진화 알고리즘을 통한 전사 조절 모티프 조합 탐색

이제근[O1,2], 정제균[1,2], 오석준[2], 장병탁[1,2,3]
[1]서울대학교 생물정보학 협동과정
[2]서울대학교 바이오정보기술 연구센터
[3]서울대학교 컴퓨터공학부
{jkrhee, jgjoung, sjo, btzhang}@bi.snu.ac.kr

Search of Transcriptional Motif Combination using Evolutionary Algorithms

Je-Keun Rhee[O1,2], Je-Gun Joung[1,2], S. June O[2], Byoung-Tak Zhang[1,2,3]
[1]Interdisciplinary 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

유전자 발현은 다양한 전사 인자들의 상호 작용에 의해서 조절되어진다. 이러한 전사 인자들에 존재하는 모티프는 직접적으로 조절 작용을 위한 기능을 수행한다. 또한 대부분의 경우에서 여러 모티프가 함께 유전자 발현 기작을 위하여 조절 작용을 한다. 따라서 이러한 모티프들이 어떤 조합으로 함께 전사 과정에 관여하는지 여부를 밝히는 작업은 중요한 일이다. 본 논문에서 진화 연산을 응용하여, 다양한 조건 하에 전사 과정에 중요하게 작용하는 모티프들의 조합을 알아보았고, 그 결과를 기본적인 k-Means 알고리즘 등과 비교하여 제안한 방법이 유전자들의 상관관계에 있어서 보다 우수한 결과를 보임을 알 수 있었다.

1. 서 론

유전자 발현(gene expression)은 매우 복잡하고 다양한 기작에 의해서 조절된다. 이 중 전사(transcription) 과정에서의 조절 기작은 DNA에 결합하는 RNA 중합 효소(RNA polymerase)와 함께 상호작용하는 전사 인자(transcription factor)들, 그리고 이들의 조합에 의해서 주로 발현이 조절된다고 할 수 있다.

유전자 발현의 조절 기작과 전사 인자들 간의 상호 관계에 대해서는 오래 전부터 분자생물학적인 실험들을 통해 조금씩 밝혀져왔다. 하지만 이러한 방법으로는 유전자 발현을 조절하는 여러 가지 요인들이 서로 어떻게 작용하는지 한 번에 종합적으로 밝혀내기 어렵다. 이러한 한계 때문에 최근에는 여러 연구들이 마이크로어레이(microarray) 결과를 분석하여 유전자 발현의 조절 요인들을 종합적으로 밝혀내고자 하고 있다. 한 가지 예로, 유전자 발현 데이터에서 함께 발현되는 유전자들을 서로 묶은 후, 이를 이용하여 함께 작용하는 모티프(motif)들을 찾음으로서 유전자들간의 전체적인 관계를 알아보는 결과가 있었다[1]. 또한 *Saccharomyces cerevisiae*에서 각 모티프들 사이의 발현 관계에 대한 점수를 계산하여, 유전자 발현을 조절하는 모티프 조합을 확인해보려는 노력도 있었다[2].

본 논문에서는 세포 주기를 비롯한 여러 조건하에서 어떤 모티프들이 상호 조합으로 조절 기작에 참여하는지를 알아보기 위하여 진화 알고리즘을 이용하였다. 진화 알고리즘은 병렬탐색을 수행함으로서 최적화에 있어서 좋은 성능을 보이는 것으로 알려져 있다. 우선 전체적인 알고리즘의 개요에 대해 설명하고, 이를 위해 필요한 알고리즘을 제안할 것이다. 그리고 3절에서는 실험 데이터를 이용하여 실질적인 실험을 해 보고, 그 결과를 기존의 결과들과 비교해보았다.

