생화학 시스템의 동적 모델링을 위한 S-tree 기반의 진화연산

조동연O 장병탁
서울대학교 컴퓨터공학부 바이오지능연구실
{dychoO, btzhang}@bi.snu.ac.kr

S-tree-Based Evolutionary Computation for Dynamic Modeling of Biochemical Systems

Dong-Yeon Cho0 Byoung-Tak Zhang
Biointelligence Laboratory
School of Computer Science and Engineering, Seoul National University

요 약

시간이 흐름에 따라 생화학 시스템이 변화하는 것을 기록한 데이터로부터 이 시스템의 상태 전이 및 시스템을 구성하는 각 생화학 물질간의 관계를 모델링하기 위한 방법으로 S-tree 구조를 제안한다. 이것은 주로 생화학 시스템의 동적 특성을 모델링 하기 위하여 연구되어 온 S-system을 나무 구조로 표현한 것이다. 본 논문에서는 진화 연산을 통해 주어진 시계열 데이터를 잘 설명하는 S-tree의 구조 및 그 변수들을 동시에 효과적으로 탐색하는 방법을 개발하였다. 이 방법에서는 구조 탐색을 위해 유전 프로그래밍(genetic programming)에서 사용되어 온 나무 구조의 교차 및 돌연변이 연산과 더불어 다양한 형태의 구조 탐색 연산자들을 도입하였고, 또한 동시에 알맞은 변수 값들을 찾기 위하여 확률적 돌연변이 연산을 통한 언덕오르기(hill-climbing)를 수행한다. 제안된 방법을 효모의 혐기성 발효 데이터에 적용한 결과 주어진 시스템을 성공적으로 모델링할 수 있었다.

1. 서론

바이오인포매틱스(bioinformatics) 분야가 발전함에 따라 생명과학 연구의 방향이 개별적인 현상을 연구하던 기존의 방식으로부터 벗어나 많은 유전자나 단백질과 같은 생체 분자들의 상호 작용을 연구하는 쪽으로 옮겨가고 있다. 또한 이러한 경향은 생물학을 시스템 관점에서 연구하는 시스템 생물학(systems biology)의 발전을 촉진시키고 있다. 이러한 관점에서 볼 때, 어떤 복잡한 생화학 시스템을 수학적인 모델을 통하여 표현하고 이를 분석함으로써 그 시스템의 내재적인 기능을 이해하고 환경의 변화에 따른 반응 양상의 예측하는 것은 의미있는 일이라 하겠다.

그러나 시간이 흐름에 따라 생화학 시스템이 변화하는 것을 기록한 데이터로부터 이 시스템의 상태 전이 및 시스템을 구성하는 각 생화학 물질간의 관계를 설명해 줄 수 있는 적합한 모델을 만드는 것은 어려운 일이다. 즉, 어떠한 방식으로든 물질간의 관계를 나타낼 수 있는 구조를 찾아내야 하고, 서로 어느 정도로 영향을 미치는가를 표현하는 변수 값들을 결정해 주어야 한다.

이와 같이 주어진 데이터에 대한 알맞은 생화학 시스템 모델을 찾기 위한 시도로서 진화 연산이 사용되어 왔다. 예를 들면 시간에 따른 유전자 조절 네트워크의 변화를 잘 설명하기 위해서 연립 미분 방정식의 형태로 모델을 설정하고 그 방정식의 형태 및 변수 값들을 결정하기 위한 방법으로 유전 프로그래밍 기법이 사용되었다[2]. 하지만 시스템의 크기가 커질수록 임의의 형태로 주어지는 수많은 연립 미분 방정식 중에서 주어진 데이터에 맞는 구조와 변수 값들을 찾는 것은 거의 불가능 하다.

