유전자 조절 네트웍 구축을 위한 진화알고리즘 기법

정제균[O1,2] 오석준[2] 남진우[1,2] 장병탁[1,2,3]
[1]서울대학교 생물정보학 협동과정
[2]서울대학교 바이오정보기술 연구센터
[3]서울대학교 컴퓨터공학부
{jgjoung[O], sjaugh, jwnam, btzhang}@bi.snu.ac.kr

Evolutionary Algorithm to Construct Regulatory Genetic Network

Je-Gun Joung[1,2] Sirk June Augh[2] Jin-Wu Nam[1,2] 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 네트웍 구축은 다양한 생물학적 실험 결과를 통하여 유전자 간의 관계를 모델링하는 작업이다. 현재 유전자 섭동(perturbation) 실험은 대규모 유전자 조절 네트웍(regulatory genetic network) 구축을 위한 중요한 데이터로 인식되고 있다. 하지만 유전자 섭동 실험에 의한 결과는 하나의 유전자가 다른 유전자에 대하여 직접적 또는 간접적인 영향을 주는 지에 대한 정보를 파악하기 어렵다. 본 논문은 이러한 문제점을 해결하기 위하여 섭동 실험에 의한 결과로부터 생성된 복잡한 유전자 관계를 실제 생물학적 네트웍 형태로 단순화시키는 진화알고리즘을 제안하고자 한다. 실험은 진화 알고리즘이 임의의 복잡한 네트웍에 대하여 다양한 후보 네트웍 해를 제시해 줄 수 있는 결과를 보여 주고 있다.

1. 서 론

현재 유전자 네트웍을 구축하기 위하여 가장 많이 이용되는 기술은 유전자의 발현(Expression)을 근거로 하는 실험들이다. 유전자칩(DNA chip)의 발달로 유전자에 대한 데이터 분석이 용이해짐에 따라 클러스터링 기법등의 다양한 분석 방법이 개발되었다[1]. 하지만 유전자 칩에 의한 클러스터링은 하나의 유전자가 다른 유전자에 영향을 주는 지에 대한 정보를 유추하기 어렵다. 따라서 최근에 유전자 섭동 실험 결과에 의한 유전자 네트웍 구축의 연구가 활발하게 진행되고 있다[2].

유전자 섭동은 특정 유전자를 변이(mutation), 과도 발현(over expression), 번역의 억제 (inhibition of translation)과 같은 방법을 사용하여 유전자의 활동성(activity)를 측정하는 기술이다. 이러한 유전자 섭동 실험에 의한 결과는 유전자칩에 의한 결과보다 정확하게 유전자 네트웍을 구축할 수 있다는 장점을 가지고 있다. 그러나 유전자 섭동 실험에 의한 결과는 하나의 유전자가 다른 유전자에 대하여 직접적 또는 간접적인 영향을 주는지에 대한 정보를 파악하기 힘들다.

본 논문은 이러한 문제점을 해결하기 위하여 섭동 실험에 의한 결과로부터 생성된 복잡한 유전자 관계를 실제 생물학적 네트웍 형태로 단순화시키는 진화 알고리즘을 제안하고자 한다.

2. 유전자 조절 네트웍의 개념

하나의 세포에는 무수히 많은 유전자가 존재한다. 이렇게 많은 유전자는 각각의 고유 기능을 가지고 있으며, 다른 유전자들과의 직접적이거나 간접적인 단순한 관계를 가지고 복잡한 전체 유전자의 조절 시스템을 만들어 간다. 이러한 유전자 조절 네트웍은 각각의 독립적인 유전자들이 다른 유전자의 활동에 영향을 미칠 수 있는가에 대한 관계를 그래프로 표현한 것이다. 유전자 조절 네트웍에서 전사 조절 인자(transcription fa ctor)의 유전자 발현 조절, 단백질 활성 조절 등은 직접적인 관계를 나타내며, 전사 조절인자에 의해 발현된 유전자의 단백질이 또 다른 유전자의 발현을 조절하거나, 활성 조절 인자에 의해 활성화된 전사 조절인자가 다른 유전자 발현을 조절하는 관계등은 간접적인 관계로 말할 수 있다.

