계량 $\mathbf{1}'|\boldsymbol{z}| = \sum_{k=1}^{m} |z_k|$의 분포를 추정하여 유전자군에 대한 유의확률 p-value를 계산하는 몬테카를로 방법을 사용한다. 여기에서 $\mathrm{Cov}(\boldsymbol{z})$로는 그 추정량 $\widehat{\mathbf{R}}$, 또는 $\widehat{\mathbf{R}}$의 대각원소 이외의 원소들을 그 평균으로 대체한 $\overline{\mathbf{R}}$을 사용한다.

표 **3.1.** 4가지 처리군의 표본의 수와 특이 발현된 유전자의 수

구분	A	B	C	D
처리	saline	cisplatin	testosterone^{+} saline	testosterone^{+} cisplatin
표본의 수	3	2	3	3
특이발현 유전자 수	362		152	

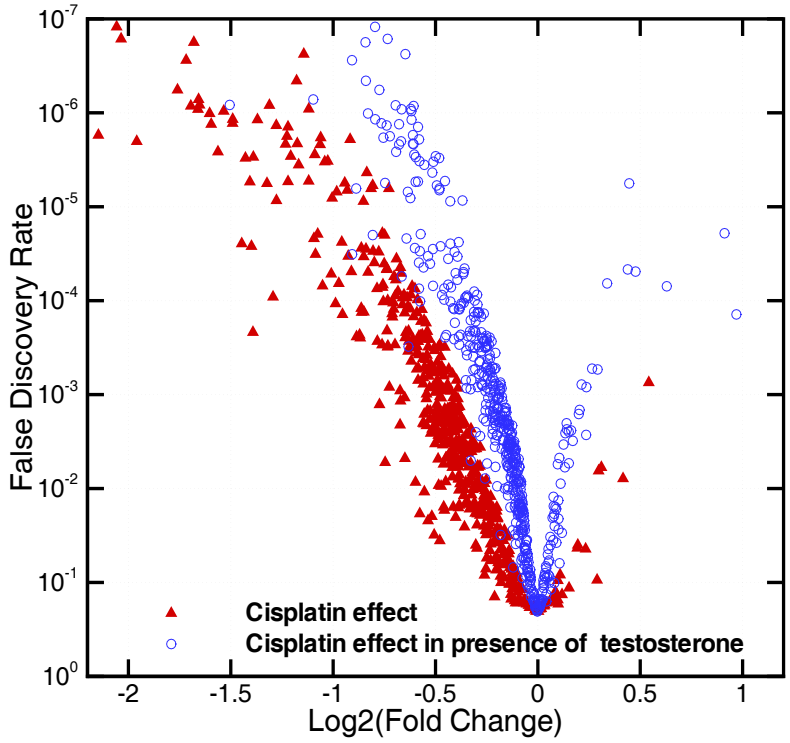

그림 **3.1.** Testosterone 사전 처리가 cisplatin 처리에 의한 유전자 발현의 변화를 경감시킴.

3. 적용 사례

본 논문에서 사용한 마이크로어레이 자료는 미국 식품의약국의 독성학 연구소(National Center for Toxicological Research; U.S. FDA)에서 개발된 미토콘드리아(mitochondria)와 관련된 534개의 유전자 발현을 측정하는 MitoChip (Desai 등, 2007)을 사용하여 쥐들의 여러 약물에 대한 반응을 실험한 자료이다.

표 3.1은 12마리의 쥐에게 4가지 종류의 약물처리를 하여 testosterone을 사용한 사전 처리(pre-treatment)가 cisplatin으로 인한 신장의 손상에 어떠한 영향을 미치는지 알아보는 실험을 나타낸다 (Li 등, 2009). 약물 처리에 영향을 받지 않는 housekeeping 유전자들과 식물인 arabidopsis의 유전자들의 평균 발현을 공변량으로 사용하는 표준화 방법을 사용하여, testosterone을 사용한 사전 처리 유무에 따른 cisplatin 처리효과를 고정효과 선형모형 아래에서 대비효과(constrast)로 측정하였다.

