유전자 발현 분석을 위한 공진화적 바이클러스터링 기법

정제균[O,1,2] 김수진[1,2] 장병탁[1,2,3]
[1]서울대학교 생물정보학 협동과정
[2]서울대학교 바이오정보기술 연구센터
[3]서울대학교 컴퓨터공학부
{jgjoung[O], sjkim, btzhang}@bi.snu.ac.kr

Gene Expression Analysis by Co-evolutionary Biclustering

Je-Gun Joung [O,1,,2] Soo-Jin Kim [1,2] Byoung-Tak Zhang [1,2,3]
[1]Graduate Program in Bioinformatics, Seoul National University
[2]Center for Bioinformation Technology, Seoul National University
[3]School of Computer Science and Engineering, Seoul National University

요 약

마이크로어레이는 전체 유전체 수준의 mRNA 발현 여부에 대한 측정이 가능하다는 점에서 분자생물학의 실험 도구로서 가장 강력한 도구 중에 하나로 부각되어 있다. 현재까지 마이크로어래이의 결과로부터 유사한 발현 패턴을 찾기 위한 여러 가지 바이클러스터링 알고리즘들이 개발되어 왔다. 하지만 대다수의 알고리즘들이 최적의 바이클러스터들을 찾기 보다는 일정 수준의 가능한 바이클러스터의 결과만을 제시하고 있다. 본 논문에서는 다른 개체집단들과 상호 진화하는 공진화적 학습에 의한 진화연산 기법을 통하여 유전자-조건의 매트릭스로 부터 열과 행을 동시에 클러스터링하는 공진화적 바이클러스터링 알고리즘(co-evolutionary biclustering algorithm: CBA)을 제안하고자 한다. CBA는 유전자발현 데이터에서 유전자-조건의 상호의존적인 부성분들로 구성된 최적화 문제에 적합한 계산방식이라고 할 수 있다. 인간 유전자 발현 데이터에 대한 실험 결과, 제시한 알고리즘은 이전의 알고리즘에 비해 발견한 바이클러스터의 패턴 유사도에 있어서 우수한 성능을 보이고 있다.

1. 서 론

마이크로어레이는 전체 유전자 수준에서 mRNA의 발현 패턴을 관찰할 수 있다는 장점을 가지고 있다[1]. 마이크로어레이 데이터는 서로 다른 조건하에서 유전자 발현 수준들로 구성된 매트릭스로 보여질 수 있다. 이러한 데이터를 가지고 가장 먼저 해볼 수 있는 분석들 중의 하나로 클러스터링을 들 수 있다. 클러스터링 분석 중에서 바이클러스터링[2]은 조건의 부집합하에서 같은 패턴을 가진 유전자들을 파악할 수 있고, 이는 생물학적으로 더 적합한 분석이라는 점에서 기존의 클러스터링에 비해 장점을 가지고 있다. 하지만 지금까지 많은 바이클러스터링 기법들이 개발 되었음에도 불구하고, 그들 알고리즘들은 최적의 바이클러스터보다는 가능한 바이클러스터들을 찾아준다는 단점을 가지고 있다[2][3][4].

본 논문에서는 이러한 문제점을 해결하기 위하여 공진화의 개념에 의해서 바이클러스터를 찾는 바이클러스터링 알고리즘(Co-evolutionary Biclustering Algorithm: CBA)을 제안한다. 제안한 알고리즘은 유전자와 조건의 두 개체집단을 가지고 학습을 하게 되는 특징을 가지고 있다. 여기서 공진화적 학습은 유전자와 조건의 두 도메인의 관점에서 목표함수를 최적화하는데 있다. 바이클러스터링 문제는 상호의존적인 부요소(subcomponent)의 진화에 대한 논점을 반영하고 있다[6]. 만약 풀고자 하는 어떤 문제가 서로 독립성이 없는 부요소로 분해될 수 있다면, 각각을 따로 진화되도록 하는 것에는 특별히 지장이 없다. 하지만 우리의 문제에 있어서, 바이클러스터링은 유전자와 조건이라는 상호 의존적인 관계를 형성하고 있다. 그래서 이 문제에 있어서, 두 객체들 중 하나가 변한다면, 이는 다른 부요소들과 연계된 적합도의 탐색지형(landscape)의 변형을 유발한다. 따라서, 결론적으로 진화알고리즘 관점에서 볼 때, 바이클러스터링 문제에 대해서 제안된 알고리즘은 부요소들간의 상호작용을 허락함으로써 결합된 탐색지형에서 탐색을 수월하게 할 수 있다는 장점을 가지게 된다.

