시계열 마이크로어레이 데이터 마이닝을 위한 분별력 있는 유전자 선정 방법

이민수[O,1], 박승수[1], 강성희[2], 박웅양[3]
이화여자대학교 컴퓨터학과[1], 명지대학교 방목기초교육대학[2], 서울대학교 의과대학[3]
ssue@ewhain.net[O,1], sspark@ewha.ac.kr[1], kangsh@mju.ac.kr[2], wypark@snu.ac.kr[3]

Selection of Discriminative Genes for Data Mining of Time-series Microarray Data

Min Su Lee[O,1], Seung Soo Park[1], Sung Hee Kang[2], Woong Yang Park[3]
Department of Computer Science and Engineering, Ewha Womans University[1]
Bangmok College of Basic Studies, Myongji University[2]
Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine[3]

요 약

본 논문에서는 시계열 마이크로어레이데이터 마이닝을 위한 전처리 작업으로 시계열 마이크로어레이 데이터에 특징 추출 방법 및 상관관계 분석을 이용하여 분화 과정에 대해 분별력 있는 유전자들을 선정하기 위한 방법을 제안하고, 줄기세포가 신경세포로 분화하는 과정에서 특이적으로 발현되는 유전자들을 찾기 위한 시계열 마이크로어레이 데이터 분석 과정을 하나의 예로 제시한다. 분석 결과, 제안한 방법이 분화 특이적으로 발현되는 분별력 있는 유전자들, 분화 과정에서 공통적으로 발현되는 유전자들, 그리고 경계선에 존재하는 유전자들을 통해서 줄기세포 신경분화의 특징들을 규명하는데 매우 유용함을 보였다.

1. 서 론

마이크로어레이(Microarray) 기술의 발전으로 처리 조건이나 환경에 따른 수만여 개 유전자들의 발현 양상을 보다 손쉽게 동정해볼 수 있게 되었다. 이 기술을 질병 진단 및 예측에 활용하기 위해 최근 각종 질병이나 노출 환경에 따라 특이적으로 발현하는 유전자를 찾기 위한 실험들이 이루어지고 있다.

더 나아가 시간 변화에 따른 유전자 발현량 변화를 동정해 볼 수 있는 시계열 마이크로어레이 실험을 통해 암의 진행상황이나 세포 분화 과정에 따른 유전자 발현 변화를 동정해볼 수도 있다. 시계열 마이크로어레이 데이터를 이용하면 시간에 따른 유전자 발현 양상의 추이를 살펴볼 수 있으므로 초기 상태에서 최종 상태로 진행하는 과정에 지속적으로 관련되는 유전자들을 발견할 수 있고, 은닉 마코프 모델(Hidden Markov Model)이나 베이지안 네트워크(Bayesian Network) 등의 알고리즘을 적용하면 유전자들 사이의 조절 매커니즘을 추정해볼 수도 있다는 장점이 있다. 더 나아가 시계열 마이크로어레이 실험에 시계열 데이터 마이닝 (Sequential Data Mining) 기법들 중 유사성 탐색 (Similarity Search)이나 순차 패턴 마이닝 (Sequential Pattern Mining)을 적용하면 질병의 진행 단계나 세포 분화 과정 등에 대한 특징을 규명하거나 예측 또는 분류 모델을 구축할 수 있다.

마이크로어레이 데이터는 속성(Attribute)에 해당하는 유전자들은 그 수가 수만 개 혹은 수십만 개에 달하는데 비해, 실험의 비싼 비용 및 샘플의 한정된 용량으로 인해 표본(Instance)에 해당하는 슬라이드 개수는 매우 적다는 특징이자 한계점을 가지고 있다. 따라서 연산 비용이 많이 소요되는 데이터 마이닝의 전처리(Preprocessing) 작업으로 표본 개수에 비해 너무 많은 속성들을 분석 목적에 맞는 개수로 줄여주는 작업이 수행되어야 한다.

일반적으로 사용되는 특징 추출(Feature Selection) 방법을 시계열 마이크로어레이 데이터에 적용하여 특정 질병의 발병 과정이나 특정 세포의 분화 과정에 관련된 유전자 리스트를 뽑으면, 도메인 특성 상 그 안에는 다른 질병의 발병 과정이나 다른 세포의 분화 과정에서도 공통적으로 발현되는 유전자들이 포함되어 있다. 따라서 특정 질병 발병과정이나 세포 분화 과정에서만 특이적으로 발현되거나 다른 과정과는 발현 양상이 명확하게 구분되는 유전자 리스트를 선택하기 위한 작업이 추가로 요구된다.

