다중 진화 알고리즘에 의한 유전자 조절 네트워크의 효율적인 탐색

김기영O 조동연 장병탁
서울대학교 컴퓨터공학부 바이오지능연구실
{kykimO, dycho, btzhang}@bi.snu.ac.kr

Efficient Identification of Gene Regulatory Networks by Multi-Stage Evolutionary Algorithms

Kee-Young KimO, Dong-Yeon Cho, and Byoung-Tak Zhang
Biointelligence Lab, School of Computer Science & Engineering, Seoul National University

요 약

DNA 마이크로어레이 기술의 발전으로 유전자 발현에 대한 많은 양의 정보가 쏟아지게 되었고, 이러한 정보들을 이용하여 유전자 조절 네트워크를 수학적으로 모델링하는 것이 시스템 생물학의 중요 관심사로 떠오르고 있다. 본 논문에서는 실험에서 얻어낸 데이터를 유전 프로그래밍을 이용한 기호 회귀를 통해 데이터 지점을 조정하고 유전 프로그래밍의 결과 함수를 이용해 각 지점에서의 미분값을 얻어내었다. 그 뒤, 불리안 네트워크를 표현하는 이진 배열과 S-시스템을 표현하는 실수 배열을 결합한 해를 사용하는 유전 알고리즘으로 앞에서 얻은 데이터를 이용해 원하는 S-시스템의 구조와 매개변수를 구해내었다.

1. 서론

DNA 마이크로어레이 기술의 발전으로 유전자 발현에 대한 많은 양의 정보가 쏟아지게 되었고, 이러한 정보들을 이용하여 유전자 조절 네트워크를 수학적으로 모델링하는 것이 시스템 생물학의 중요 관심사로 떠오르고 있다. 유전자 조절 네트워크를 모델링하는 방법으로 많이 쓰이는 것 중 하나가 S-시스템이다[1]. S-시스템은 특정한 형태를 취하는 미분 방정식의 집합으로 표현되며, 네트워크의 구조와 네트워크의 노드간의 비선형적인 관계도 나타낼 수 있기 때문에 유전자 조절 네트워크를 모델링하는데 적합하다. 하지만, S-시스템은 너무 많은 매개변수를 가지는 단점이 있다. 이런 단점을 극복하기 위해, S-시스템을 탐색할 때는 진화 연산이 많이 사용된다. 진화 연산은 동시에 많은 수의 매개변수를 탐색해 줄 수 있는 장점이 있다[2, 3].

그러나 진화 연산을 이용해 S-시스템을 탐색한다 해도 두 가지 어려움이 있다. 첫째는 S-시스템에 의해 표현될 수 있는 유전자 조절 네트워크의 구조가 너무나 다양하기 때문에 그 구조를 찾기가 어렵다는 점이고, 둘째는 특정 해에 의해 주어진 S-시스템을 평가하기 위한 수치 적분(numerical integration) 과정의 시간이 오래 걸린다는 것이다. 특히, 지역 최적화(local optimization) 기법을 사용하게 되면, 한 해(chromosome)를 생성하기 위해 수백에서 수천에 이르는 수치 적분 과정이 필요하게 되므로 수행 속도가 매우 늦어지게 된다.

이러한 단점을 보완하기 위해, [4]에서는 먼저 주어진 데이터를 인공신경망(artificial neural network)으로 학습시킨 다음, 그 인공신경망을 이용해 전체 유전자의 시간에 따른 발현 그래프를 얻어내었다. 실험 데이터로부터 전체 그래프를 생성하여 해의 데이터와 비교하는 것으로 이렇게 하면 각 지점(point)에서의 미분값을 미리 알 수 있으므로 수치 적분 과정 없이도 해를 평가할 수 있어 속도를 향상시킬 수 있다. 뿐만 아니라 이렇게 되면 각각의 유전자를 독립적으로 다루어 하나의 큰 문제를 작은 문제로 나누는 효과를 볼 수 있으므로 문제의 복잡도(complexity)를 크게 낮출 수 있다. 하지만 이 경우에는 구한 작은 문제들의 해를 나중에 하나로 합쳐주는 과정이 필요하게 된다.

