진화연산에 기반한 유전자 발현 데이터로부터의 유전자 상호작용 네트워크 구성

Construction of Gene Interaction Networks from Gene Expression Data Based on Evolutionary Computation

정 성 훈, 조 광 현*
(Sung Hoon Jung and Kwang-Hyun Cho)

Abstract : This paper investigates construction of gene (interaction) networks from gene expression time-series data based on evolutionary computation. To illustrate the proposed approach in a comprehensive way, we first assume an artificial gene network and then compare it with the reconstructed network from the gene expression time-series data generated by the artificial network. Next, we employ real gene expression time-series data (Spellman's yeast data) to construct a gene network by applying the proposed approach. From these experiments, we find that the proposed approach can be used as a useful tool for discovering the structure of a gene network as well as the corresponding relations among genes. The constructed gene network can further provide biologists with information to generate/test new hypotheses and ultimately to unravel the gene functions.

Keywords : microarray, gene interaction networks, time-series gene expression data, evolutionary computation

I. 서론

분자생물학 및 유전공학의 발달과 더불어 유전자 사이의 상호작용을 규명하는 유전자 (상호작용) 네트워크의 탐구에 대한 연구가 많이 진행되고 있다 [1-14]. 유전자 네트워크의 탐구란 유전자 칩(DNA microarray)과 같은 고효율(high throughput)실험으로 얻어지는 대규모 유전자 발현 시계열(time-series)데이터로부터 유전자들 상호간의 작용메커니즘을 밝히는 작업으로서 유전자의 기능을 규명하는데 매우 중요한 연구이다 [3-14]. 이러한 작업은 유전자 상호작용에 의한 결과물(유전자 발현 데이터)을 이용하여 역으로 그 원인이 되는 유전자 상호작용을 유추해내는 과정이며 일종의 역공학 (reverse engineering) 이라고 볼 수 있다.,

세포 및 분자수준에서의 기본적인 생명현상은 다음과 같다. 세포 내외부의 자극이 신호로 전달되면 세포 핵 내부 유전자들의 전사(transcription)를 통해 mRNA가 생성되고 생성된 mRNA가 번역(translation)되어 단백질(protein)이 생성되는 일련의 과정들이 일어나게 된다. 생성된 단백질은 다시 유전자를 조절하기도 하고 다른 단백질에 상호작용하기도 하며 신진대사(metabolism)를 조절하게 된다. 이와 같은 과정에 있어서 유전자 사이의 상호작용을 규명하는 것은 궁극적으로 관련 신진대사의 메커니즘을 밝히거나 유전자의 역할을 규명하는데 있어서 기본적으로 필요한 과정으로서 단백질 기능의 분석 및 신약개발 등에 유용한 정보를 제공할 수 있다.

* 책임저자(Corresponding Author)
논문접수 : 2004. 9. 20., 채택확정 : 2004. 10. 26.
정성훈 : 한성대학교 정보공학부(shjung@hansung.ac.kr)
조광현 : 서울대학교 의과대학 의학과 및 서울대학교 생명공학연구원(ckh-sb@snu.ac.kr)
※ 본 연구는 2003년도 한성대학교 교내연구비로 수행되었습니다.

유전자 발현데이터로부터 유전자 네트워크를 재구성하기 위해서는 크게 다음과 같은 두 가지 과정이 필요하다. 첫 번째는 유전자 네트워크의 구조를 결정하는 모델링 과정이다. 이는 실제 유전자 네트워크를 알 수 없는 상황에서 실제 유전자 네트워크와 유사한 모델을 유추해내는 작업으로서 부울 네트워크 (Boolean networks) [3], 선형 네트워크 (linear networks) [4,5,9,11], 베이시안 네트워크 (Bayesian networks) [6], 미분방정식 (differential equations) [15], 회귀 신경망 (recurrent neural networks) [1,2,13,14] 등의 방법으로 접근되어져 왔다. 두 번째 과정은 유전자 발현 데이터를 이용하여 첫 번째 과정에서 정립된 유전자 네트워크 모델의 구성 파라미터를 추정해내는 것이다. 그동안 유전자 네트워크 모델의 파라미터를 결정하는 방법으로 SA(simulated annealing)법, 유전자 알고리즘 (genetic algorithm; GA), 하강 경사법 (gradient descent methods), 선형 회귀 분석 (linear regression methods) [4,5,8,9,11] 등이 사용되어져왔다.