그림1(a)은 간단한 유전자 조절 네트웍의 개념도를 그린 것이다. 유전자1은 유전자2의 발현을 조절하는 전사 조절 단백질이기 때문에 유전자1 과 유전자2는 직접적인 관계가 있다. 또한 유전자 2는 인산화 단백질(protein kinase)을 발현시켜 유전자3의 단백질을 활성화시킴으로써 유전자3과 직접적인 관계를 갖는다. 유전자 3과 4도 유전자 조절 단백질의 조절 관계로 직접적인 관계를 갖는다. 이렇게 직접적인 관계를 간략히 그래프로 표시하면 그림1(b)와 같은 네트웍을 얻게 된다.

3. 유전자조절 네트웍 구축을 위한 진화알고리즘

진화 알고리즘은 자연 세계의 진화 과정을 모방해 문제 풀이 또는 모의 실험에 이용하는 연구의 한 방법이다. 진화 알고리즘은 풀고자 하는 문제에 대한 해들을 염색체로 표현한 다음 이들을 점차적으로 변형함으로써 점점 더 좋은 해들을 생성해 낸다. 진화 알고리즘 중에서 유전 알고리즘(genetic algorithm)

은 염색체가 고정 길이의 이진 문자열(binary string)로 표현된다[3].

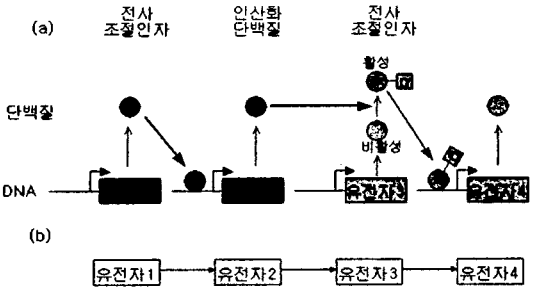

그림 1. (a)간단한 유전자 조절 네트웍의 개념도 (b)직접적인 유전자 조절 네트웍

각각의 해는 개체(individual)라고 부르고, 이 해들의 집합을 개체군(population)이라고 부른다. 세대가 지날수록 점점 좋은 해를 생성하기 위해 유전 알고리즘에서는 교차(crossover)와 돌연변이(mutation)를 연산자로 사용하게 된다.

```
개체군 크기 Npop 선택
진화 횟수 Ng 선택
개체군 초기화
최대 진화 횟수 Ng일 때 까지 반복:
  1. 적합도(fitness)에 근거하여 확률적으로
     두 부모를 선택
  2. 교차 연산에 의해서 두 부모로부터 자식을 생성
  3. 돌연변이 연산에 의해서 무작위로 값을 변경
  4. 개체군 크기가 될 때까지 1번부터 반복
```

그림 2. 유전 알고리즘의 흐름

유전 알고리즘은 방대한 문제 공간에 대하여 하나의 개체가 아닌 개체군 단위의 병렬적인 탐색을 수행한다는 특징을 가지고 있다. 또한 교차 연산과 돌연변이를 통해서 다른 알고리즘에 비해 전역탐색을 쉽게 할 수 있는 장점이 있다. 그래서 TSP(traveling salesman problem)와 같은 다양한 그래프 최적화 문제에 적용되어 좋은 성능을 보이고 있다[4].

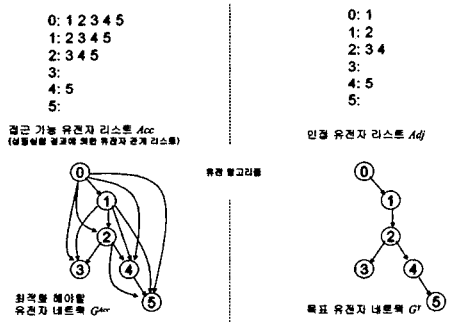

그림 3. 유전 알고리즘에 의한 유전자 조절 네트웍 구축의 개념도

그림 1은 유전 알고리즘에 의한 유전자조절 네트웍 구축에 대한 개념을 보여 주고 있다. 유전 알고리즘으로 학습해야 할 최종 목표는 주어진 생물학 실험에 의한 데이터로부터 네트웍의 복잡도가 최적화된 유전자 네트웍을 구축하는 것이다. 생물학적 실험에 의해 생성된 데이터는 각각의 유전자에 대하여 연결선으로 접근가능한 유전자 리스트 *Acc*인데, 우리는 이러한 리스트로부터 유전 알고리즘을 통하여 인접 유전자 리스트를 생성하고자 한다.