본 논문에서는 시계열 마이크로어레이 데이터 마이닝을 위한 사전 작업으로도 활용될 수 있는 특정 진행 과정에 특이적인 유전자들을 구분하기 위한 분석 방법을 제안하고, 줄기세포가 신경세포로 분화하는 과정에 분별력 있는 유전자들을 찾기 위한 시계열 마이크로어레이 데이터 분석 과정을 하나의 예로 제시한다.

2. 분화 특이적인 유전자들을 구분하기 위한 분석 방법

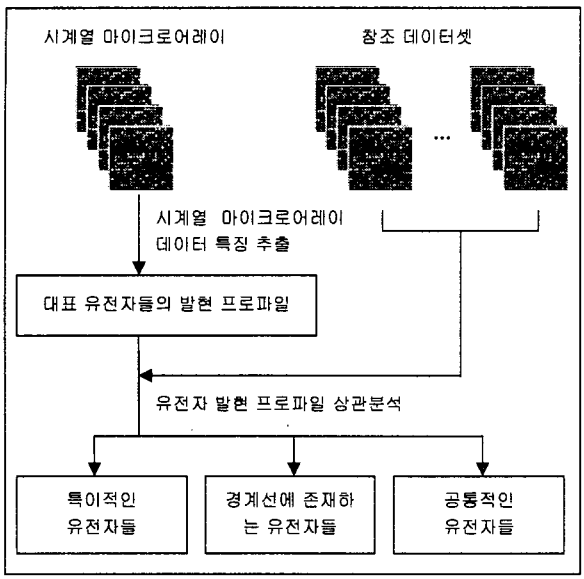

그림 1. 시계열 마이크로어레이 데이터에서의
분별력 있는 유전자 선정을 위한 과정

특정 분화 과정에 분별력 있는 유전자들을 찾기 위해서는 특징 추출을 통해 얻은 해당 모델을 대표하는 유전자들의 프로파일을 다른 분화 과정에서의 발현 프로파일들과 비교함으로써, 특징으로 선택된 유전자들이 실제 다른 데이터에서는 발현 양상이 어떠한지, 실제로 그 모델을 분별하는데 사용할 수 있는지 등을 검증하는 작업을 수행해야 한다(그림 1). 그러기 위해서는 분석 대상이 되는 시계열 마이크로어레이 데이터와 관련있는 다른 과정에 대한 시계열 마이크로어레이 데이터를 통합하여 분석하여야 한다. 분석 대상이 되는 시계열 마이크로어레이 데이터에서 특징추출 알고리즘을 적용하여 그 데이터를 대표할 수 있는 유전자들의 집합을 선택한 후, 참조 시계열 마이크로어레이 데이터셋들과 각 유전자 발현 프로파일들을 상관성을 분석한다. 다른 분화 과정에서의 발현 프로파일과 강한 음의 상관관계를 가지는 유전자들을 해당 분화 과정을 분별할 수 있는 유전자로 선택하여 분화의 특징 및 원인 규명 등에 이용하고, 양의 상관관계가 높은 유전자들은 여러 분화 과정에서 공통적으로 나타나는 유전자들로 분화과정의 공통적으로 나타나는 특징을 분석하며, 나머지 유전자들은 경계선에 존재하는 유전자들로 분류할 수 있다.

3. 신경세포 분화에 대해 분별력 있는 유전자 선정

시계열 마이크로어레이에서 분별력 있는 유전자들을 선정하기 위한 분석의 예로, 본 논문에서는 배아줄기세포를 신경세포로 분화시키는 과정에서 특이적으로 발현되는 유전자를 찾는 것을 목적으로 하는 실험 데이터를 사용하였다[1]. 줄기세포를 신경세포로 분화하는 과정에서 발현되는 유전자들에는 일반적으로 다른 분화 과정에서도 발현되는 유전자들을 포함하고 있다. 따라서 신경세포로의 분화에만 관련된 유전자들을 발견하기 위해서는 배아 줄기세포가 신경세포 외의 다른 세포로 분화될 때에도 발현되는 유전자들을 제거해주는 작업이 필요하다.