그러나 현재까지는 상기 어떤 방법도 실제 유용한 결과를 도출해내지 못하고 있는데, 그 이유는 대략 다음과 같은 세 가지 요인에 기인한다. 첫째, 실험으로 얻어지는 유전자 발현데이터의 실험 샘플링 개수(<20)가 결정해야할 모델 파라미터의 개수(>500)에 비하여 매우 적다는 것이다. 따라서 결과적으로 추정되는 모델 파라미터의 값이 부정확하게 된다. 두 번째 요인으로는 역 공학 관점으로부터 주어진 허용 오차 범위 내에서 유전자 발현 시계열데이터를 잘 모사하는 유전자 네트워크의 모델이 다수 개 존재한다는 것이다. 결국 주어진 하나의 유전자 발현 시계열 데이터로는 비교적 정확한 파라미터를 구하는 것이 어려울 수밖에 없다. 마지막으로 일반적인 모델링 문제에서와 마찬가지로 모델 자체의 구조적 한계로 인하여 가정한 모델구조에 포함되지 않은 효과가 주어진 데이터에 이미 영향을 주었을 가능성이 있다

는 것이다. 물론 모델링에서 고려하지 않은 효과는 미미하다는 전제로 간과될 수 있지만 때로는 그 효과가 무시할 수 없을 만큼 작용할 수 있고 이러한 영향은 모델의 파라미터 결정에 큰 문제로 대두 될 수 있다. 이와 같은 이유들로 인하여 현재로서는 아직 확실히 만족할 만한 결과를 얻고 있지 못하고 있다. 단, 이러한 문제점들은 서로 다른 조건하에서의 실험 데이터를 여러 차례 습득하여 이용한다던지 기존에 미리 알려진 사전지식을 추가한다던지 보다 진보된 알고리즘으로 처리하는 등의 방법으로 어느 정도 보완이 가능하다.

본 논문에서는 유전자 네트워크 모델로 많이 사용되고 있는 회귀신경망을 사용하며 모델 파라미터를 결정하는 방법으로 진화연산 (보다 자세하게는 GA [16-18]와 진화 프로그래밍 (evolutionary programming; EP) [19])을 사용한다. GA는 범용 검색 (혹은 최적화) 알고리즘으로서 보통 문제의 해를 비트 문자열로 표현하여 집단적으로 진화시키는 과정을 통해 전역 최적 해를 찾아나가는 방법이다. 이에 비하여 EP는 해를 있는 그대로 사용하는 방법으로서 해가 최적 해에 근접한 정도에 역 비례하여 잡음(다른 표현으로 섭동)을 섞어 최적 해를 찾아나가기 때문에 실수해를 가질 때 유용한 방법이다.

우리는 먼저 역 공학 측면에서 제안하는 접근법이 유효한지를 보이기 위하여 인위적인 유전자 네트워크를 가정하고 이 유전자 네트워크로부터 유전자 발현 데이터를 생성한 뒤 이를 이용하여 역으로 유전자 네트워크를 구성하는 실험을 한다. 그 다음 실제로 실험을 통하여 얻은 유전자 발현 데이터를 이용하여 유전자 네트워크를 구성한다. 우리의 방법과 유사한 방법으로 기존 Wahde [13,14]의 결과가 있어서 실험 결과를 이와 비교해본다. 결과적으로 본 논문에서 제안하는 방법이 더욱 좋은 결과를 보였으나 그 원인을 정량적으로 비교분석하기에는 Wahde의 논문에서 방법을 기술하는 설명이 부족하여 확정적인 결론을 내리지는 못하였다. 또한 본 논문에서만 사용한 EP는 GA보다 더욱 좋은 결과를 보였다. 이를 통하여 GA와 EP가 모두 유전자 네트워크 구성에 효과적으로 사용될 수 있으며 특히 EP가 보다 적합함을 알 수 있었다.