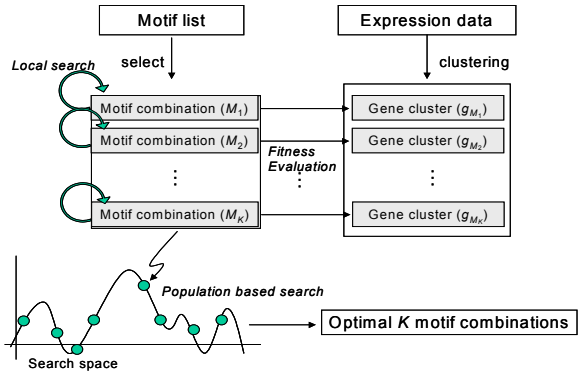

그림 1. 상호간에 조합으로 전사를 조절할 수 있는 모티프 탐색 방법의 전체적인 개요

2. 전사 모티프 조합 탐색을 위한 진화 알고리즘

2.1 발현프로파일을 근거로한 모티프 조합 탐색

우선 유전자 발현에 있어서 모티프 간의 상호 조합의 효과를 파악하기 위하여 마이크로어레이 실험데이터가 필요하다. 만약 상류지역(upstream region)에 같은 모티프 조합을 가지고 있는 유전자들이 있을 때, 그 유전자들의 발현 프로파일들도 서로 유사하다면, 상호 조절할 가능성이 크다고 판단할 수 있다. 이러한 상호 조절에 영향을 주는 모티프 조합을 추출하기 위하여 여기서는 진화 알고리즘을 이용하였다. 각 개체군은 서로 연관된 모티프 집합(set)으로 표현되며, 한 세대가 진행될수록 보다 적합한 모티프 조합이 만들어지게 된다. 결국 최종적으로 가장 최적의 모티프 집합들이 만들어질 수 있는 것이다. 그림 1에서는 진화 알고리즘에 의하여 모티프

조합을 최적화하는 전체적인 개요에 대하여 개략적으로 설명하고 있다.

2.2 알고리즘 설계

그림 2는 여기서 제시하고 있는 기본 알고리즘을 보여준다. 초기의 개체군은 랜덤하게 N개만큼 생성된다. 여기에서 각 개체는 우선 적합도 함수(fitness function)를 근거로 하여 좋은 해(solution)인지를 평가받게 된다. 그런 다음, 개체군(population)에서 랜덤하게 개체들이 선택(selection)되어 유전 연산자(genetic operator)들을 통하여 다음 세대(generation)에 해로서 전달되게 된다. 이러한 선택 과정은 완전한 랜덤이라기 보다는 확률적인 선택 방법을 이용한다. 즉 적합도가 높은 개체군일수록 진화 연산을 위해 선정될 가능성을 높게 만드는 것이다.

```
procedure
    begin
    initialize population Pop of size #Popsize;
    for each individual do Fitness Evaluation; end
    t := 0
    while (not termination condition) do
        for i := 1 to #recombinations do
            select two parents Ia, Ib randomly;
            Ic = Recombine(Ia, Ib);
            add individual Ic to Pop;
        end
        for i := 1 to #mutations do
            Ic := Mutation(Ic);
        end
        for each individual do Fitness Evaluation; end
        for i := 1 to #local search do
            Ic = Local-Search(Ic);
        end
        t := t + 1
    end
end
```

그림 2 재조합, 변이, 지역 탐색을 이용하여 우리의 실험에 사용된 알고리즘

유전 연산자들은 다음 세 가지로 요약할 수 있다. 먼저, 선택된 개체군은 p_c의 확률을 가지고 교차(crossover) 연산이 수행된다. 교차 연산의 결과 두 부모로부터 새로운 두 개의 개체군이 만들어지게 된다. 이 실험에서 교차 연산은 랜덤하게 선택된 교차점을 기준으로 서로 교환되도록 하였다. 교차 연산은 그림 1에서 보이는 모티프 조합 M_i들의 교환을 수행한다.

또한, 변이(mutation) 연산은 선정된 개체군에서 특정한 지점에서 변이가 일어나도록 한다. 하지만 이 연산은 세대가 지남에 따라 항상 일어나는 것이 아니라, 매우 낮은 확률로 발생되도록 하였다. 변이 연산은 모티프 조합 M_i에 속한 모티프 색인(index)들의 개수를 조정한다.

한편 지역탐색(local search)은 모티프 색인들을 변환함에 따라 적합도가 계속 높아지는 방향으로 언덕 오르기 탐색(hill-climbing search)을 수행한다.

2.3 적합도 함수

각 개체에 대한 적합도는 다음과 같이 정의된다.

$$Fitness = \alpha EC + S$$

EC(expression coherence)는 서로 연관성 있는 유전자들이 발현 공간(expression space) 상에서 얼마나 밀집되어 있는지 여부를 나타내주는 수치이며, S(separation)는 각각의 집단(group)이 서로 얼마나 분리되어 있는지 여부를 표현한다.