또한, 우리의 알고리즘은 자식 세대의 개체집단을 만드는 과정에서 일반적인 진화알고리즘이 아닌 확률적 정보를 활용하여 전역적 탐색을 수행한다. 이러한 방식의 탐색 방법을 분포추정알고리즘 (Estimation of Distribution Algorithm: EDA)라고 하며, 일반적인 진화알고리즘에 비해서 성능이 우수하다고 평가 받고 있다[7].

CBA의 특성과 성능을 평가하기 위하여 우리는 인공데이터와 실제 마이크로어레이 데이터로 테스트해 보았다. 그 결과, MSR (mean squared residue) 스코어에 있어서 이전 알고리즘의 기본 스코어 보다 31.5% 향상을 가져왔다.

2. 마이크로어레이의 바이클러스터링

$G=\{g_1, g_2, \ldots, g_N\}$을 유전자 집합, $C=\{c_1, c_2, \ldots, c_M\}$을 조건 집합이라고 하자. 유전자들과 조건들로 이루어진 데이터 E는 실수값들로 구성된 $N \times M$ 의 매트릭스로 보여질 수 있다. 매트릭스에 있어서 각각의 엔트리 e_{ij} 는 특정 조건 c_j 하에서 특정 유전자 g_i의 발현 레벨을 의미한다.

바이클러스터는 전체 매트릭스에서 행이나 열이 응집성을 보이는 부매트릭스를 말한다. 행 인덱스의 집합을 I이라고 하고, 열 인덱스의 집합을 J라고 할 때, 바이클러스터는 $|I| \le |N|$ 과 $|J| \le |M|$인 (I, J) 이라고 표기할 수 있다. 우리의 경우, 바이클러스터링은 mean squared residue (MSR) score [3]를 최소화하는 바이클러스터를 찾는데 목적이 있다.

$$H_{IJ} = \frac{1}{|I||J|} \sum_{i \subset I, j \subset J} h_{ij}^2 , \quad (1)$$

여기서 I와 J에 의해서 결정되는 바이클서스터에서 각각의 엔트리 e_{ij} 의 residue는

$$h_{ij} = e_{ij} - e_{iJ} - e_{Ij} + e_{IJ}, \quad (2)$$

으로 정의되고, 각 항목은

$$e_{iJ} = \frac{\sum_{j \subset J} e_{ij}}{|J|}, \quad e_{Ij} = \frac{\sum_{i \subset I} e_{ij}}{|I|}, \quad e_{IJ} = \frac{\sum_{i \subset I, j \subset J} e_{ij}}{|I||J|}, \quad (3)$$

으로 정의된다. 여기서 e_{iJ}는 i 행에 있어서 J 엔트리들의 평균값이다. 나머지도 수식도 같은 방식으로 해석될 수 있다.

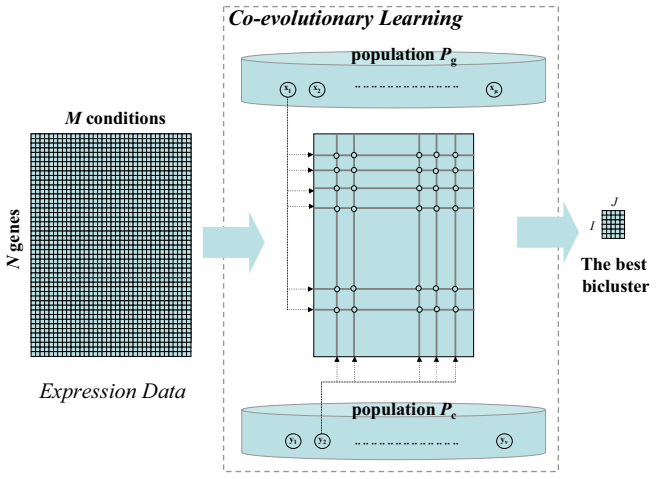

그림. 1. 유전자 발현 분석을 위한 공진화적 바이클러스터링 알고리즘의 개요도

3. 바이클러스터링에 대한 공진화알고리즘

그림 1 은 유전자 분석을 위한 공진화적 바이클러스터링에 대한 개요도를 보이고 있다. 알고리즘의 핵심 아이디어는 유전자 집합과 조건 집합에 대한 두 개의 개체집단이 공진화를 한다는 것이다. 두 개체 집단내의 각 개체는 유전자와 조건 집합에서 g_i 과 c_j 의 인덱스로 이루어진 집합이다.

그리고 적합도 함수는 두 개체집단의 개체들이 결합했을 때 얻어진 MSR 점수를 가지고 계산된다. 그런 다음, 두 개체집단은 이전 세대의 개체집단의 통계적 정보에 의해서 업데이트되게 된다. 세대가 감에 따라서 두 개체집단은 공진화 기법을 통해서 유전자들과 조건들의 확률들이 변하게 되는데, 이러한 확률들은 바이클러스터를 형성하는데 있어서 중요도를 의미한다. 다음은 CBA의 전체적인 알고리즘의 절차를 보이고 있다.