본 논문의 구성은 다음과 같다. 2절에서는 유전자 네트워크 모델에 대하여 설명한다. 3절에서는 GA와 EP로 유전자 네트워크 모델의 파라미터를 결정하는 방법을 설명한다. 4절에서는 인위적으로 생성된 유전자 발현 시계열 데이터로부터 유전자 네트워크를 구성하는 과정과 실제 실험 유전자 발현 데이터로부터 유전자 네트워크를 구성하는 과정을 각각 설명한다. 마지막으로 5절에서 본 논문의 결론을 맺는다.

II. 유전자 네트워크 모델의 구조정립

유전자 네트워크는 부울 네트워크, 선형 네트워크, 베이시안 네트워크, 미분방정식, 회귀신경망 등으로 모델링이 시도되어져 왔다. 본 논문에서는 이 가운데 비선형성이 강한 실제 유전자 네트워크를 잘 모델링 할 수 있고 비교적 결과 해석에도 유용한 회귀신경망 모델을 사용한다. 회귀신경망 모델은 다음과 같은 미분방정식으로 표현된다.

$$\tau_i \frac{dx_i(t)}{dt} = g\left(\sum_{j=1}^{I} W_{i,j} x_j(t) + b_i \right) - x_i(t). \quad (1)$$

(1)에서 $x_i(t)$는 유전자 i의 시간 t에서의 발현 정도를 나타내며 함수 $g(\cdot)$는 활성화 함수 (activation function) 이고 $W_{i,j}$는 유전자 j가 유전자 i를 조절하는 정도이고 b_i와 τ_i는 각각 유전자 i의 바이어스와 시간상수 (time constant)이다. 보통 활성화함수로서 시그모이드 (Sigmoid) 함수 $g(z) = (1 + e^{-z})^{-1}$가 많이 사용된다.

$$z = \sum_{j=1}^{I} W_{i,j} x_j(t) + b_i$$

라고 놓으면 (1)은 다시 다음과 같이 표현된다.

$$\frac{dx_i(t)}{dt} = \frac{1}{\tau_i}\left(\frac{1}{1+e^{-z}} - x_i(t) \right). \quad (2)$$

컴퓨터로 쉽게 다루기 위하여 이산시간영역으로 변환하면 결국 다음과 같은 식이 된다.

$$x_i[t+\Delta t] = x_i[t] + \frac{\Delta t}{\tau_i}\left(\frac{1}{1+e^{-z}} - x_i[t] \right) \quad (3)$$

(3)에서 $\frac{\Delta t}{\tau_i} \to 0$ 이면 (3)은 (2)로 근접해 간다. 실제로 $\tau_i \gg \Delta t$ 이므로 (3)을 근사치로 사용할 수 있다. (3)에서 보면 $g(z) \in (0,1)$이기 때문에

$$(g(z) - x_i[t]) \in (-x_i[t], 1 - x_i[t])$$

이 된다. 만약 $x_i[0]$이 $(0,1)$사이의 값을 가지면 $(g(z) - x_i[t]) \in (-1,1)$이 된다. 또한 (3)에서 만약 $x_i[t] = 0$ 이라고 가정한다면

$$x_i[t+\Delta t] \in \Delta t / \tau_i (1 + e^{-z})$$

이 되기 때문에 결국은 $x_i[t+\Delta t] \in (0, \Delta t/\tau_i)$가 된다. 유사하게 (3)에서 만약 $x_i[t] = 1$ 이라면

$$x_i[t+\Delta t] \in (1 - \Delta t/\tau_i, 1)$$

된다. 결국 만약 $x_i[0]$이 (0,1)사이의 값을 가지고 $\tau_i > \Delta t$ 이라면 $x_i[t+\Delta t]$는 0과 1사이의 값으로 한정된다.

III. 유전자 네트워크 모델의 파라미터 추정

유전자 네트워크 모델의 파라미터를 추정하기 위하여 SA법, 유전자 알고리즘, 하강 경사법, 선형 회귀 분석 등의 방법이 적용되어져 왔다 [4,5,8,9,11]. 본 논문에서는 최적화 문