따라서 생화학 시스템 이론(Biochemical System Theory)[2]에서 제공하는 어느 정도 정형화된 형태의 모델로서 주어진 시스템을 설명하는 것이 더 적합하다. 이러한 것 중 대표적인 것이 2절에서 설명할 S-system이다. 유전자 네트워크를 표현하는 S-system의 구조 및 변수를 학습하기 위한 유전 프로그래밍 기법이 [3]에서 제시 되었으나, 실수 값을 갖는 변수를 표현하기에는 단말 노드의 형식에 제약이 있어 부적합하다. 또한 유전 알고리즘(genetic algorithm)을 사용하고 모델의 복잡도를 제한하는 방법으로 S-system을 학습하는 방법도 제시되었다[4]. 그러나 이 방법에서도 유용한 구조를 직접 찾을 수 없기 때문에 너무 많은 시간이 소요되고, 또한 모델의 복잡도를 제한하는 정도에 따라서 그 성능이 심하게 영향을 받는다.

본 논문에서는 S-system의 구조를 효과적으로 표현할 수 있는 나무 구조의 S-tree 표현 방법을 제안하고, 주어진 데이터로부터 S-tree의 구조 및 그 변수 값들을 효율적으로 탐색하기 위한 진화 연산 기법을 개발하였다. 제안된 방법을 두 종류의 생화학 시스템 데이터에 적용한 결과 주어진 데이터가 표현하는 시스템을 성공적으로 모델링할 수 있었다.

논문의 구성은 다음과 같다. 2절에서는 생화학 시스템의 동적 모델링을 위한 S-system을 개략적으로 설명하고, 이것의 나무 구조 형태인 S-tree 표현 기법을 제안한다. 3절에서는 S-tree의 구조 및 변수 값들을 탐색하기 위한 진화 연산 기법을 설명하고, 제안된 방법을 적용한 실험 결과를 4절에서 보인다. 끝으로 5절에서는 결론과 함께 앞으로의 연구 방향을 제시한다.

2. S-system과 S-tree

S-system은 다음과 같은 정형화된 형태의 연립 미분 방정식으로 표현된다.

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}}, \quad i,j = 1,2,\ldots,n$$

여기서 X_i는 생화학 시스템을 구성하는 각 물질의 농도를 나타내며, 이 중 n개는 시간에 따라 그 농도가 변하는 물질이고 m개는 항상 같은 농도를 유지하는 물질이다. 또한 실수 값을 갖는 g_{ij}와 h_{ij}는 X_i에 대한 X_j의 영향을 정량적으로 표시한다. 그리고 음수가 아닌 실수 값으로 나타나는 a_i와 β_i는 반응 속도 상수를 의미한다. 그리고 위 식에서 첫 번째 부분은 X_i를 증가시키는데 영향을 주는 것들을 나타내며, 두 번째 부분은 반대로 감소시키는데 영향을 주는 것들을 나타낸다. 결국 S-system을 이용하여 어떤 생화학 시스템을 표현하기 위해서는 총 $2n(n+1)$개의 변수 값들을 결정해 주어야 한다.

그러나 그림 1의 위쪽 행렬 표현에서 보이는 바와 같이 S-system을 정의하기 위한 변수 값들 중에서 g와 h는 원소의 값들이 대부분 0인 희소 행렬(sparse matrix)로 나타나게 된다. 즉, 이것은 생화학 시스템을 구성하는 각 물질들은 그 시스템 내에서 몇 개 정도의 물질들과만 상호 작용을 하고 있다는 사실을 암시한다. 따라서 총 $2n(n+1)$개의 변수 값들을 모두 나타낼 필요 없이 0이 아닌 부분을 찾아내어 그 변수 값들을 정확하게 결정하는 것이 중요하다.

이를 위하여 본 논문에서는 그림 1의 아래쪽에서 보여주는 바와 같은 S-tree 구조를 제안한다. 먼저 루트 노드(R)는 S-system을 구성하는 미분 방정식의 개수(n) 만큼 자식 노드를 갖는다. 즉, 각 서브 트리는 S-system을 구성하는 n개의 미분 방정식을 표현하게 된다. 그 다음 X_i 노드들은 각 미분 방정식에서 첫 번째 부분과 두 번째 부분을 나타내기 위한 2개의 자식을 갖게 된다. 이 때 그 연결 가중치로는 각각 a_i와 β_i의 값을 표현한다. $g_{i.}$와 $h_{i.}$ 노드는 g와 h 행렬의 각 열에서 0이 아닌 원소의 개수 만큼을 그 자식으로 갖게 되며, 단말 노드가 그 값이 0이 아닌 원소를 나타낸다. 그리고 그 연결 가중치가 변수의 값을 표현한다.