측정된 534개의 유전자들 중 많은 수가 표 3.1에서 볼 수 있듯이 처리군들(B, D)에서 대조군들(A, C)과

표 **3.2.** Fisher's exact test와 Delongchamp의 방법에 의한 유전자군에 대한 처리효과의 유의성 검정

유전자군	유전자 갯수	p-value < 0.05	Fisher's exact test	Delongchamp	p-value < 0.05	Fisher's exact test	Delongchamp
		cisplatin only			cisplatin on testosterone		
Complex1	27	24	0.0083	0.0002	7	0.6501	0.0864
Complex2	3	2	0.7486	0.0501	0	1.0000	0.1683
Complex3	7	4	0.8367	0.0034	2	0.6172	0.0876
Complex4	12	8	0.6486	0.0014	3	0.6861	0.0918
Complex5	13	13	0.0053	0.0001	2	0.9124	0.1376
lipid metabolism	42	25	0.8982	4.2E-5	14	0.2392	0.0030
TCA cycle	15	11	0.4204	0.0037	1	0.9926	0.1511
total	534	359			147		

다르게 발현되었다(FDR < 0.05). 그림 3.1은 534개의 유전자들의 false disconvery rate(FDR)을 fold change에 대응하여 나타낸 volcano plot이다. 여기에서 fold change는 cisplatin 처리군(B, D)에서 유전자 발현량의 대조군(A, C)에서 유전자 발현량에 대한 비율이다. 따라서 volcano plot에서 하나의 유전자에 대응하는 점의 위치가 fold change는 0에서 멀수록(그림 좌우 방향) false discovery rate는 작을수록(그림 위쪽 방향) 유의하게 처리군과 대조군에서 다르게 발현된 유전자이다. 붉은 점들(A와 B의 비교)과 푸른 점들(C와 D의 비교)의 위치를 비교하여 보면 testosterone 사전 처리를 하지 않은 집단(A, B)에서 많은 수의 유전자들이 cisplatin 처리군(B)에서 대조군(A)보다 약하게 발현되었고(fold change < 1), testosterone 사전 처리는 그 효과를 경감시킴을 알 수 있다. 즉 testosterone 사전 처리를 한 집단(C, D)에서 많은 수의 유전자들이 처리군(D)에서 대조군(C)보다 약하게 발현되었지만 그 정도는 testosterone 사전 처리를 하지 않았을 때(A, B)보다 약해졌음을 알 수 있다.

생물체의 에너지원인 ATP를 생성하는 세포 내 기관인 미토콘드리아의 역할과 연관된 생물학적 경로인 TCA cycle, 지방 대사(lipid metabolism) 및 산화적 인산화 과정의 다섯 복합체들(oxidative phosphorylation complexes I-V)의 7개 유전자군에 대한 Cisplatin 처리 효과의 유의성(p-value)을 Fisher's exact test와 Delongchamp의 방법을 사용하여 검정하였다. 표 3.2를 보면 Delongchamp의 방법으로 유의성 검증을 하였을 때 testosterone 사전 처리는 cisplatin 처리효과를 감쇠시키는 volcano plot의 결과와 동일한 해석을 할 수 있다. 그러나, Fisher's exact test로는 유전자군에 대한 처리효과의 유의성을 검정하기보다는 그 유전자군에서의 특이 발현 유전자들의 비율이 전체에서의 특이 발현 유전자들의 비율과 비슷한지 여부를 검정한 것임을 볼 수 있다.

MitoChip과 같이 소규모의 마이크로어레이의 경우 이 실험에서와 같이 많은 비율의 유전자들이 특이 발현할 경우가 많다. 이 경우 Fisher's exact test와 같이 gene permutation을 하는 랜덤검정법으로는 유전자군에 대한 처리효과를 직접 검정하기는 어렵다. 그리고, 이 실험에서와 같이 표본의 수가 작을 경우 sample permutation을 하는 랜덤검정법으로는 서로 치환할 수 있는 동등한 표본의 쌍이 적어서 정밀한 유의성 검정을 할 수 없으므로 적용할 수 없다.