제에 효과적으로 사용되어온 GA와 진화프로그래밍방법을 사용한다. Wahde는 본 논문의 접근법과 유사하게 논문 [8]에서 회귀신경망 유전자 네트워크 모델의 파라미터를 결정하기 위하여 GA를 사용하였다. 현재로서는 대부분의 유전자 네트워크가 규명이 되지 않은 상황이므로 유전자 발현 데이터로부터 유전자 네트워크를 구성해도 얼마나 잘 구성되었는지를 판단할 기준이 적절하지 않다. 그래서 Wahde는 인위적인 유전자 네트워크를 가정하여 이로부터 인위적인 유전자 발현데이터를 생성하고 이 데이터를 이용하여 유전자 네트워크를 재구성해봄으로써 가정하였던 인위적인 유전자 네트워크와 비교 평가하였다. 그러나 결과로 얻어진 유전자 네트워크 파라미터의 표준편차가 매우 커서 이로부터 직접 유전자 사이의 관계를 규명하기에는 불충분하였다. 이러한 결과는 추정해야할 파라미터의 개수에 비하여 주어진 데이터가 적은 것이 근본적인 원인이지만 그 외에도 GA가 비트 열로써 해를 표현하여 최적화되기 때문에 발생하는 문제일 것으로 여겨진다.

본 논문에서 우리는 이러한 문제를 고찰하기 위하여 GA와 더불어 진화프로그래밍 기법을 사용하여 유전자 네트워크의 파라미터를 추정해 본다. 4절의 실험결과에서 보듯이 진화프로그래밍이 GA 보다 더 좋은 결과를 보여줌을 알 수 있다. 또한 제안하는 GA가 Wahde [8]의 결과보다도 우수함을 알 수 있다. 그러나 Wahde의 논문에서 사용된 GA가 상세히 기술되어있지 않기 때문에 왜 본 논문에서 제안하는 GA가 더 좋은 성능을 보이는지를 정량적으로 비교 분석하기는 어렵다.

GA로 최적화 문제를 수행하려면 가장 먼저 대상 해를 염색체로 표현하는 코딩 작업이 필요하다. 본 논문에서는 다음과 같은 방법으로 코딩하였다. 만약 P개의 파라미터가 있고 각 파라미터를 K비트로 표현한다면 $P\times K$개의 비트로 염색체를 구성한다. 이 경우에 각 염색체는 파라미터가 가질 수 있는 실수 범위를 2^K 개로 분할한 값만을 가질 수 있다. 그러므로 파라미터 값을 보다 정교하게 찾아내기 위해서는 큰 K 값을 사용해야하나 K 값이 너무 크면 탐색 공간이 커져서 전역 해를 찾기 어렵고 또한 지역해에 빠질 가능성이 매우 커진다. 여기서 K를 1로 한다면 2개의 파라미터 값만을 가지기 때문에 부울 네트워크와 유사하게 모델링이 될 것이다.

표 1. GA의 파라미터.
Table 1. Parameters for the GA.

파라미터	값
교 확률 p_c	0.6
교배 포인트 개수	40
돌연변이 확률 p_m	0.01
개체군 개체 수	20
개체비트 길이	(24 * 30) bits
종료조건	0.001

본 논문에서는 K를 30으로 설정하여 실험하였다. 이 경우 하나의 염색체 길이가 매우 커져 (24개의 파라미터일 경우 24 * 30 = 720 비트) 단순한 교배 (crossover) 메커니즘으로는 염색체의 유전자가 잘 섞이지 못한다. 따라서 다중 포인트 교배 방법을 사용하였다. 본 논문에서 사용한 GA의 파라미터 값들은 표 1에 정리되어 있다.

종료조건은 개체 중 최고로 적합도가 높은 개체가 해에 근접한 정도를 따져서 결정한다. 본 논문에서는 주어진 유전자 발현 데이터와 GA가 찾은 가장 좋은 유전자 네트워크가 만드는 유전자 발현 데이터와의 평균자승오류 (MSE: mean square errors) 가 종료조건 보다 작은 경우에 종료한다. 평균자승오류는 다음과 같이 주어진다.

$$\delta_j = \frac{1}{TN}\sum_{i=1}^{N}\sum_{t=0}^{T}(x_i[t]-x_i^d[t])^2. \qquad (4)$$