각 개체에서 k번째 모티프 조합을 M_k라고 하였을 때, g_{M_k}는 M_k를 포함하는 유전자들이고, EC는 다음과 같이 표현된다.

$$EC = \frac{1}{K}\sum_{k=1}^{K} C(g_{M_k})$$

여기에서의 EC값은 다음의 식에 의해서 표현된 상관계수(correlation coefficient)들의 평균값이다.

$$C(g_{M_k}) = \frac{1}{P}\sum_{i=1}^{J_k}\sum_{j=i+1}^{J_k} r(v(g_{M_k}^{i}), v(g_{M_k}^{j}))$$

여기에서 r은 발현 프로파일(profile)에 따른 유전자 쌍(pair)사이의 유사성을 나타내는 것으로 Pearson 상관계수에 의해 측정된다. 또한 P값은 가능한 유전자 쌍에 대한 전체 개수를 나타내며, J_k는 그룹 내의 유전자 총 수를 나타낸다. 여기에서 같은 클러스터(cluster)에 속하는 유전자의 수는 경계값(threshold) T보다 더 큰 값을 가져야만 한다. 만일 경계값보다 더 작다면 EC는 '0'의 값을 가지도록 한다.

한편 S값은 다음과 같이 표현된다.

$$S = \frac{1}{Z}\sum_{i=1}^{K}\sum_{j=i+1}^{K} d(v(\widehat{g_{M_i}}), v(\widehat{g_{M_j}}))$$

이 식에서 $v(\widehat{g_{M_i}})$은 모티프 조합 M_i를 포함한 유전자 발현 특성의 평균이며, d(distance)는 $1-r$의 값이고, Z는 가능한 클러스터의 총 수이다.

3. 실험 설정 및 결과

실험을 위하여 Pilpel이 추출했던 효모(yeast)의 모티프 정보들을 이용하였다[2]. 이 자료들은 이미 알려진 37개의 모티프와 모티프로 추정되는 329가지의 자료로 구성되며, AlignACE 프로그램을 사용하였다.

실험 결과를 확인하기 위해서는 효모의 세포 주기 동안의 800개의 ORF의 마이크로어레이 분석 결과를 사용하였다[3]. 또한 모티프들이 서로 간에 공동 상승효과를 나타내는지 여부를 확인하기 위하여 6000개 이상의 ORF에 대하여, 포자 형성(sporulation), 열에 의한 충격(heat-shock), 혐기적 성질에서 호기적 성질로의 일시적인 변환(diauxic shift) 등의 환경에서의 유전자 발현 데이터를 이용하였다[4].

실험에서 개체군은 100개를 기본으로 하고, 총 200세대까지 세대수가 진행되도록 하였다. 또한 교차가 일어날 확률(p_c)은 0.9, 변이가 일어날 확률(p_m)은 0.01로 하였다. 지역 탐색에 대해서는 20번 반복하는 동안 수행되도록 하였다. 적합도를 결정하는 데에 있어서, α값은 휴리스틱(heuristic) 값으로서 0.8로 결정하였다. 클러스터

K의 값은 5이며, 유전자 수에 대한 경계값 T는 10으로 설정하여 실험하였다.

먼저 실험을 통하여 유전자들 간의 관계를 확인해보고, 이를 k-Means 알고리즘과 비교하여 보았다. 그림 3은 이 실험에 대한 결과를 보여주고 있다.

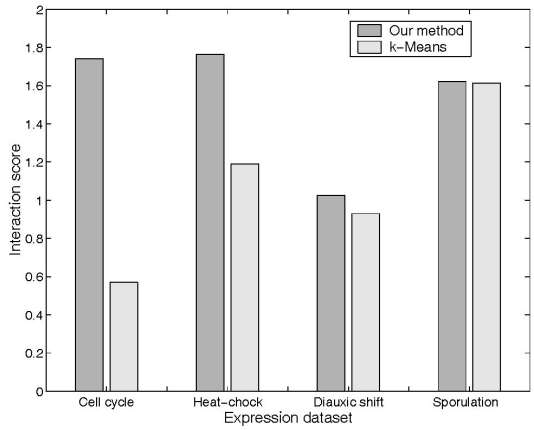

그림 3. 단백질 상호 작용에 대한 우리의 실험 결과와 k-Means 알고리즘의 비교

이 결과는 같은 집단에 속하는 유전자들의 상호 관계를 파악하기 위하여 단백질 상호작용 점수($PI\ score$)를 측정해 본 것이다. 효모에 대한 단백질 상호작용 정보는 MIPS 데이터베이스로부터 얻었으며, 이 점수는 $PID(M)$과 $PID(R)$간의 차이를 이용하여 계산된다.

$$PI\ score = \log(PID(M)/PID(R))$$

$PID(M)$은 모티프 조합을 포함하는 유전자들에서 단백질 상호작용 밀도(protein interaction density)를 나타내는 것이며, $PID(R)$은 랜덤하게 선택된 집합에서의 단백질 상호작용 밀도를 의미한다. k-Means 클러스터링에서 클러스터 k의 수는 30으로 하였다. 이 실험의 결과, k-Means 알고리즘보다 본 논문에서 제시한 알고리즘에 의해 얻어진 클러스터 내의 유전자들에 해당하는 단백질 사이에, 더 많은 상호작용이 존재함을 볼 수 있었다. 따라서 진화알고리즘에 의해 탐색된 모티프 조합들은 전사 조절에 있어서 상호 작용에 의한 영향을 줄 가능성이 상당히 높다는 것을 말해주고 있다.

