Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

대조적인 지역 클러스터 식별

Identification of Contrasting Local Clusters

이건명[1] · 이선아[1] · 황경순[1] · 이찬희[2]
Sun A Lee, Kyung Soon Hwang, Keon Myung Lee, Chan Hee Lee

[1]충북대학교 전기전자컴퓨터공학부, 충북BIT지방연구중심대학사업단
E-mail: kmlee@cbnu.ac.kr
[2]충북대학교 생명과학부

요 약

마이크로어레이 데이터는 여러 샘플들의 대량의 유전자들에 대한 발현정도를 표현하며, 이에 대한 분석을 통해서 생명현상에 대한 이해와 분석이 이루어지고 있다. 생명현상이 유전자의 발현에 많은 영향을 받는 것이 알려져 있기 때문에 실험 샘플 집단내에서 또는 실험 샘플 집단간에서 발현 특성이 대조적으로 나타나는 유전자의 집단을 추출하는 것이 유용한 경우가 있다. 이 논문에서 관심영역으로 선택된 영역에 대해서 대조적인 패턴을 갖는 집단을 알고리즘적으로 선택하는 방법을 제안한다.

키워드 : 클러스터링, 마이크로어레이, 생명정보학

1. 서 론

마이크로어레이 데이터 분석은 대규모로 유전자의 발현 패턴을 분석함으로써 유전자 수준에서의 생명현상에 대한 연구를 위해 사용되고 있다. 마이크로어레이는 유리, 필터 또는 실리콘 판 위에 유전자를 검출할 수 있는 많은 수의 프로우브를 붙여 놓거나 합성하여 놓아서, 동시에 많은 유전자에 대한 발현량을 측정할 수 있도록 한 것이다. 발현 패턴을 비교할 때 발현 정도가 대조적으로 나타나는 클러스터를 식별하는 것이 분석에서 유용한 경우들이 있다. 예를 들면, 특정 약물을 투여한 것에 대한 샘플에 대해서 발현량이 높은 유전자 집단에 대응하여, 발현량이 낮은 유전자 집단을 추출하면, 약물투여에 영향을 받는 유전자 집단을 식별할 수 있게 된다. 샘플집단 간에 하나의 샘플집단에 대해서 발현량이 높고, 다른 집단에서는 발현량이 낮은 것을 식별하게 되면, 집단간에 일관성있게 차별화되는 유전자 집단을 선택할 수 있게 된다.

2. 관련 연구

계층적 클러스터링 기법을 사용하여 유전자 집단과 샘플 집단에 대해서 각각 클러스터링을 하게 되면, (그림 1)과 같은 클러스터링된 결과를 얻을 수 있다. 그런데 그림에서 사각형으로 표시된 것과 같이 대조적인 지역 클러스터를 선택하기 위해서는 분석자가 시각적으로 선택해야 하고, 또한 대조되는 부분들이 하나의 집단으로 클러스터되어 있지 않은 경우가 많아 선택하는데 노력이 많이 요구된다.

비교되는 집단이 미리 지정되는 경우에는 집단간에 해당 유전자 또는 샘플들에 대해서 t-test 등과 같은 통계적 검증을 하여 차이가 나는 개체들을 확인할 수 있다. 그러나 t-test의 특성상 차이의 여부만을 판정해 주기 때문에 분석에서 요구되는 크게 대조되는지에 대해서는 충분히 판단할 수 없기 때문에 기대하는 결과를 얻기 곤란하다.

대조할 집단을 지정한 경우에 퍼지 패턴을 적용하여 대조되는 개체를 찾는 방법이 제안되어 있다.[1] 이 방법에서는 유전자 발현정도를 나타내는 축에 대해서는 대조되는 집단을 대표하는 퍼지 소속함수를 정의하여, 해당하는 퍼지 소속함수에 대한 만족정도를 판정하는 방법으로 정의된 패턴을 만족하는 개체를 선택할 수 있도록 한다. 이 방법은 대조할 기준 집단을 미리 지정하기 때문에 분석자가 대조 집단을 미리 선정해야 하는 부담이 있다. 따라서 이 방법은 약물 실험 등과 같이 비교 대상이 되는 집단이 분석시점에 미리 지정되는 경우에 유용한 것으로, 임의로 선택한 지역 클러스터에 대조되는 지역 클러스터를 찾는 데는 직접 적용하기 곤란하다.