(4)에서 N은 총 유전자의 개수, T는 총 샘플링 개수, $x_i[t]$는 GA로 구해진 유전자 네트워크가 생성한 i번째 유전자의 시간 t에서의 발현정도, $x_i^d[t]$는 주어진 유전자 발현 데이터이다. 총 유전자의 수가 N일때 결정해야할 총 파라미터의 개수는 유전자들 사이의 $W_{i,j}$ 가 $N\times N$개 이고 바이어스 및 시간 상수가 각각 N개이므로 총 $N(N+2)$이다. 총 데이터 개수가 $N\times T$이므로 사실상 합리적인 샘플링개수는 $T\rangle N+2$ 이어야 한다. 각 염색체의 적합도는 식 (4)에 주어진 MSE를 이용하여 다음과 같이 구한다.

$$f_j = \frac{1}{\delta_j+1} \qquad (5)$$

GA와 더불어 EP를 이용하여 파라미터를 결정하였다. 진화프로그래밍은 가능 해를 비트열로 표현하지 않고 가능해 그 자체의 값을 직접 진화시키는 방법이다. EP에서는 무작위 적으로 생성된 초기개체에 Gaussian 분포의 섭동(perturbation)을 더해주어 최적 해를 찾아간다. 단, 섭동은 무작위탐색(random search)이 되지 않도록 하기 위해서 최적해에 근접한 정도를 따져 최적해에 근접할수록 적게 섭동되도록 한다. 이러한 메커니즘은 GA의 돌연변이에 해당된다고 볼 수 있는데 진화프로그래밍에는 GA의 교배에 해당하는 연산은 없다. GA에서는 보통 룰렛 휠 방법으로 다음세대를 생성할 부모를 선택하는데 비하여 EP에서는 다른 개체와 특정 횟수에 걸쳐서 경합을 하고 이긴 숫자가 많은 순서로 정렬하여 결과가 좋은 순서대로 선택한다.

표 2. 진화프로그래밍의 파라미터.
Table 2. Parameters for the EP.

파라미터	값
개체군 개체수	100
하나의 개체	24개의 실수 값
경합 상대 수	40
종료조건	0.001

다른 개체와 경합시에 사용되는 것은 얼마나 그 개체가 최적해에 근접해 있나 하는 것 등이며 본 논문에서는 (4)의 MSE를 사용하였다. 그러므로 MSE가 작은 개체가 선택되는 것이다. 표 2는 진화프로그래밍에서 사용된 파라미터 값들을 보여준다.

IV. 실험 결과

본 논문에서는 먼저 인위적인 유전자 네트워크로부터 얻은 유전자 발현 데이터를 이용하여 유전자 네트워크를 구성한 결과를 설명하고 다음으로 실제 실험에서 주어진 유전자 발현 데이터로부터 유전자 네트워크를 구성하는 과정을 설명한다. 우리가 사용한 인위적인 유전자 네트워크는 Wahde[8]가 사용한 것과 동일한 것을 사용하였다. 표 3은 우리가 사용한 유전자 네트워크의 파라미터를 보여준다.

표 3의 파라미터를 갖는 유전자 네트워크가 생성한 인위적인 유전자 발현 데이터는 그림 1과 같다.

인위적인 유전자 발현 데이터로부터 GA와 진화프로그래밍 방법을 사용하여 유전자 네트워크의 파라미터를 구해 보았다. 그러나 서론에서 언급한 바와 같이 결정해야할 파라미터의 개수에 비하여 데이터의 개수가 많지 않고 에러 범위 안에서 동일한 유전자 발현을 보일 수 있는 네트워크가 많이 있을 수 있기 때문에 한 번의 실험만으로 파라미터를 찾을 수 없다. 그래서 Wahde의 논문에서와 같이 50번을 실험하여 평균 및 표준편차를 구하였다.

표 4를 보면 3개의 결과 모두 표준편차가 매우 큼을 볼 수 있다. 그러나 시간상수는 세 실험 모두에서 유사하게 구해졌고 전반적인 경향측면에서는 세 실험이 어느 정도 유사한 경향성을 보였다.

표 3. 인위적인 유전자 네트워크의 파라미터.
Table3. Parameters of the artificial gene network.

