유전자알고리즘을 이용한 유전자 조절네트워크 추론

Gene Regulatory Network Inference using Genetic Algorithms

김태건[1], 정성훈[2]

[1] 서울시 성북구 한성대학교 정보통신공학과
E-mail: like.sunshine@gmail.com
[2] 서울시 성북구 한성대학교 정보통신공학과
E-mail: shjung@hansung.ac.kr

요 약

본 논문에서는 유전자 발현데이터로부터 유전자 조절네트워크를 추론하는 유전자 알고리즘을 제안한다. 근래에 유전자 알고리즘을 이용하여 유전자 조절네트워크를 추론하려는 시도가 있었으나 그리 성공적이지 못하였다. 우리는 본 논문에서 유전자 조절네트워크를 보다 효율적으로 추론할 수 있게 하기위하여 새로운 유전자 인코딩 기법을 개발하여 적용하였다. 선형 유전자 조절네트워크로 모델링 된 인공 유전자 조절네트워크를 사용하여 실험한 결과 대부분의 경우에 있어서 주어진 인공 유전자 조절네트워크와 유사한 네트워크를 추론하였으며 완전히 동일한 유전자네트워크를 추론하기도 하였다. 향후 실제 유전자 발현데이터를 이용하여 추론해 보는 것이 필요하다.

Key Words : Genetic Algorithms, Gene Regulatory Network Inference

1. 서 론

분자생물학 및 공학기술의 결합으로 마이크로어레이(Microarray) 실험기법이 등장함에 따라 수천수만의 유전자를 동시다발적으로 실험하여 대규모 유전자발현 데이터(Gene Expression Data)를 얻을 수 있게 되었다 [1]. 이에 따라 유전자발현데이터로부터 유전자 상호 간의 작용메커니즘을 밝히는 작업의 필요성이 대두되었다. 이러한 필요성에 유전자 조절네트워크를 추론하는 여러 가지 방법이 개발되었는데 그 중 한 방법이 유전자 알고리즘을 이용하는 방법이다 [2,3,4,6].

유전자 알고리즘을 이용한 추론 방법에서는 유전자 조절네트워크를 인코딩하여 개체로 나타내고 이 개체를 진화시켜 추론을 하게 된다. 이 때 개체별 적합도 평가는 인코딩된 개체를 디코딩하여 후보 유전자 조절네트워크를 만들고 이 유전자 조절네트워크가 만드는 유전자 발현 데이터와 실험으로 얻은 유전자 발현 데이터가 얼마나 유사한지를 척도로 평가하게 된다. 즉 후보 유전자 조절네트워크가 만드는 유전자 발현데이터가 실험으로 얻은 유전자 발현데이터와 유사할수록 큰 적합도를 갖게 된다. 그러나 실험기법상의 제약으로 유전자발현 데이터가 유전자 수에 비해 극히 적기 때문에 만족스러운 정도의 결과를 얻지 못하고 있다 [2]. 또한 유전자 조절네트워크의 특성이 반영되지 않은 인코딩 방법으로 인하여 효과적인 탐색이 이루어지지 않는 문제점이 있다.

본 연구에서는 유전자 조절네트워크의 특성을 반영한 인코딩 방법을 사용하여 유전자 조절 네트워크를 추론하는 방법을 제안한다. 본 논문에서 사용한 인코딩 방법은 인공유전체(Artificial Genome) [7]를 응용한 인코딩 방법을 사용한 것으로서 이렇게 하면 실제 유전자 네트워크와 같이 유전자 간의 관계가 0 인 개체로 많이 진화되게 되고 유전자 간의 상호영향 정도가 급격하게 변하지 않는 등의 장점이 있다.

우리는 본 논문에서 제안한 방법의 효용성을 실험하기 위하여 선형 유전자 조절네트워크로 모델링된 인공 유전자 조절네트워크를 이용하여 실험하였다. 즉 인공의 선형 유전자 조절네트워크로 인공 유전자 발현데이터를 만든 다음 이것을 이용하여 유전자 알고리즘으로 유전자 네트워크를 추론하였다. 실험결과 대부분의 경우에 주어진 인공 유전자 조절네트워크와 유사한 것을 찾았으며 똑같은 것을 찾는 경우도 많이 있었다. 이런 결과로 볼 때 본 논문에서 제

안한 방법이 효과적임을 알 수 있었다. 향후 실제 유전자 발현데이터를 이용하여 추론해 보는 연구가 필요하다.

2. 유전자 조절네트워크 추론

유전자 조절네트워크는 각 유전자간의 상호관계로 표현된다. 그림 1은 4개의 유전자를 갖는 유전자 조절네트워크의 한 예이다. 유전자간의 상호관계는 발현을 촉진시키는 관계(Activation)와 억제시키는 관계(Inhibition)로 표현된다 (그림에서 화살표는 촉진을 의미하며 막대(bar)는 억제를 표현하고 링크의 숫자는 가중치를 의미한다).