Algorithm 1 Co-evolutionary biclustering

Initialize $Pop_G(0) := \{x_1(0), ..., x_\sigma(0)\}$, $Pop_C(0) := \{y_1(0), ..., y_\tau(0)\}$;
Evaluation $Pop_G(0) : \{A(x_1(0)), ..., A(x_\sigma(0))\}$
$Pop_C(0) : \{A(y_1(0)), ..., A(y_\tau(0))\}$
While ($t > gen_{max}$) **do**
Selection: $Pop'_G(t) := s(Pop_G(t))$,
$Pop'_C(t) := s(Pop_C(t))$;
Probability Update: $P_G(t+1) := u(P_G(t))$,
$P_C(t+1) := u(P_C(t))$;
Generation: $Pop_G(t+1) := \phi(P_G(t+1))$,
$Pop_C(t+1) := \phi(P_C(t+1))$;
Evaluation: $Pop_G(t+1) : \{\#A(x_1(t+1)), ..., A(x_\sigma(t+1))\}$,
$Pop_C(t+1) : \{\#A(y_1(t+1)), ..., A(y_\tau(t+1))\}$;
$t := t+1$;
end.

두 개체집단의 다음 세대 $Pop(t+1)$은 다음 식과 같이 전 세대의 확률 분포에 의해서 샘플링하는 방식으로 만들어진다.

$$Pop_G(t+1) = \phi(p_1, p_2, \ldots, p_N), \quad (4)$$
$$Pop_C(t+1) = \phi(p_1, p_2, \ldots, p_M).$$

그리고 각각의 확률은 다음 식과 같이 적합도가 높은 개체 S_g , S_c개에 유전자 g_i나 조건 c_j가 어느 정도 포함되었는지를 근거로 결정된다.

$$p_i = (1-\zeta)p_i + \zeta \frac{f_{g_i}}{\sum_{k=1}^{S_g} f_{g_k}}, \quad i \subset I, \quad (5)$$

$$p_j = (1-\eta)p_j + \eta \frac{f_{c_j}}{\sum_{k=1}^{S_c} f_{c_k}}, \quad j \subset J,$$

여기서 ζ 와 η 는 확률을 갱신하는데 있어서 조절에 관여하는 파라미터들이고, f 는 g_i나 c_j의 빈도를 나타낸다. 각 개체의 적합도는 식 (6)과 같이 유전자에 대한 개체일 경우, 조건에 대한 개체들 중에서 적합도가 높은 개체들과 결합했을 때의 점수를, 조건의 개체일 경우는 그 반대로 적용하여 계산한다.

$$A(x_i) = \frac{\sum_{k=1}^{Z_g} H^B_{x_i y_k}}{Z_g}, \quad A(y_j) = \frac{\sum_{k=1}^{Z_c} H^B_{x_k y_j}}{Z_c}. \quad (6)$$

여기서 H^B는 높은 유사도를 가진 바이클러스터들의 점수(MSR의 낮은 점수)를 의미한다. 그리고 Z_g는 Z_c 는 낮은 점수를 가진 개체들의 개수를 의미한다. 여기서 Z는 세대가 감에 따라 다음 식에 의해 바뀌게 된다.

$$Z = \begin{cases} Z\nu & \text{if } Z > 1, \\ 1 & \text{otherwise.} \end{cases} \quad (7)$$

여기서 $\nu \in (0,1]$의 값이다. ν의 값이 클수록 선택되는 식 (6)에서 언급된 개체의 수는 빨리 줄어든다.

4. 실험 결과

우리는 공진화적 바이클러스터링 알고리즘을 두 가지 다른 데이터들에 적용해 보았다. 먼저 인공데이터에 대해서는 랜덤으로 200 X 20의 매트릭스를 생성한 다음, 행에 있어서 약간의 잡음을 첨가한 10 X 5의 유사도가 높은 부매트릭스를 그림 2와 같이 전체 매트릭스에 삽입하였다.

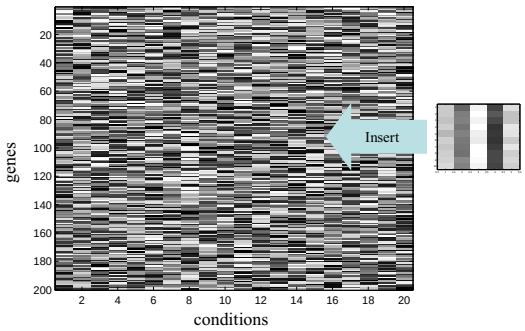

그림. 2. 인공 데이터: 전체 매트릭스에 유사도가 있는 요소들을 포함하는 블록을 삽입시킴.