본 논문에서는 [4]에서 제시된 인공신경망 대신에 유전 프로그래밍(genetic programming) 기법을 이용한다. 유전 프로그래밍을 이용하여 인공신경망에서 필요한 히든 레이어, 히든 유닛의 갯수 조절 과정이 없어졌고, 데이터도 100개나 필요하였으나 여기서는 26개의 데이터만을 사용해 데이터의 수를 현실적으로 실험 가능한 정도로 줄였다 또한, 유전자 조절 네트워크의 구조를 효율적으로 찾아주기 위해 구조만을 탐색하는 이진 배열을 유전 알고리즘의 해에 포함시켜 보다 효율적으로 탐색할 수 있도록 하였다

2. 유전자 조절 네트워크의 효율적인 탐색을 위한 다중 진화 알고리즘

2.1 S-시스템

S-시스템은 식 (1)과 같은 연립미분 방정식으로 표현된다. 이것은 n 개의 서로 다른 물질이 상호 영향을 주고받는 관계를 동적으로 표현할 수 있다[1, 5].

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}} \quad (1 \le i, j \le n) \qquad (1)$$

여기서 a_i 와 β_i 는 속도 상수(rate constant)라 하고, g_{ij} 와 h_{ij} 는 활동 차수(kinetic order)라 부른다.

S-시스템을 유전 알고리즘으로 찾아내는 과정은 우선, 유전 알고리즘을 통해 얻어진 임의의 해를 수치 적분 과정을 통해 각 유전자의 시간에 따른 발현량 그래프로 바꾸고 이를 실험을 통해 미리 얻은 데이터와 식 (2)를 통해 비교해 각 해의 적합도(fitness)를 측정한다.

$$E = \sum_{i=1}^{n}\sum_{t=1}^{T}\left(\frac{X_i(t) - \hat{X}_i(t)}{X_i(t)}\right)^2 \qquad (2)$$

(T =데이터의 개수, X = 참 값, $\hat{X}$ = 수치 적분 값)

2.2 유전 프로그래밍을 이용한 기호 회귀

그러나 서론에서 언급했듯이 S-시스템을 매번 수치 적분을 통해 구하는 것은 매우 느리므로, 본 논문에서는 다음과 같이 두 단계를 거쳐 문제를 해결한다. 첫 단계는 유전 프로그래밍[6]을 이용한 기호 회귀(symbolic regression)를 통해 주어진 데이터의 곡선을 구하는 것이다. 이를 통해 각 지점에서의 미분값을 얻어낸다. 두 번째 단계는 각 지점에서의 데이터와 학습한 곡선에서 얻은 미분값을 이용해 유전 알고리즘을 최적화하여 최종적으로 S-시스템을 찾아내는 것이다.

유전 프로그래밍으로 주어진 26개의 데이터를 지나는 곡선을 구하고, 이 곡선에서 새로운 100개의 점을 얻었다. 각 점에서의 미분값도 수치적인 방법을 이용해 구했다.

2.3 하이브리드 유전 알고리즘을 이용한 S-시스템 탐색

두 번째 단계에서는, 유전자 조절 네트워크의 구조와 매개변수를 동시에 찾아줄 수 있는 하이브리드 유전 알고리즘을 개발하였다. 유전 알고리즘의 적합도 함수는 식 (3)과 같다. 식 (3)에서 $\ddot{X}$ 을 식에 곱해주는 것은 그래프에서 상대적으로 변화율이 큰 언덕이나 계곡 부분을 정확하게 예측하는 것이 올바른 해를 찾는데 중요한 조건이 되기 때문이다.

$$E = \sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{X}_i(t)\left(\frac{\dot{X}_i(t) - X'_i(t)}{\dot{X}_i(t)}\right)^2 \qquad (3)$$

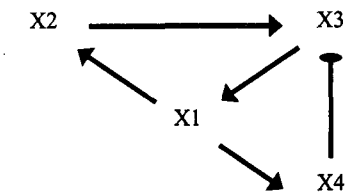
($\dot{X}$ = 미분값, $\ddot{X}$ = 이계미분값, X' = S-시스템으로 얻은 값)