$\alpha = (5.0 \;\; 10.0 \;\; 8.0 \;\; 10.0 \;\; 10.0)$, $\beta = (10.0 \;\; 10.0 \;\; 10.0 \;\; 10.0 \;\; 10.0)$

$$g = \begin{pmatrix} 0.0 & 0.0 & 1.0 & 0.0 & -1.0 \\ 2.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & -1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 2.0 & 0.0 & -1.0 \\ 0.0 & 0.0 & 0.0 & 2.0 & 0.0 \end{pmatrix}, \quad h = \begin{pmatrix} 2.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 2.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & -1.0 & 2.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 2.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 2.0 \end{pmatrix}$$

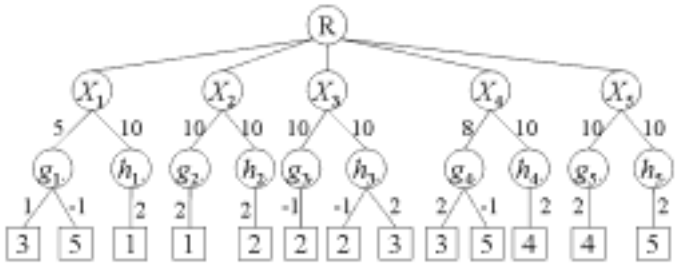

그림 1. S-system의 행렬 표현 및 그에 상응하는 S-tree의 예 (n=5, m=0)

3. S-tree의 탐색을 위한 진화 연산

시간의 흐름에 따라 관측되는 생화학 시스템의 시계열 데이터에 적합한 S-system을 나타내는 S-tree를 얻기 위해서는 0이 아닌 원소의 개수 및 그 위치를 정확하게 표현하는 S-tree의 구조와 연결 가중치로 표현되는 각 변수의 값들을 찾아내야 한다. 먼저 구조의 학습을 위하여 유전 프로그래밍에서 나무 구조의 탐색을 위해 사용되는 교차 연산을 사용한다. 즉, 두 개의 부모 트리를 고른 후에 각 부모에서 임의로 선택된 서브 트리를 서로 교환한다. 이 때 교차점은 같은 X_i 노드나 $g_{i.}$ 또는 $h_{i.}$ 노드가 선택된다. X_i가 선택된 경우에는 전체 시스템 중 하나의 미분 방정식이 교차되는 것이고 $g_{i.}$나 $h_{i.}$ 노드가 선택되는 것은 하나의 열이 교차되는 것이다. 또한 임의의 서브 트리를 선택하여 그것을 임의로 생성된 다른 서브 트리로 교체하는 돌연 변이 연산도 사용된다. 여기에 추가적으로 단말 노드를 하나 추가하거나 삭제하는 연산과 단말 노드의 값, 즉 0이 아닌 위치를 나타내는 값을 바꾸는 연산도 추가 되었다. 이러한 돌연변이 연산자들은 현재 개체군 내에 없는 새로운 구조를 만들어 낼 수 있기 때문에 구조 탐색에 유용하게 사용될 수 있다.

a_i와 β_i의 값 그리고 0이 아닌 g_{ij}와 h_{ij}의 값을 나타내는 가중치는 언덕 오르기(hill-climbing)에 의하여 조정된다. 가중치 변경은 주어진 S-tree를 후위 순회하며 이루어지게 되는데 이 때 사용되는 변경 값은 평균이 0이고 표준 편차가 1인 정규 분포에서 추출된다. 새롭게 추출된 값과 원래의 변수 값을 더한 후 새롭게 조정된 가중치에 의하여 표현된 S-system에 더 좋은 적합도를 갖는다면 그 값을 수용하고 그렇지 않으면 원래의 가중치 값을 유지한다.

S-tree의 적합도는 다음과 같이 주어진 데이터와의 상대적인 제곱 오차로 표현한다.