이를 위하여 그림 4와 같이 유전 알고리즘의 각 개체를 암호화(encoding)할 수 있다. 개체의 각 비트(bit)는 인접 노드로서 서로 연결이 되는지를 나타낸 것이다.

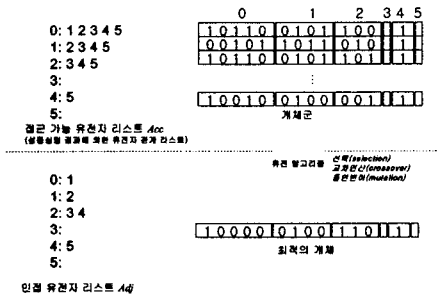

그림 4. 유전자 조절 네트웍 구축을 위한 유전 알고리즘의 유전자 암호화 방식

각각의 i 개체에 대한 적합도 함수는 식(1)과 같이 두 개의 항목으로 구성되어 있다. 첫 번째 항목은 각각의 노드가 접근 가능한 다른 노드들에 얼마나 접근 가능한 지를 측정한 것이다. 여기서는 식 (2)와 같이 T_{Acc}를 노드 접근도라고 정의한다. 그리고 두 번째 항목은 네트웍이 얼마나 적당한 연결선으로 구성되어 있는 지를 측정한 것이다. 식 (3)의 T_{Edge}는 연결선의 복잡도라고 정의한다.

$$Fitness(i) = \alpha T_{Acc}(i) + \beta T_{Edge}(i) \quad (1)$$

$$T_{Acc}(i) = N_{Acc}(i) / N_{Acc}^{Total} \quad (2)$$

$$T_{Edge}(i) = (1 - (|N_{Node}^{Total} - N_{Edge}(i) - 1|) / N_{Node}^{Total}) \quad (3)$$

α와 β는 두 항목에 대하여 균형을 맞추기 위한 상수들이다. 여기서 α=1-β이다. α가 크면 접근도에 중점을 두게 되고 반면에 β가 크면 연결선의 복잡도에 중점을 두게된다. α와 β값은 반복 실험을 통하여 최적화해야 한다.

4. 실험 및 결과

유전자 조절 네트웍 구축을 위한 실험 데이터는 20개의 유전자에 대하여 섭동 실험에 대한 가상적인 실험의 결과로 생각되는 접근 가능 유전자 리스트 *Acc*를 무작위 샘플링한 것이다. 실험 데이터는 먼저 그림 6의 (a)에 보이는 인접 노드에 의한 그래프를 해로서 생각하고, 이를 바탕으로 그림 5과 같이 접근 가능 유전자 리스트에 의한 복잡한 그래프를 생성한다. 즉, 그림 5의 데이터를 통하여 그림 6의 (a)에 있는 단순한 그래프를 유전 알고리즘을 통한 학습 목표로 설정한다.

개체군의 크기는 200, 세대수는 100, 돌연변이 확률은 0.01로 설정하였고, 노드 접근도에 대한 상수 α와 연결선 복잡도에 대한 상수 β는 0.7, 0.5, 0.3로 다양하게 변화시켜 보았다.

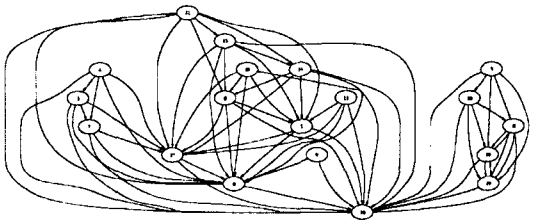

그림 5. 섭동 실험 결과로 생성된 전체 네트웍의 가상 데이터

유전 알고리즘을 통하여 최종 세대에서 결과가 나온 최적해를 그림 6의 (b)에서 보여 주고 있다. 그림 6의 (a)와 (b)의 네트웍 구조가 유사함을 알 수 있다. 사실 생물학적으로 그림 5의 섭동 실험 결과에 의한 네트웍에 대하여 그림 6의 (a)와 같은 네트웍이 정답이라고 단정지을 수는 없다. 따라서 유전 알고리즘은 (b)와 같은 여러 가지 후보 네트웍을 제안해 줄 수 있다는 장점이 있다. 또한 유전자들 간의 직접적인 연결에 대한 사전 지식이 쌓임에 때라 이를 적합도 함수에 반영할 수도 있다.