유전 알고리즘의 해는 불리안 네트워크 형태로 유전자 조절 네트워크의 구조를 나타낸 이진 배열과 S-시스템을 나타내는 실수 배열 두 가지를 사용한다. 이렇게 구성된 해는 교차, 돌연변이 연산에서는 불리안 네트워크를 이용해 유전자 네트워크의 구조를 탐색하고, 지역 최적화 연산에서는 실수 배열을 이용해 교차, 돌연변이 연산을 통해 찾아낸 유전자 조절 네트워크의 구조에서 S-시스템의 각 매개변수를 최적화한다. 교차, 돌연변이 연산 모두 부모로 선택된 두 해에서 한번에 한 행에만 변화를 주도록 되어있다. 그것은 작은 변화로도 크게 결과가 달라지는 S-시스템의 특성상, 서로 독립인 한 행 내에서만 변화가 일어나도록 조절해준 것이다.

3. 실험 및 결과

3.1 데이터

제안된 방법의 성능을 검증하기 위하여 [4]에서 다루었던 인공적인 유전자 조절 네트워크를 사용하였다. 이는 4개의 가상 유전자가 서로 영향을 주고 받는 것을 모델링한 것으로서, S-시스템으로 그림 1의 (a)와 같이 나타내어진다. 여기에 초기값 X_1=1.4, X_2=2.7, X_3=1.2, X_4=0.4를 주어 푼 다음 시간단위 0에서 5까지의 범위에서 0.2 간격으로 26개의 점을 얻어 실험 데이터 값으로 이용하였다. (그림 2)



(a) 그래프 형태의 표현

$$\frac{dX_1}{dt} = 12X_3^{0.8} - 10X_1^{0.5} \qquad \frac{dX_2}{dt} = 8X_1^{0.5} - 3X_2^{0.75}$$

$$\frac{dX_3}{dt} = 3X_2^{0.75} - 5X_3^{0.5}X_4^{0.2} \qquad \frac{dX_4}{dt} = 2X_1^{0.5} - 6X_4^{0.8}$$

(b) S-시스템 형태의 표현

그림 1. 문제로 사용된 유전자 조절 네트워크

3.2 유전 프로그래밍의 결과

유전 프로그래밍 과정은 Frayn이 구현한 GPLib[7]에 약간의 수정을 가하여 진행하였다. 사용된 연산자는 가감승제(+, -, *, /)와 거듭제곱, 지수, sin함수를 사용하였다. 해집단의 크기는 3000, 세대는 2000이며, 크기 4인 토너

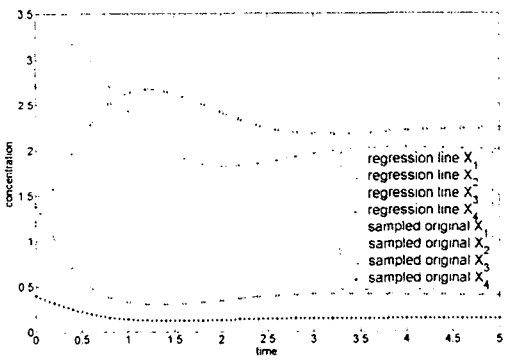


그림 2. 실험 데이터와 유전 프로그래밍으로 얻은 곡선

먼트 선택을 사용하고, 교차율은 0.35, 돌연변이율은 0.5로 설정하였다.

유전 프로그래밍을 통해 얻은 그래프와 사용된 26개의 데이터를 그림 2에 나타내었다. 여기서 얻은 100개의 점과 그 기울기는 다음 하이브리드 유전 알고리즘 단계의 입력으로 사용된다.