그림 3.2는 생쥐에 14일 동안 600ppm농도의 usnic acid를 처리한 후 간세포의 미토콘드리아 속에서 작용하는 여러 가지 대사 경로들 사이의 상호작용을 보여준다 (Joseph 등, 2009). 굵은 화살표는 Delongchamp의 방법으로 유전자군 분석을 하였을 때 usnic acid 처리에 의해 유전자군의 작용이 화살표의 방향으로 통계적으로 유의하게 바뀌었음을 나타낸다. 그림 3.2에서 보이는 것처럼 usnic acid는 미토콘드리아 안으로 양성자를 수송하여 막간공간에서의 양전자 농도를 낮춰서 ATP생성을 방해한다. 그 결과로 전자 전달 연쇄(electron transport chain)에서 복합체V를 제외한 나머지 복합체들(I-IV)과 관

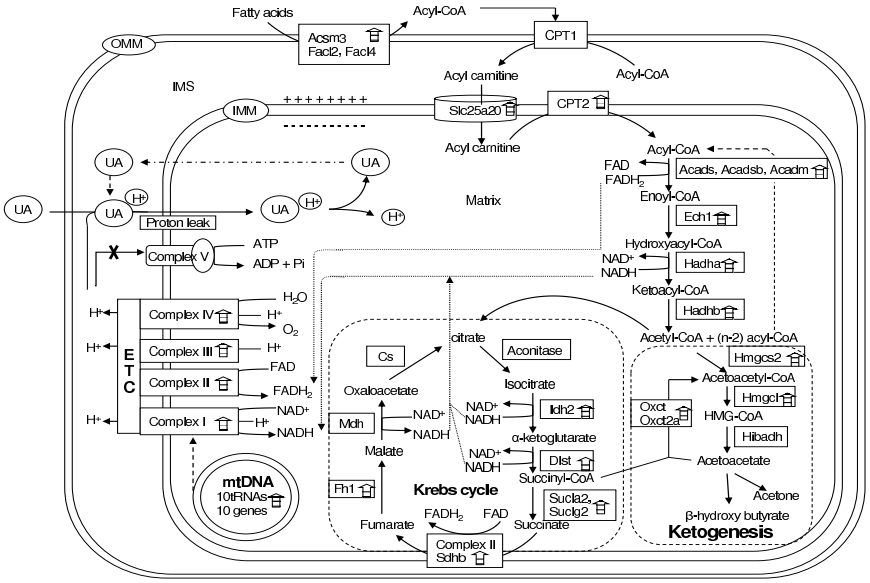

그림 **3.2.** usnic acid 처리에 의한 간세포의 미토콘드리아 속의 대사 경로들의 상호작용

련된 유전자들의 발현량이 증대된다. 이러한 변화는 또한 fatty acyl-CoA로의 지방산 활성, fatty acyl-CoA의 전달, 지방산 산화 경로(fatty acid β-oxidation)와 크렙스 회로(Krebs cycle)과 연관된 유전자들의 별현량 증가와 관련된다. 이러한 미토콘드리아 경로와 관련된 유전자들의 과발현은 전자 전달을 강화하기 위하여 복합체 I과 II에 NADH와 FADH2를 공급하고, 그 결과 usnic acid에 의한 양성자 농도의 저하을 개선해서 ATP를 생성시킨다. 본 연구자들은 유전자 온톨로지(Gene Ontology) 분석을 이용하여, 미토콘드리아 관련 유전자 발현 자료로부터 다양한 기능적 범주들을 분류함으로써 여러 가지 미토콘드리아 경로의 상호작용을 잘 이해할 수 있게 되었고, 간세포 미토콘드리아에 대한 usnic acid 효과의 기전을 규명할 수 있었다.

4. 소프트웨어

적용 사례의 유전자 집단 분석에 사용된 SAS 프로그램(GOanalysis)은 저자의 홈페이지(http://cafe.daum.net/go.analysis)에서 내려 받아 사용할 수 있다. GOanalysis는 3개의 SAS data set을 입력으로 받아들여서 Delongchamp의 방법으로 계산한 유전자군에 대한 유의성을 나타내는 p-value를 SAS data set으로 출력한다.

2개의 입력 SAS data set 중 하나는 SAS를 이용하여 일반화 선형모형(GLM procedure)으로 구한 각 유전자들에 대한 처리군들 간의 대비효과(contrast)의 추정값과 그 유의성을 나타내는 p-value들을 출력한 SAS data set이고, 다른 하나는 관찰값들을 고정효과 선형모형으로 적합한 잔차들을 출력한 SAS data set이다. 나머지 하나의 입력 SAS data set은 유전자군의 이름들을 go라는 열에, 그 유전자군에 속하는 유전자들을 gene이라는 열에 가지는 SAS data set이다. 이 정보는 주로 유전자 온톨로지 웹페이지(http://www.geneontology.org/)와 같은 곳에서 얻을 수 있고 GOanalysis는 실험에 사용된 마이크로어레이에 심어진 유전자들만 속해있는 유전자군들의 처리효과에 대한 유의성을 계산하여 출력한다.