두 번째 데이터는 실제 마이크로어레이 실험 데이터로서 DLBCL (diffuse large B-cell lymphoma) 종양의 인간 유전자 발현 데이터이다[8]. 이 데이터는 4,026 유전자와 96 조건으로 이루어져 있다. 발현에 있어서 누락된 부분은 랜덤값으로 채워져 있다. 실험 설정은 Table 1에 명시되어 있다.

Table 1. 실험 설정

Parameters	인공 데이터	실제 데이터
Pop_G의 수	800	10000
Pop_C의 수	80	300
세대 수	150	150
Z_g, Z_c	0.3, 0.3	0.1, 0.3
v#(유전자, 조건)	(0.9, 0.95)	(0.9, 0.95)
ζ, η#	0.1, 0.1	0.1, 0.1
w_x, w_y	10, 5	30, 20

그림 3과 같이 우리의 알고리즘은 그림 2에서 명시한 부매트릭스를 정확하게 발견하였다.

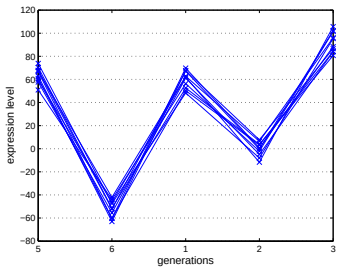

그림. 3. 인공데이터로부터 해답의 탐색 결과

그림 4는 인간 유전자 발현 데이터에 CBA를 적용한 결과를 보이고 있다. 적합도가 가장 좋은 개체를 세대 수 증가에 따라 관찰했을 때, 유전자에 대한 적합도와 조건에 대한 적합도는 좋아지고 있다. 스코어도 마찬가지로 좋아지고 있으며, 특히 기존의 Cheng논문[3]에서 제시하고 있는 기본스코어 (1200: 왼쪽 그림의 점선) 보다 아주 좋은 값을 보이고 있다.

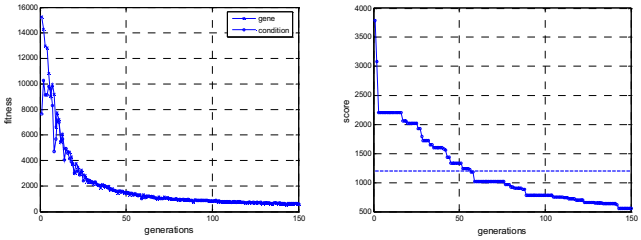

그림. 4. 인간 유전자 발현 데이터: 왼쪽은 세대수 vs. 적합도, 오른쪽은 세대수 vs. 스코어를 보이고 있다.

5. 결론 및 향후과제

본 논문은 공진화적 학습 방법을 통하여 유전자 발현 매트릭스의 열과 행을 동시에 클러스터링하는 바이클러스터링, CBA를 제안하였다. CBA는 진화알고리즘 관점에서 볼 때, 바이클러스터링 문제에서 주어진 문제를 분해하여 해결하기 때문에, 특정 조건들에서 유전자 패턴을 찾아내는 데 있어서 효과적이다. CBA를 두 발현데이터에 적용해본 결과, 기존의 알고리즘 보다 고수준의 바이클러스터를 찾을 수 있다는 것을 확인할 수 있었다.

감사의 글

이 논문은 과학기술부 국가지정연구실 사업(NRL)에 의하여 지원되었음.

참고 문헌

[1] DeRisi,J.L., Iyer.V.R., Brown.P.O., Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale, *Science*, Vol. 278, pp. 680-686, 1997.

[2] Madeira,S.C., Oliveira,A.L., Biclustering Algorithms for Biological Data Analysis: A Survey, *IEEE/ACM Tran. on Comp. Bio. and Bioinfor.*, Vol. 1, pp. 24-45, 2004.

[3] Cheng,Y., Church,G., Biclustering of Expression Data, *In Proceedings ISMB*, pp. 93-103, 2000.

[4] Yang, J., Wang,W., Wang,H., Yu.P., Enhanced Biclustering on Expression Data, *In Proceedings of the 3rd IEEE Conf. on Bioinfor. and Bioeng.*, pp. 321–327, 2003.

[5] Hillis,D.W.: Co-evolving Parasites Improve Simulated Evolution in an Optimization Procedure, *Physica D*, Vol. 42, pp. 228-234, 1990.

[6] Potter, M., De Joung, K., Cooperative Coevolution: An Architecture for Evolving Coadapted Subcomponents, *Evol. Comp.*, Vol. 8, pp. 1-9, 2000.

[7] Larranaga,P., Lozano,J.A., Estimation of Distribution Algorithms, A New Tool for Evolutionary Computation, Kluwer Academic Publishers (2001).

[8] Alizadeh,A.A. et al., Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling, *Nature*, Vol. 403, pp. 503-510, 2000.