마지막으로 우리는 네 가지의 다른 발현 환경에서 모티프 조합으로 인한 유전자 발현의 상호 상승효과를 확인해보았다. 그림 4는 우리의 알고리즘을 이용하여 얻어진 모티프 조합의 결과를 보여주며, 특정한 생물학적 조건들과 중요하게 관련되어 있음을 알 수 있다.

예를 들어 세포 주기 자료로부터 얻어진 결과를 보면, MCB와 SCB의 두 모티프가 전사 과정에서 함께 작용하여 유전자 발현의 상승작용을 일으킨다는 사실을 알 수 있다. MCB와 SCB에는 각각 MBF (MCB-binding factor)와 SBF (SCB-binding factor)라는 전사 인자가 결합할 수 있다. 이 MBF와 SBF는 세포 주기 상의 G1 기에서 S기 넘어가는 과정에 사용되는 유전자의 발현을 조절하는 역할을 수행하는 것으로 알려져 있다[5]. 결국 우리의 실험 결과로서 얻어진 모티프들의 작용이 유전자 발현의 조절에 중요한 역할을 수행하는 사실을 확인할 수 있다.

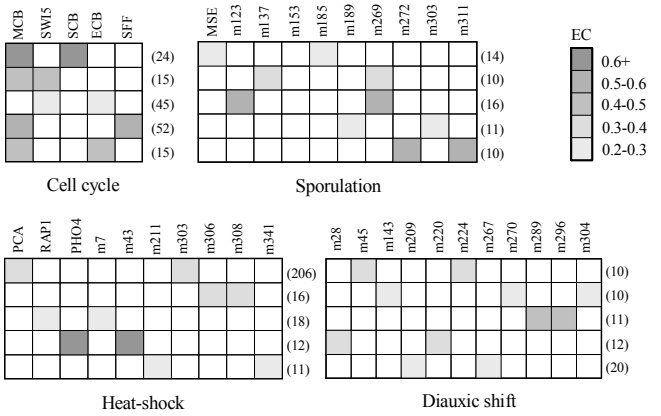

그림 4. 여러 다른 조건에서의 모티프 상호 조합. 회색으로 칠해진 부분은 상승효과를 가지는 모티프들을 가리키며, ( ) 는 이 모티프 조합을 포함하는 유전자의 개수를 나타낸다.

4. 결 론

본 논문에서는 진화 연산을 이용하여 유전자 발현 과정에 영향을 미치는 전사 인자의 모티프들이 어떤 조합으로 작용하는지 여부를 알아보았다. 진화 알고리즘을 이용하여 각 유전자들을 높은 상관관계가 있는 집단으로 군집화함으로서 함께 작용하는 모티프 조합들을 쉽게 분리할 수 있었다.

향후 연구로는 모티프 조합에 대한 최적화 방법에 있어서 우리가 제안한 알고리즘을 보다 개선하고, 생물학적으로 더욱 신뢰할 수 있는 결과를 보일 수 있도록 해야할 것이다.

감사의 글

이 논문은 과학기술부의 국가지정연구실 사업(NRL)과 Systems Biology 사업(M10309000002-03B5000-00110)에 의하여 지원되었음.

참고 문헌

[1] S.Tavazoie, J.D.Hughes, M.J.Cambell, R.J.Cho, and G.M.Church. Systematic determination of genetic network architecture. *Nature genetics*. 22: 281-285, 1999.

[2] Y.Pilpel, P.Sudarsanam, and G.M.Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*. 29: 1-7, 2001.

[3] P.T.Spellman, G.Sherlock, M.Q.Zhang, V.R.Iyer, K.Anders, M.B.Eisen, P.O.Brown, D.Botstein, and B.Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell* 9: 3273-3297, 1998.

[4] M.B.Eisen, P.T.Spellman, P.O.Brown, and D.Botstein. Cluster analysis and display of genome- wide expression patterns. *Proc. Natl. Acad. Sci., USA* 95:14863-14868, 1998.

[5] I.Simon, J.Barnett, N.Hannet, C.T.Harbison, N.J.Rinadi, T.L.Volkert, J.J.Wyrick, J.Zeitlinger, D.K.Gifford, T.S.Jaakkola, and R.A.Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697-708, 2001.