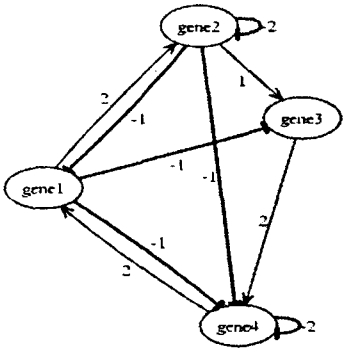

그림 1. 유전자조절네트워크

유전자 알고리즘을 이용하여 유전자 조절네트워크를 추론하기 위하여 유전자 조절네트워크가 유전자 알고리즘의 개체로 인코딩되어야 한다. 본 논문에서 제안하는 인코딩 방법은 인공유전체(Artificial Genome)[7]를 응용한 방법으로 그림 2와 같다.

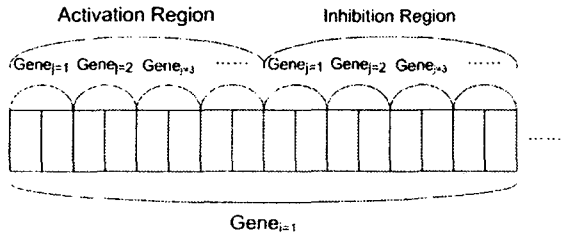

그림 2. 유전자 조절네트워크의 인코딩

그림 2에서처럼 유전자는 자신을 포함하여 다른 유전자를 촉진하는지 억제하는지를 2개의 비트로 표현한다. 2개의 비트의 값은 1의 개수로 주어지기 때문에 2비트가 가질 수 있는 값은 0, 1, 2이다 (01과 10은 1의 개수가 모두 하나로서 같은 값 1을 갖는데 이는 DNA의 코돈(codon)과 같은 일종의 생물체의 여분(Redundancy)기능으로 볼 수 있다). 다른 유전자에 대하여 촉진과 억제를 모두 인코딩하기 때문에 실제 가중치는 촉진 가중치에서 억제가중치를 뺀 값이 된다. 촉진 가중치에서 억제가중치를 빼면 발생할 수 있는 경우의 수는 0이 세 개, 1과 -1이 각각 두 개, 2와 -2가 각각 하나씩 발생한다(결국 양수는 촉진, 음수는 억제를 나타낸다). 이러한 가중치는 실제 유전자 조절네트워크가 성긴(sparse) 특성을 갖기 때문에 매우 유리한 인코딩 방법이다. 또한 가중치의 절대값이 큰 것 (2 또는 -2)보다는 작은 것(1 또는 -1)이 더 많이 생기는 것도 실제 네트워크의 특성을 반영할 수 있다는 점에서 유리하다.

유전자 조절네트워크를 추론하는 유전자 알고리즘은 위에서 설명한 방식으로 인코딩된 개체를 무작위로 생성하여 진화시킴으로서 주어진 유전자 발현데이터에 가장 유사한 유전자 발현 데이터를 만드는 유전자 조절네트워크를 찾게 된다. 그렇기 때문에 어떤 개체의 적합도는 해당 개체를 디코딩하여 유전자 조절네트워크를 만들고, 만들어진 유전자 조절네트워크를 이용하여 유전자 발현데이터를 만든 후 이 유전자 발현데이터가 주어진 유전자 발현데이터와 얼마나 유사한지를 측정하여 구한다. 그렇지만 실제 응용에 있어서는 주어진 유전자 발현데이터가 실제 찾으려는 유전자 수보다 상당히 적은 문제점이 있다. 이렇게 되면 상당히 많은 후보 유전자 조절네트워크가 주어진 유전자 발현데이터와 유사한 데이터를 만들 수 있어 만족할 만한 결과를 얻기 힘들게 된다.

본 논문에서 제안한 유전자 알고리즘의 성능분석을 위해 식 1에 주어진 것과 같은 인공 선형 유전자 조절네트워크를 이용하여 유전자발현데이터를 인공적으로 생성하였다 [5]. 인공유전자 발현데이터를 만들기 위해 사용한 W 값은 표 1과 같으며 인공 유전자발현 데이터는 표2에 그림은 그림 3에 있다.

$$\frac{dx_i}{dt} = \sum_{j}^{N} W x_{ij} \qquad (1)$$

표 1. 인공유전자조절네트워크의 **W**

j \ i			
0	-1	0	2
2	-2	0	0
-1	1	0	0
-1	-1	2	-2

표 2. 인공 유전자 발현 데이터

time	gene1	gene2	gene3	gene4
0.0	1.00	1.00	1.00	1.00
0.5	0.33	0.35	1.94	0.80
1.0	-0.53	0.78	2.44	0.16
1.5	-0.60	1.31	2.31	-0.35
2.0	-0.09	1.36	1.93	-0.33
2.5	0.32	1.06	1.75	-0.01
3.0	0.28	0.81	1.87	0.20
3.5	0.00	0.83	2.07	0.15
4.0	-0.18	1.00	2.13	-0.02
4.5	-0.12	1.11	2.05	-0.11
5.0	0.03	1.08	1.95	-0.06
5.5	0.10	0.99	1.94	0.03
6.0	0.05	0.94	1.98	0.06
6.5	-0.03	0.97	2.03	0.02
7.0	-0.05	1.01	2.03	-0.02