$W_{i,j}$				b_i	τ_i
20.0	5.0	0.0	0.0	0.0	10.0
25.0	-5.0	-17.0	0.0	-5.0	5.0
0.0	10.0	20.0	-20.0	-5.0	5.0
0.0	0.0	10.0	-5.0	0.0	15.0

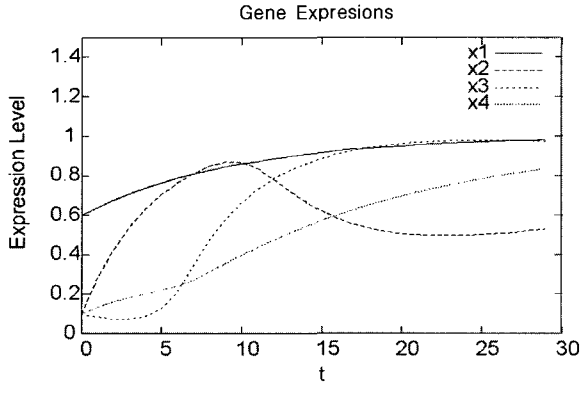


그림 1. 인위적인 유전자 발현 시계열 데이터.
Fig. 1. Artificial gene expression time-series data.

표 4. 실험결과 비교 (a) Wahde의 GA (b) GA (c) EP.
Table4. Comparison of experiments (a) Wahde's GA (b) GA (c) EP.

$W_{i,j}$				b_i	τ_i
11(11)	7.8(12)	7.5(12)	0.8(13)	3.7(4.8)	11(1.2)
14(7.7)	-16(7.3)	0.45(13)	-2.4(14)	1.8(4.6)	7.5(1.8)
5.9(14)	11(11)	6.1(13)	3.2(14)	1.9(5.8)	9.1(0.7)
3.0(11)	1.5(13)	5.3(11)	-16(7.0)	1.6(5.7)	16(5.3)

(a)

$W_{i,j}$				b_i	τ_i
17(9.3)	11(13)	6.7(15)	4.0(16)	3.1(5.5)	10(1.1)
20(7.8)	-13(5.3)	-11(10)	-9.4(12)	5.1(4.7)	5.2(0.5)
-12(9.3)	16(7.9)	12(9.5)	13(11)	-4.8(4.6)	5.8(0.6)
-4.5(13)	11(12)	14(11)	1.4(15)	-0.4(5.4)	17(1.8)

(b)

$W_{i,j}$				b_i	τ_i
14(13)	9.1(15)	7.9(14)	7.1(17)	3.7(6.1)	10(1.1)
24(6.1)	-15(4.2)	-18(8.2)	-8.2(10)	8.0(2.9)	5.1(0.3)
-8.7(8.4)	15(8.6)	12(10)	15.2(14)	-8.3(2.7)	5.1(0.6)
6.3(15)	12(14)	11(14)	4.4(15)	-1.3(7.3)	18(1.8)

(c)

위의 실험을 직접적으로 비교하는 것은 난해해 보인다. 이를 위해 하나의 지수를 도입하여 실험결과를 비교한다. 이 지수를 $\psi_{P,P'}$이라고 정의한다.

$$\psi_{P,P'} = \frac{1}{N(N+2)} \sum_{n=1}^{N(N+2)} \frac{1}{[1+\eta(\rho_{P_n}-\rho_{P'_n})^2]} \times \frac{1}{[1+(1-\eta)(\sigma_{P_n}-\sigma_{P'_n})^2]} \quad (6)$$

(6)에서 P는 인위적인 유전자 네트워크의 파라미터를, P'는 GA나 진화프로그래밍으로 구성된 유전자 네트워크의 파라미터를 각각 의미한다. 즉 두 개 네트워크의 파라미터를 비교하는 지수로서 각 파라미터별로 평균 ρ 와 표준편차 σ를 사용한다. 단, 실제로 인위적인 유전자 네트워크는 가정된 것이기 때문에 각 파라미터에 평균 ρ는 존재하나 표준편차 σ는 0이다. 수식 (6)에서 η는 두 항목 평균과 표준편차를 어느 비율로 적용할 것인지를 결정해 준다. $\eta=1$이면 평균 ρ만으로 두 네트워크의 파라미터를 비교하는 것이고 $\eta=0$이면 표준편차 σ만으로 비교하는 것이다. 만약 두 개의 네트워크 파라미터가 동일하면 η에 상관없이 $\psi_{P,P'}$는 최대 1이 된다. 본 논문에서는 $\eta=0.9$로 하여 결과를 산출했는데 그 이유는 표준편차보다는 평균이 더 중요한 요소이기 때문이다. 표 5는 산출결과를 보여준다.