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

여기서 $X'_i(t)$는 S-tree에 의하여 예측된 값이고, $X_i(t)$는 실제로 관측되어 데이터로 주어진 값이다. 또한 T는 주어진 시계열 데이터 개수이다.

전체적인 진화 연산의 흐름은 다음과 같다. 총 N개의 S-tree를 임의로 생성하여 초기 개체군을 만든 후, 적합도를 평가한다. 다음으로 개체군에서 임의로 부모 2개를 선택하여 교차 연산 확률에 따라 교차 연산을 수행한 후 다시 돌연변이 연산을 그 확률에 따라 수행한다. 이렇게 하여 λ개의 새로운 S-tree를 만든 후, 각 자손에 대하여 언덕 오르기를 적용한다. 이 중 적합도가 가장 좋은 μ개를 선택하여 개체군에서 적합도가 나쁜 개체들을 대체한다. 여기서 개체군의 다양성을 보장하기 위한 방법으로 개체군 내에 같은 구조의 S-system을 표현하는 S-tree가 존재한다면 그 개체를 대체하도록 하였다. 만약 같은 구조를 나타내는 S-tree가 없다면 적합도가 나쁜 개체들 중에서 임의로 선택된 것을 대체한다. 이 과정을 정해진 세대 또는 적합도 평가 회수만큼 반복한 후 가장 제곱 오차가 적은 것을 해로 돌려준다.

4. 실험 및 결과

제안된 방법을 검증하기 위하여 다음과 같은 S-system으로 표시되는 데이터로부터 얻어진 값들을 사용하였다.

$$\frac{dX_1}{dt} = 0.8122X_2^{-0.2344}X_6 - 2.8632X_1^{0.7464}X_5^{0.0243}X_7$$

$$\frac{dX_2}{dt} = 2.8632X_1^{0.7464}X_5^{0.0243}X_7 - 0.5239X_2^{0.735}X_5^{-0.394}X_8^{0.999}X_{11}^{0.001}$$

$$\frac{dX_3}{dt} = 0.5232X_2^{0.7318}X_5^{-0.3941}X_8 - 0.0148X_3^{0.584}X_4^{0.03}X_5^{0.119}X_9^{0.944}X_{12}^{0.056}X_{14}^{-0.575}$$

$$\frac{dX_4}{dt} = 0.022X_3^{0.6159}X_5^{0.1308}X_9X_{14}^{-0.6088} - 0.0945X_3^{0.05}X_4^{0.533}X_5^{-0.0822}X_{10}$$

$$\frac{dX_5}{dt} = 0.0913X_3^{0.333}X_4^{0.266}X_5^{0.024}X_9^{0.5}X_{10}^{0.5}X_{14}^{-0.304} - 3.2097X_1^{0.198}X_2^{0.196}X_5^{0.372}X_7^{0.265}X_8^{0.265}X_{11}^{0.0002}X_{13}^{0.47}$$

이것은 효모의 혐기성 발효(anaerobic fermentation)를 나타낸 식(n=5, m=9)으로서[5] 본 논문에서는 시간에 따라 변하는 5개의 변수와 관련된 구조 및 변수만을 예측하고 그렇지 않은 나머지 9개의 변수에 대한 값들은 알고 있는 것으로 가정하였다.

X_i의 초기 농도를 (0.2, 1.5, 6.9, 0.009, 2.0)로 설정한 후, Euler법에 의하여 주어진 연립미분방정식을 총 500단위 시간 동안 시뮬레이션하여 매 5 단위 시간마다의 데이터 값을 진화 연산에서 사용하였다 (T=100). S-tree의 구조는 단말 노드가 3개 이하를 가지도록 제한했으며, 개체군의 크기는 1,000이었다. 교차 확률은 0.75, 돌연변이 확률은 0.6으로 설정하였으며, λ와 μ의 값은 각각 100과 10을 사용하였다. 그리고 a_i와 β_i는 [0, 4]에서, g_{ij}와 h_{ij}는 [-1, 1]에서 그 값을 찾도록 하였다. 끝으로, 진화 연산은 3,000세대 후 종료 되었다.