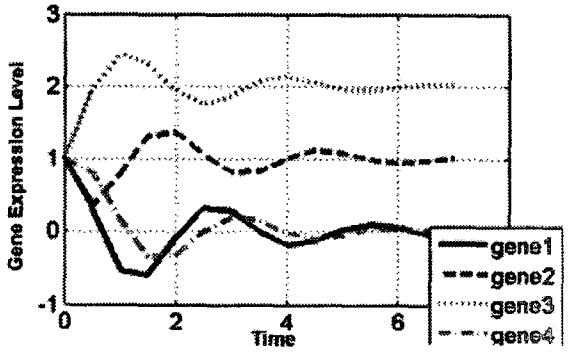

그림 3. 인공 유전자 발현 데이터

유전자알고리즘의 교배연산(crossover)은 0.45의 확률에 따라 한 개의 지점을 기준으로 두 개의 개체의 한쪽 부분을 서로 교환하였고, 돌연변이연산(mutation)은 0.08의 확률로 각 개체 중 한 값을 0 또는 1 의 값으로 설정하였다. 표 3은 유전자 알고리즘의 파라미터를 표시한다.

표 3. 유전자알고리즘의 파라미터

교배확률	0.45
돌연변이확률	0.08
개체수	300
종료조건	적합도 > 0.95
유전자 수	4

유전자알고리즘의 파라미터 설정은 유전자알고리즘의 성능을 좌우할 수 있다. 유전자알고리즘에서 지역해(local optima)에 빠지는 문제는 돌연변이 연산자를 통해 벗어날 수 있으나 돌연변이 확률이 낮을 경우 지역해에서 벗어나지 못하는 경우가 있다. 반대로 돌연변이 확률이 높으면 유전자 알고리즘이 개발(exploitation)을 하지 못하고 탐험(exploration)만을 반복하여 전역해(global optima)에 접근하지 못 할 수가 있다. 이러한 점 때문에 본 연구에서는 돌연변이 확률을 유동적으로 조절하였다. 각 개체의 적합도 값은 개체에서 W matrix를 구한 후 상미분방정식을 풀어 유전자 발현데이터를 생성하고 주어진 인공유전체발현 데이터와의 유사도로 측정한다. 유사도는 수식 2에서처럼 TSSE(Total Sum Square Error)를 기준으로 작은 값일수록 유사도가 높은 것으로 판단하고 높은 적합도 값을 주었다. 유전자알고리즘의 부모 세대에서 가장 높은 적합도의 개체는 자식 세대에서도 보존되도록 Elitism을 적용하였다.

$$fitness = \frac{1}{1+tsse} \qquad (2)$$

3. 실험결과 및 결과 고찰

서로 다른 난수를 이용하여 88번을 실험한 결과 전체 해 공간의 0.001% 의 부분을 검색한 후 주어진 인공유전자조절네트워크와 같은 유전자발현데이터 양상을 보이는 동일한 유전자 조절네트워크를 추론하였다. 돌연변이 확률은 개체 내에서 최대 적합도가 400세대 이상 변동이 없으면 최대 적합도에 변동이 있을 때까지 50세대 간격으로 0.02씩 돌연변이 확률을 높였다. 돌연변이확률이 0.4이상이 되면 0.999의 확률로 돌연변이 확률을 변동하여 local optima에 빠진 경우 탐험(exploration)을 하도록 유도하였다. 돌연변이 확률을 고정적으로 적용한 경우보다 돌연변이 확률을 진화연산 중에 변화시킨 경우가 더 짧은 세대 만에 인공유전자조절네트워크와 동일한 유전자조절네트워크를 찾았다. 향후 실제 유전자 발현 데이터를 이용하여 실제 유전자조절네트워크를 추론하는 연구가 필요하다.

참 고 문 헌

[1] D. Duggan and M. Bittner and Y. Chen and P. Meltzer and J. Trent, "Expression profiling using cDNA microarrays", nature genetics, vol. 21, 1999.

[2] H. Iba and A. Mimura, "Inference of a gene regulatory network by means of interactive evolutionary computing"

Information Sciences, Vol. 145, pp. 225-236, 2002.

[3] D. Repsilber and H. Liljenstrom and S. Andersson, "Reverse engineering of regulatory networks: simulation studies on a genetic algorithm approach for ranking hypotheses" BioSystems, Vol. 66, pp. 31-41, 2002.

[4] M. Wahde and J. Hertz, "Coarse-grained reverse engineering of genetic regulatory networks", BioSystems, Vol 55. pp. 129-136, 2000.

[5] E. Sontag and A. Kiyatkin and B. Kholodenko, "Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data", BioInformatics, Vol. 20, pp. 1877-1886, no. 12 2004.

[6] S. Ando and H. Iba, "Inference of Gene Regulatory Model by Genetic Algorithms", IEEE, Vol 1, pp. 712-719, 2001.

[7] T. Reil, "Dynamics of gene expression in an artificial genome - implications for biological and artificial ontogeny," in *Fifth European Conference on Artificial Life* pp. 457-466, 1999.