5. 결론 및 토의

적용 사례의 그림 3.2과 같이 유전자군 분석에 의해 마이크로어레이 자료로부터 세포 내부의 생물학적 경로의 변화를 직접적으로 검증할 수 있게 됨으로써 약물처리에 의한 효과의 기전을 잘 이해할 수 있게 된다. Fisher's exact test와 같은 많은 유전자군 분석 방법은 한 유전자군에 속하는 유전자들에 대한 p-value의 분포와 마이크로어레이에 심어져 있는 전체 유전자들에 대한 p-value의 분포와의 비교를 통한 간접적인 방법인데 비해, Delongchamp의 방법은 한 유전자군에 속하는 유전자들 각각에 대한 처리효과를 나타내는 p-value들을 취합하여 그 유전자군에 대한 전체적인 처리효과를 측정하는 직접적인 방법이다. 또한 유전자들 사이의 상관관계를 직접 추정하여 유전자군 분석에 반영하기 때문에 다른 랜덤검정법들과 달리 표본의 수가 적을 때도 합리적인 유전자군 분석을 할 수 있다.

몇 번의 실험 자료 분석 경험에 의하면 (Desai 등, 2007; Desai 등, 2008; Desai 등, 2009; Li 등, 2009; Kashimshetty 등, 2009) 각 유전자마다 처리효과의 방향을 미리 알기는 어려우므로 양측검정 방법으로 p-value를 구하게 된다. 이때 m개의 유전자가 속한 유전자군에 대한 검정은 요약통계량 $\mathbf{1}'\boldsymbol{z}$의 분포를 $\mathrm{MVN}(\mathbf{0}, \mathrm{Cov}(\boldsymbol{z}))$를 따르는 $(z_1, \ldots, z_m)$의 랜덤표본을 이용하여 구하는 몬테카를로 방법을 사용하여 p-value를 계산한다. 여기에서 잔차에 의한 $\mathrm{Cov}(\boldsymbol{z})$의 추정량은 $\overline{\mathbf{R}}$을 사용하는 것이 좀 더 안정적이다.

연구자들이 마이크로어레이를 이용한 실험에서 일변량 또는 이변량 t 검정이나 고정효과 선형모형에서 대비효과의 검정으로 각 유전자에 대한 약물처리효과의 유의성을 검정하였을 때, 저자의 홈페이지에 있는 GOanalysis라는 SAS 프로그램을 사용하여 유전자군 분석을 할 수 있다. 현재 GOanalysis는 serial code이므로 17375개의 유전자로 이루어진 마이크로어레이 (Ayyadevara 등, 2009)와 연관된 2708개의 유전자군에 대한 분석을 Windows XP 개인용 컴퓨터의 Intel(R) Core(TM)2 Duo CPU E8400 @ 3.00GHz로 하였을 때 41시간 46분 44초(cpu 시간: 41시간 39분 52초)가 걸렸다. 앞으로 R package와 MATLAB의 parallel code로 Delongchamp의 방법을 구현하여 짧은 시간 안에 유전자군 분석을 할 수 있게 할 계획이다.

참고문헌

0이미성, 송혜양 (2009). 유전자 연관성이 랜덤검정 P값과 유의 유전자군의 탐색에 미치는 영향, <응용통계연구>, **22**, 781–792.

Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Minguez, P., Montaner, D. and Dopazo, J. (2007). From genes to functional classes in the study of biological systems, *BMC Bioinformatics*, **8**, 114.

Ayyadevara, S., Tazearslan, C., Bharill, P., Alla, R., Siegel, E. and Reis, R. J. S. (2009). Caenorhabditis elegans PI3K mutants reveal novel genes underlying exceptional stress resistance and lifespan, *Aging Cell*, **8**, 706–725.

Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y. A., Müller, R., Meese, E. and Lenhof, H. P. (2007). GeneTrail—advanced gene set enrichment analysis, *Nucleic Acids Research*, **35**, 186–192.

Cavalieri, D., Castagnini, C., Toti, S., Maciag, K., Kelder, T., Gambineri, L., Angioli, S. and Dolara, P. (2007). Eu.Gene analyzer a tool for integrating gene expression data with pathway databases, *Bioinformatics*, **23**, 2631–2632.