AMD Athlon MP 2000+ CPU가 장착된 Windows XP 기반의 PC에서 약 24시간 동안의 수행 후에 상대적인 제곱 오차가 0.163인 S-tree가 얻어 졌다. 그림 2는 실제 S-system 식을 이용하여 얻어진 데이터와, 제안된 진화 연산에 의하여 얻어진 S-tree 구조로 표현된 S-system을 학습 데이터를 얻는 것과 같은 방법으로 시뮬레이션한 결과를 보여주고 있다. 데이터로 주어진 500단위 시간 내의 시스템 변화 양상뿐만 아니라 그 후 500단위 시간 동안, 즉 안정화 상태 후의 시스템 양상도 거의 유사하게 예측하였음을 알 수 있다.

그러나 우리가 찾은 S-tree를 좀 더 자세히 분석해 본 결과, g와 h 행렬에서 21개의 원소가 0이 아닌 값을 갖고 있었다. 이것은 위의 식으로 주어진 실제 시스템이 총 23개의 0이 아닌 원소를 갖고 있는 것과는 약간 다른 결과이다. 또한 a_i와 β_i의 값들에 대해서도 각각 0.5와 1.1 정도의 평균 제곱 오차가 발생하였다. 이러한 것들이 시스템이 안정 상태로 들어간 후의 결과에서 약간의 오차를 발생시키는 원인이라고 생각된다.

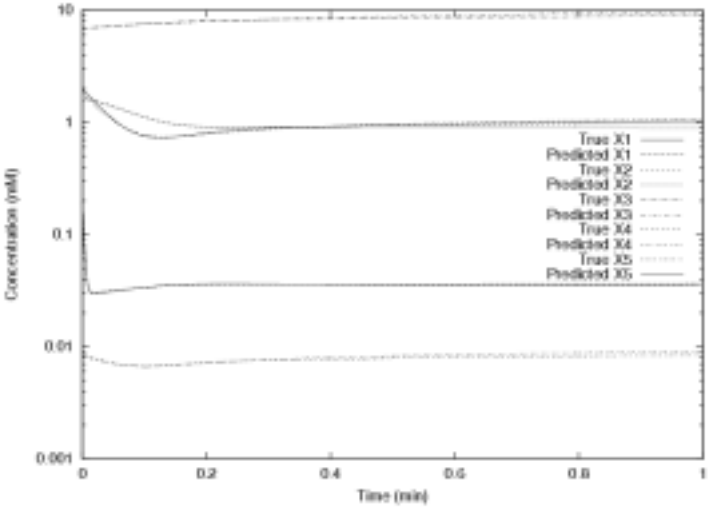

그림 2. 실제 S-system과 진화 연산을 통해 얻어진 S-tree로 표현된 시스템의 시간에 따른 변화

5. 결론

본 논문에서는 시계열 데이터에 의하여 표현되는 생화학 시스템을 효율적으로 모델링하기 위한 방법으로 S-system의 나무 구조 형태인 S-tree를 제안하였다. 또한 주어진 데이터에 알맞은 S-tree의 구조 및 변수 값들을 찾기 위한 방법으로서 여러 가지 구조 탐색 연산과 확률적 돌연변이 연산을 통한 언덕 오르기를 수행하는 진화 연산을 개발하였다. 제안된 방법을 효모의 혐기성 발효 데이터에 적용한 결과 주어진 시스템을 효과적으로 모델링 할 수 있었다.

감사의 글

본 연구는 과학기술부 국가지정연구실사업(NRL)과 Systems Biology 사업, 교육인적자원부 BK21-IT, 산업자원부 Molecular Evolutionary Computing (MEC) 과제에 의하여 지원되었음.

참고 문헌

[1] E. Sakamoto and H. Iba, Inferring a system of differential equations for a gene regulatory network by using genetic programming, *Proceedings of the 2001 Congress on Evolutionary Computation*, vol. 1, pp. 720-726, 2001.

[2] M. A. Savageau, *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*, Addison-Wesley, 1976.

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

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

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