신경세포 분화에 특이적인 유전자들만 선택하기 위하여 배아줄기세포를 신경세포로 분화할 때 발현되는 유전자 프로파일을 배아줄기세포를 랜덤하게 분화시켰을 때 나타나는 유전자 프로파일과 비교한다.

3.1. 데이터 셋 및 실험 디자인

생쥐의 배아줄기세포(1단계)를 도파민에 반응하는 신경세포로 유도 분화(Guided Differentiation, GD)시키면서 총 5단계에 걸쳐 샘플링한 샘플들과 각 단계에 대응되도록 디자인한 배아줄기세포를 배아체(Embryoid Body, EB)로 랜덤 분화(Random Differentiation, RD)시키면서 5단계에 샘플링한 샘플들을 Macrogen의 Mouse Oligo 11K Chip에 올려 11,376여 개의 유전자 발현량을 동정한 마이크로어레이 데이터를 예제로 사용한다. 이 데이터는 배아줄기세포에서의 유전자 발현 값을 기준으로 RD와 GD의 나머지 4단계에 해당하는 샘플의 유전자 발현 값의 비율을 측정하는 two-dye 실험으로 디자인 되었으며 각각 3번 반복실험을 한 총 24장의 마이크로어레이 데이터이다.

3.2. 분석

3.2.1. 데이터 정규화

GenePix Pro를 이용하여 수치화된 유전자 발현 값들은 R 시스템의 vsn 패키지를 이용하여 슬라이드 간에 전체적인 정규화 처리를 해준 후[2], print-group Lowess 정규화 처리를 해줌으로써 프린트 팁 위치에 따른 바이어스를 교정해 주었다[3].

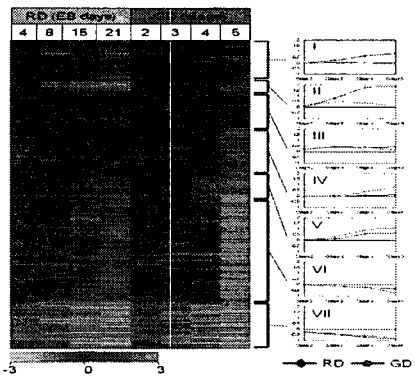

그림2. K-평균 군집화 결과

3.2.2. 발현 프로파일의 시각적 비교를 위한 군집화

정규화한 GD와 RD 데이터에 시계열 데이터의 유사성 척도로 일반적으로 사용되는 유클리드 거리(Euclidean Distance)를 사용하여 K-평균 군집화 (K-Means Clustering) 알고리즘을 적용해봄으로써 GD와 RD의 전반적인 유전자 발현 프로파일의 특징을 시각화하여 살펴보았다. 유전자 수가 많기 때문에 발현량의 변화가 적은 유전자들은 군집화 작업에서 제외하였다. 각 독립된 3번 실험에서의 유전자 발현 비율의 평균값을 구하고 적어도 하나 이상의 스테이지에서 1.5-fold 이상 발현 값이 변화한 1,884개의 유전자들을 7개의 군집으로 군집화 하였다(그림 2).

7개 군집들 중 1,282개의 유전자 (68%)들이 포함되어있는 4개의 군집들은 RD와 GD에서 비슷한 양상으로 발현량이 증가(C3와 C5)하고 발현량이 감소(C6와 C7)하는 경향을 보여준다. 이것은 GD와 RD사이의 유전자 발현 패턴이 매우 많은 부분에서 겹친다는 것을 반영한다고 할 수 있다.

3.2.3. 특징 추출 (Feature Selection)

분화 진행에 따라 특이적으로 발현량이 변화한 유전자들(Differentially Expressed Genes, DEGs)을 찾기 위해 정보 획득, 지니 지표, 주성분분석, 유전자 알고리즘 등과 같은 정보공학적 특징 추출 방법이 사용될 수도 있으나, 본 논문에서는 GD 모델에서의 특징적 유전자를 선별하기 위해 다변량 순열 검정(Multivariate Permutation Test) F-test를 이용하였다. 오류 발견 비율(False Discovery Rate)이 10% 이하이면서 90%의 신뢰도를 가지도록 평가한 결과, GD 모델에서 622개의 유전자들이 특징으로 추출되었다.