Delongchamp, R. R., Lee, T. and Velasco, C. (2006). A method for computing the overall statistical significance of a treatment effect among a group of genes, *BMC Bioinformatics*, **7**(Suppl 2), S11.

Desai, V. G., Lee, T., Delongchamp, R. R., Leakey, J. E. A., Lewis, S. M., Lee, F., Moland, C. L., Branham, W. S. and Fuscoe, J. C. (2008). NRTI-induced expression profile of mitochondrial genes in the mouse liver, *Mitochondrion*, **8**, 181–195.

Desai, V. G., Lee, T., Delongchamp, R. R., Moland, C. L., Branham, W. S., Fuscoe, J. C. and Leakey, J. E. A. (2007). Development of mitochondria-specific mouse oligonucleotide microarray and validation of data

by real-timePCR, *Mitochondrion*, **7**, 322–329.

Desai, V. G., Lee, T., Moland, C. L., Branham, W. S., Tungeln, S. V., Beland, F. A. and Fuscoe, J. C. (2009). Effect of short-term exposure to zidovudine (AZT) on the expression of mitochondria-related genes in skeletal muscle of neonatal mice, *Mitochondrion*, **9**, 9–16.

Draghci, S., Khatri, P., Martins, R. P., Ostermeier, G. C. and Krawetz, S. A. (2003). Global functional profiling of gene expression, *Genomics*, **81**, 98–104.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes, *Annals of Applied Statistics*, **1**, 107–129.

Joseph, A., Lee, T., Moland, C. L., Branham, W. S., Fuscoe, J. C., Leakey, J. E., Allaben, W. T., Lewis, S. M., Ali, A. A. and Desai, V. G. (2009). Effect of (+)-usnic acid on mitochondrial functions as measured by mitochondria-specific oligonucleotide microarray in liver of B6C3F1 mice, *Mitochondrion*, **9**, 149–158.

Kashimshetty, R., Desai, V. G., Kale, V. M., Lee, T., Moland, C. L., Branham, W. S., New, L. S., Chan, E. C., Younis, H. and Boelsterli, U. A. (2009). Underlying mitochondrial dysfunction triggers flutamide-induced oxidative liver injury in a mouse model of idiosyncratic drug toxicity, *Toxicology and Applied Pharmacology*, **238**, 150–159.

Lee, H. K., Braynen, W., Keshav, K. and Pavlidis, P. (2005). ErmineJ: tool for functional analysis of gene expression data sets, *BMC Bioinformatics*, **6**, 269.

Lee, T., Desai, V. D., Velasco, C., Reis, R. J. S. and Delongchamp, R. R. (2008). Testing for treatment effects on gene ontology, *BMC Bioinformatics*, **9**(Suppl 9), S20.

Li, S., Nagothu, K. K., Desai, V. G., Lee, T., Branham, W. S., Moland, C. L., Megyesi, J. K., Crew, M. D. and Portilla, D. (2009). Transgenic expression of proximal tubule peroxisome proliferator-activated receptor-α a in mice confers protection during acute kidney injury, *Kidney International*, **76**, 1049–1062.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *PNAS*, **102**, 15545–15550.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S. and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies, *PNAS*, **102**, 13544–13549.

A Method for Gene Group Analysis and Its Application

Taewon Lee[1] · Robert R. Delongchamp[2]

[1]Department of Information and Mathematics, Korea University
[2]Department of Epidemiology, University of Arkansas for Medical Sciences

(Received February 28, 2011; Revised April 4, 2011; Accepted April 7, 2011)

Abstract

In microarray data analysis, recent efforts have focused on the discovery of gene sets from a pathway or functional categories such as Gene Ontology terms(GO terms) rather than on individual gene function for its direct interpretation of genome-wide expression data. We introduce a meta-analysis method that combines p-values for changes of each gene in the group. The method measures the significance of overall treatment-induced change in a gene group. An application of the method to a real data demonstrates that it has benefits over other statistical methods such as Fisher's exact test and permutation methods. The method is implemented in a SAS program and it is available on the author's homepage(http://cafe.daum.net/go.analysis).

Keywords: Microarray, Gene Ontology, gene group analysis, p-value, SAS program.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2011-0015128).

[1]Corresponding author: Assistant professor, Department of Information and Mathematics, Korea University, Jochiwon, Chungnan 339-700, Korea. E-mail: taewon70@korea.ac.kr