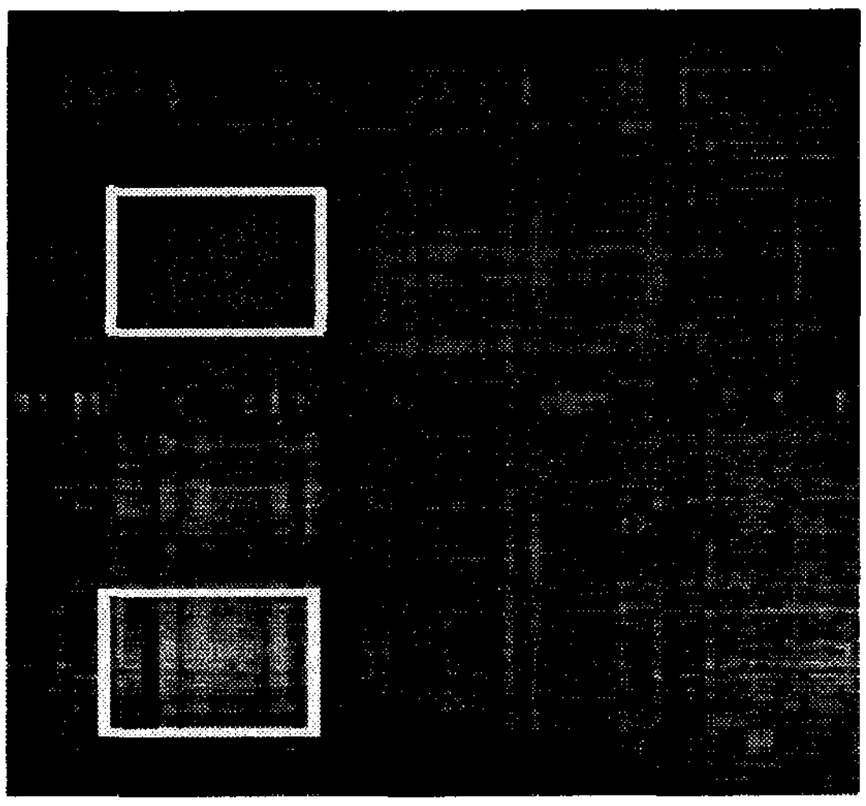

그림 1. 마이크로어레이 클러스터링 결과에 대해서 대조적 지역클러스터를 사각형으로 표시한 예

3. 제안한 대조적 지역 클러스터링 기법

제안한 방법에서는 일단 전체적으로 마이크로어레이 데이터를 계층적 클러스터링한 다음, 결과에 대해서 분석자가 대조적인 지역 클러스터를 찾기 위해 사용될 관심영역을 지정하도록 한다. 제안한 방법은 지정된 관심영역을 기준으로 대조되는 영역을 찾기 위해 다음과 같은 과정을 사용한다.

대조적인 지역 클러스터를 찾기 위해서 우선 마이크로어레이 데이터에 대해서 유전자 및 샘플에 대한 계층적 클러스터링을 수행한다. 입력 마이크로어레이 데이터 행렬은 M으로 나타내고, 클러스터링 결과로 정렬된 발현정도값 행렬은 E으로 나타낸다. 행렬에서 각 행은 하나의 유전자에 대응하여, 각 열은 하나의 샘플에 대응하고, 원소 e_{ij}는 i번째 유전자 g_i의 j번째 샘플에서 발현정도를 나타낸다.

$$E=\begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ e_{21} & e_{22} & \cdots & e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nm} \end{pmatrix}$$

행에 대응하는 유전자 이름은 $G=(g_1,g_2,\cdots,g_n)$로 나타내고, 열에 대응하는 샘플 이름은 $S=(s_1,s_2,\cdots,c_m)$으로 나타낸다. 편의상 유전자 i의 전체 샘플에 대한 발현정도는 $G_i=(e_{i1},e_{i2},\cdots,e_{im})$로 나타내고, 샘플 j의 전체 유전자 집합에 대한 발현정도는 $S_j=(e_{1j},e_{2j},\cdots,e_{nj})^t$로 나타낸다.