(a) (b)

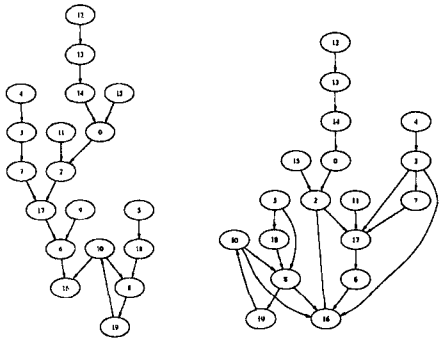

그림 6. (a) 목표 유전자 네트웍, (b) 유전 알고리즘 결과에 의한 유전자 네트웍

그림 7은 유전 알고리즘의 세대에 따른 적합도의 추이를 보이고 있다. 각각의 선은 20번 수행한 결과를 평균한 것이다. 이에 대한 실험 결과는 노드 접근도 상수를 변경하면서 최고 적합도의 변화를 살펴 본 것인데 모든 적합도가 동일하게 세대가 감에 따라 좋아지는 것을 알 수 있었다. 이 그림에서 노드 접근도가 0.3일 때, 즉 연결선 복잡도가 0.7일 때 가장 좋은 접합도를 나타내고 있다.

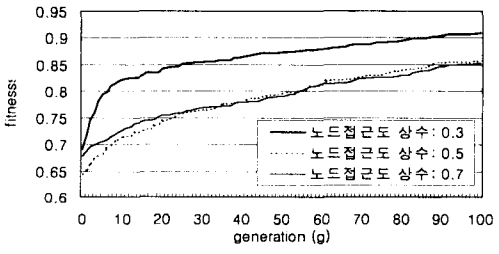

그림 7. 유전 알고리즘의 세대에 따른 적합도 결과

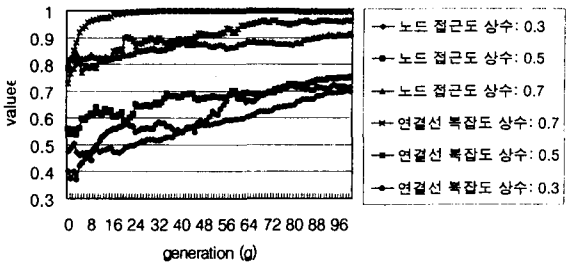

그림 8. 유전 알고리즘의 세대에 따른 노드 접근도와 연결선 복잡도의 결과

그림 8은 유전 알고리즘의 세대에 따른 노드 접근도와 연결선 복잡도의 관계에 대한 결과를 보이고 있다. 이 실험 결과는 각각의 상수값을 변경함에 따라 적합도에 영향을 주고 있음을 보여주고 있다. 예를 들어, 노드 접근도 상수 α가 0.7, 일 때 연결선 복잡도에 대한 상수 β는 0.3을 가지게 된다. 이 때, 노드 접근도는 높은 값을 가지는 반면, 연결선 복잡도는 상대적으로 낮은 값을 보이고 있다.

5. 결 론

본 논문에서는 섭동에 의한 생물학 실험 결과를 통한 유전자 조절 네트웍 구축을 위한 유전 알고리즘의 학습 방법에 대해서 살펴보았다. 실험 결과를 통하여 우리는 유전 알고리즘이 해로써 가능한 유전자 조절 네트웍을 제시해 줄 수 있다는 결론을 얻게 되었다.

향후 연구로서 인공 데이터가 아닌 효모(yeast)와 같은 종에 대한 섭동 실험 결과에서 나온 실제 생물학 데이터를 적용할 계획이다. 그러기 위해서는 우리는 먼저 데이터에 대한 잡음(noise)에 민감한지를 테스트해야 할 것이다. 그리고 대규모 데이터를 고려해 볼 때 노드의 개수도 고려해야 할 것이다.

감사의 글

이 논문은 과학기술부의 국가지정연구실 사업과 IMT-2000 과제에 의하여 지원되었음.

참고문헌

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

[2] Andreas Wagner, How to reconstruct a large genetic network from n gene perturbations in fewer *n*2 easy steps., *Bioinformatics*, Vol. 17, pp. 1183-1197, 2001.

[3] Goldberg, D. E., Genetic algotithms in search, *Optimization, and Machine Learning.*, Addison -Wesley, 1989.

[4] Homaifar, A., Guan, S., and Liepins, G. E. , A new approach on the traveling salesman problem by genetic algrithoms., *Procceedings of the Third International Conference for Genetic Algorithms*, pp. 460-466, 1998.