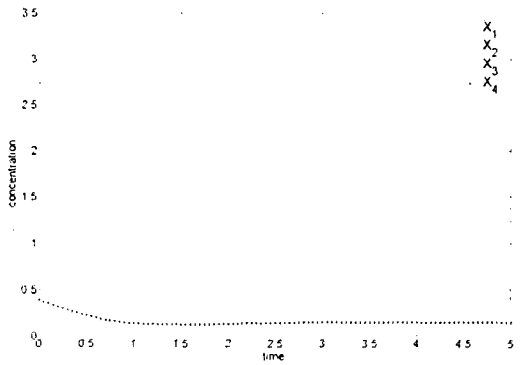
3.3 하이브리드 유전 알고리즘의 결과

유전 프로그래밍으로 얻은 100개의 점과 그 기울기를 이용하여 하이브리드 유전 알고리즘을 통해 S-시스템을 탐색한다. 해집단의 크기는 10,000, 세대는 1,000,000이다. 선택 연산으로는 해의 다양성을 유지하기 위해 제한된 토너먼트 선택(restricted tournament selection)[8]을 사용한다. 교차율은 0.8, 돌연변이율은 0.3이다. 매개변수 최적화를 위한 지역 최적화 알고리즘으로는 (1+1) 진화 전략[9]을 사용한다. 하이브리드 유전 알고리즘 수행 결과, 목표 S-시스템의 구조와 매개변수들을 찾아내는데 성공하였으며, 그 결과를 그림 3에 나타내었다. 그림 3의 (b)에서는 X_1이 X_2를 활성화시키고 다시 X_2가 X_3를 활성화시키는 모습을 보여주고 있어 그림 2에서 주어진 데이터의 특성을 만족시키고 있다.

$$\alpha=\begin{pmatrix}14.614\\8.123\\3.088\\2.866\end{pmatrix}\quad g=\begin{pmatrix}0.0 & 0.0 & -0.623 & 0.0\\0.486 & 0.0 & 0.0 & 0.0\\0.0 & 0.761 & 0.0 & 0.0\\0.254 & 0.0 & 0.0 & 0.0\end{pmatrix}$$

$$\beta=\begin{pmatrix}12.670\\3.076\\5.583\\5.391\end{pmatrix}\quad h=\begin{pmatrix}0.398 & 0.0 & 0.0 & 0.0\\0.0 & 0.753 & 0.0 & 0.0\\0.0 & 0.0 & 0.486 & 0.231\\0.0 & 0.0 & 0.0 & 0.443\end{pmatrix}$$

(a) 하이브리드 GA를 통해 얻은 S-시스템



(b) 예측된 S-시스템의 시간에 따른 변화 그래프

그림 3. 하이브리드 유전 알고리즘의 결과

4. 결론

지금까지 유전자 조절 네트워크를 찾기 위한 다단계 진화 알고리즘을 제시하였다. 시간이 오래 걸리는 수치 적분 과정을 없애기 위해 유전 프로그래밍을 통한 기호 회귀 단계를 두었고, 이중적인 해 구조와 개량된 적합도 함수를 가진 하이브리드 유전 알고리즘을 이용하여 유전자 조절 네트워크의 구조와 매개변수를 동시에 찾아내었다. 제안된 방법을 인공적인 유전자 조절 네트워크에 적용해 본 결과 목표로 한 S-시스템을 찾을 수 있음이 입증되었다.

감사의 글

본 연구는 교육인적자원부 BK21-IT, 산업자원부 차세대 신기술 개발 사업의 분자 진화 컴퓨팅 과제 및 과학기술부 국가지정연구실 과제에 의하여 일부 지원되었다. 또한 이 연구를 위해 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에도 감사드린다.

참고문헌

[1] Voit, E.O., *Computational Analysis of Biochemical Systems*, Cambridge University Press, 2000.

[2] Ando, S., Sakamoto, E., and Iba, H., "Evolutionary modeling and inference of gene network", *Information Sciences*, vol. 145, no. 3, pp. 237-259, 2002.

[3] Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., and Tomita, M., "Dynamic modeling of genetic networks using genetic algorithm and S-system", *Bioinformatics*, vol. 19, no 5, pp. 643-650, 2003.

[4] Almeida, J.S. and Voit, E.O., "Neural network-based parameter estimation in S-system models of biological networks", *Genome Informatics*, vol. 14, pp. 114-123, 2003.

[5] Savageau, M.A., *Biochmical System Analysis: A Study of Function and Design in Molecular Biology*, Addison-Wesley, 1976.

[6] Koza, J.R., *Genetic Programming: On the Programming of Computers by Natural Selection*, MIT Press, 1992.

[7] http://www.cs.bham.ac.uk/~cmf/GPLib/GPLib.html

[8] Harik, G.R., "Finding multimodal solutions using restricted tournament selection", *Proceedings of the Sixth International Conference on Genetic Algorithms*, pp. 24-31, 1995.

[9] Bäck, T., *Evolutionary Algorithm in Theory and Practice*, Oxford University Press, 1996.