3.2.4. 상관분석을 통한 분별력 있는 유전자 선정

3.2.2.절에서 살펴본 바와 같이 GD 모델과 RD 모델은 많은 부분에서 데이터 발현 양상이 매우 비슷하다. GD 모델에서의 특징적인 유전자들로 선별된 리스트에는 다른 분화 과정에서도 비슷한 양상으로 발현되는 유전자들이 포함되어있을 수 있으므로, GD 모델에 특이적인 유전자들을 고르기 위해서 GD 모델의 특징적 유전자들의 발현 프로파일을 RD모델에서의 발현 프로파일과 비교하는 상관분석을 수행하였다.

Spearman 상관분석은 모수적(Parametric) 방법인 Pearson 상관분석과는 달리 값 자체보다는 두 변수가 가지는 값들의 순위 간에 직선적 관계가 있는지 여부를 확인함으로써 두 변수 사이의 연관성의 강도를 측정하는 비모수적 (Non-parametric) 방법이다. 상관계수는 -1에서 1사이의 값을 가지며 상관계수가 1에 가까울수록 양의 상관관계를, -1에 가까울수록 음의 상관관계를, 0에 가까울수록 아무 상관관계가 없음을 나타낸다.

우리는 Spearman 상관계수(ρ_s)에 기반해서 GD 모델을 대표하는 622개의 유전자들의 발현 프로파일을 RD 모델에서의 유전자 발현 프로파일과 비교하였다. 그 결과 Spearman 상관계수에 따라 GD 모델을 대표하는 유전자들을 다음과 같이 나눌 수 있었다.

- $\rho_s \geq 0.6$, 두 분화 모델의 공통된 유전자: 622개의 유전자들 중 절반 이상인 400개(64.3%)의 유전자가 RD모델에서도 비슷한 발현 패턴을 보였다.
- $\rho_s \leq -0.4$, GD 모델에 특이적인 유전자: 66개의 유전자는 GD 모델에서만 특이적으로 발현된 것으로 나타났다.
- $-0.4 < \rho_s < 0.6$, 경계에 걸친 유전자: 이 구간의 유전자들은 K-평균 군집화의 C3과 많이 중복되는 유전자들로서 대부분이 발현 프로파일이 증가했다가 마지막 단계에서 감소하는 경향을 보였다. 즉 신경세포로의 분화에 따라 발현량이 감소하는 유전자들을 포함한다.

3.3. 생물학적 검증

이와 같은 세 구간에 해당하는 유전자 - GD 특이적 (Serpini1 ρ_s=-0.4, Rab33a ρ_s=-0.8), 경계성 (Cdk4 ρ_s=0.2, P4ha2 ρ_s=-0.2), 공통 유전자 (Sox4 ρ_s=0.6)에 대해 in vivo와 in vitro 실험을 통해 생물학적 검증을 하였다. 그 결과, GD 모델과 RD 모델에서 공통적으로 발현되는 유전자들은 신경 계열 뿐만 아니라 다른 조직에서도 많이 발현되는 것으로 나타났으며, GD 특이적인 유전자들은 배아 말기부터 발현량이 증가하다가 성체의 뇌에서도 계속 발현이 되고 있으며 뇌 관련 기관에서만 강하게 발현되는 것으로 나타났다. 경계성 유전자들은 배아에서는 발현되다가 성체의 뇌에서는 발현량이 감소하고, 뇌 이외의 다른 기관에서도 발현이 많이 되는 것으로 나타났다.

4. 결론

본 논문에서 시계열 마이크로어레이 데이터를 이용해서 특정분화과정을 분별하는데 사용할 수 있는 유전자들, 여러 분화 과정에서 공통적으로 발현되는 유전자들, 경계선에 존재하는 유전자들을 구분하고 특징화 할 수 분석 방법을 제안하였고 생물학적으로 그 효용성을 검증하였다. 본 논문에서 제안한 분석 방법은 질병의 발병 원인을 규명하여 이를 예방하거나 다양한 세포 분화 과정을 이해하는 등의 연구 분야에 활용할 수 있을 것이다.

참고문헌

[1] Lee MS, et al. 'Selection of Neural Differentiation-Specific Genes by Comparing Profiles of Random Differentiation', Stem Cells, 2006 (accepted)

[2] Huber W, et al. 'Variance Stabilization Applied to Microarray Data Calibration and the Quantification of Differential Expression', Bioinformatics 18(Suppl 1):S96-S104, 2002.

[3]Yang YH, et al. Normalization for cDNA Microarray Data' Technical Report, UC Berkeley, 2001