클러스터링 결과 E로부터 관심 영역을 선정한다. 관심영역(interest region)은 비슷한 발현정도를 갖는 클러스터링 결과에서 사각형 영역으로서, 해당 부분과 유사한 유전자 및 샘플 집단을 군집화하기 위한 기준(reference)에 대한 정보를 제공하는 역할을 한다. 선택된 관심영역을 다음과 같은 부분 행렬$B_{[a,b:c,d]}$로 나타낸다.

$$B_{[a,b:c,d]}=\begin{pmatrix} e_{ac} & \cdots & e_{ad} \\ \vdots & \ddots & \vdots \\ e_{bc} & \cdots & e_{bd} \end{pmatrix}$$

선정된 관심영역 $B_{[a,b:c,d]}$에 대한 평균 m과 행별 편차의 평균 rv_m을 계산한다.

$$m=mean\{e_{st}|e_{st}\in B_{[a,b:c,d]}\}$$
$$rv_m=mean\{variance_i\{e_{ic},\cdots,e_{id}\}|i=a\ldots b\}$$

각 행 s의 $G_B=(g_a,\cdots,g_b)$에 대응하는 원소들의 m으로 부터의 거리에 대한 평균 절대 거리 ad_s와 평균 절대편차 av_s를 계산한다.

$$ad_s=mean_{t=c\ldots d}\{|e_{st}-m|\}$$
$$av_s=\frac{\sum_{t=c}^{d}|e_{st}-m|}{d-c+1}$$

거리의 값 d를 증가시켜가면서, 해당 거리 구간에 절대 편차값이 대응되는 행의 개수를 계산한다. 원소의 m으로 부터의 최대 거리를 l이라 하고, 구간을 n로 나누는 경우, 평균 표준편차 av_s가 $rv_m\cdot\tau$이내이면서, 각 구간 $\left[i\frac{l}{n},(i+1)\frac{l}{n}\right)$에 ad_s값이 해당하는 행의 개수를 사용하여 히스토그램을 작성한다.

관심영역 $B_{[a,b:c,d]}$에 대조가 되는 영역의 최소 크기를 사용자가 C로 지정한다고 가정하자. 이전 단계에서 구한 히스토그램으로부터 빈도수가 C 이상인 위치들을 선택한다. 선택된 위치들 중에서 가장 오른쪽에 있는 위치에 속하는 행들의 관심영역 $B_{[a,b:c,d]}$에 대응하는 위치의 값들이 대조군으로 선정된다.

관심영역으로 선택되는 행(유전자)의 집합에 대해서 열이 복수개로 클러스터화 된 경우에는 각 클러스터에 대해서 대조적인 유전자 집단을 선정하는 것이 요구된다. 이를 위해서 앞에서 설명한 대조영역 선정하는 방법을 각 클러스터에 대해서 적용하여 히스토그램을 계산한다. 히스토그램을 중첩시켜서 각 위치별로 최소값을 선택하고, 빈도수가 C 이상인 위치들을 선택한다.

선택된 위치별로 해당 위치의 유전자 이름 집합들의 교집합을 계산하여, 교집합의 크기가 C 이상인 것들을 선택한다. 각 선택된 교집합은 관심영역에 대조적인 발현특성을 갖는 유전자 집단이 된다.

4. 결론

마이크로어레이 데이터는 다수의 샘플에 대한 대량의 유전자 발현정보를 포함하고 있기 때문에, 분석자들의 주된 관심은 이들로부터 추가적인 분석을 위한 유용한 실마리를 효과적으로 찾는 것이다. 이 논문에서는 이러한 실마리 정보를 효과적으로 추출하기 위한 방법으로 시각적인 분석과 통계적 처리로 곤란한 대조적인 집단을 추출하는 효과적인 방법을 제안하였다. 마이크로어레이 데이터에서 지역 클러스터의 구조를 추출하는 것은 분석자들의 분석 부담을 줄일 수 있는 효과적인 기능이기 때문에, 향후 여러 마이크로어레이 분석도구에 반영될 수 있을 것으로 기대된다.

참 고 문 헌

[1] 이건명, 이선아, 이승주, 김원재, 김용준, 배석철, "마이크로어레이 데이터의 게놈 수준 분석을 위한 퍼지 패턴 매칭에 의한 유전자 필터링", 한국퍼지및지능시스템학회 2007추계학술대회 논문집, 제17권, 제2호, pp.145-148, 2007.11.

[2] D. Nam, S.-Y. Kim, Gene-Set Approach for Expression Pattern Analysis, Briefing in Bioinformatics, Jan., 2008.

[3] S. Draghici,"Data Analysis Tools for DNA Microarrays," Chapman & Hall/CRC, 2003.

[4] D. W. Mount, "Bioinformatics: Sequence and Genome Analysis," Cold Spring Habor Lab Press, 2004.

[5] G. B. Forgel, D. W. Corne, "Evolutionary Computation in Bioinformatics," Morgan Kaufmann Publishers, 2003

[6] W. L. Martinez, A. R. Martinez, "Exploratory Data Analysis with MATLAB," Chapman&Hall/CRC, 2005.