표 5. 비교 지수.
Table 5. Comparison of the measure.

실험 ($\eta = 0.9$)	ψ
Wahde's GA [8]	0.042
GA	0.134
EP	0.138

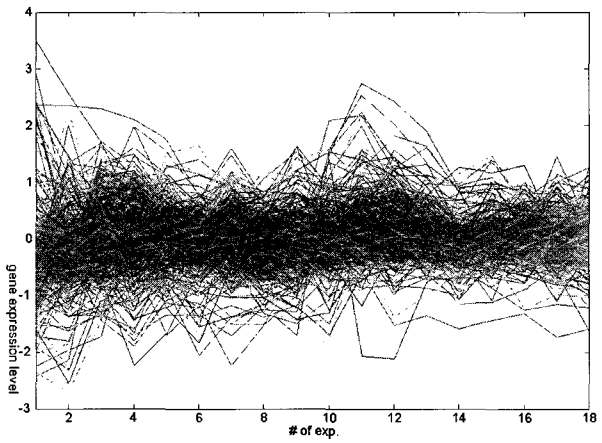


그림 2. Spellman의 효모 데이터(511x18).
Fig. 2. Yeast data profiles of Spellman data(511x18).

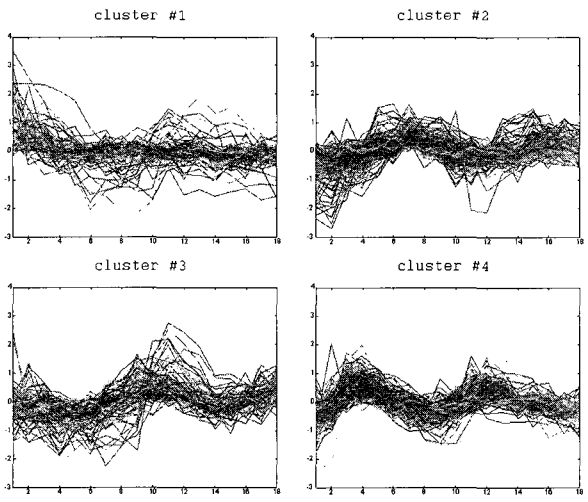


그림 3. 4개의 클러스터 데이터.
Fig. 3. Data profiles of four clusters.

표 5를 보면 본 논문에서의 GA 가 Wahde [8]의 결과보다 우수함을 알 수 있다. 그러나 앞서 지적한 바와 같이 Wahde의 논문에서는 사용된 GA에 대한 상세한 기술이 결여되어 있어서 (즉 사용된 파라미터 값들) 결과의 차이에 대한 정량적인 비교 분석이 현재로서는 어렵다. GA의 경우 문제를 코딩하는 방법, 교배방법 및 교배와 돌연변이 확률 등에 따라 결과가 많은 영향을 받기 때문이다. 그러나 이들에 비하여 본 논문에서 처음 적용한 진화프로그래밍 방법은 더욱 우수한 결과를 보임을 알 수 있다.

또한 진화프로그래밍 방법이 GA보다 매우 빠르게 최적 해에 근접하므로 효율성 측면에서도 더 우수함을 알 수 있었다.

다음으로 실제 유전자 발현 데이터를 이용하여 유전자 네트워크를 구성하는 시도를 하였다. 우리가 사용한 데이터는 Spellman[20]의 데이터로서 많은 연구자들이 사용하는 데이터이다. Spellman의 데이터는 6178행 82열의 데이터로서 이 중에 세포분열과 관련된 유전자로 판명된 511개의 유전자를 먼저 추출하였고 여러 실험 중에 하나인 알파실험결과 18개를 이용하였다. 그림 2는 추출된 511x18 데이터를 모두 출력한 것이다.

먼저 유전자의 개수가 많기 때문에 이를 K-means 클러스터링 방법을 사용하여 4개의 유전자 그룹으로 클러스터링하였다. 이렇게 하여 511x18 데이터를 4x18데이터로 만들었다. 본 논문에서 전체 클러스터의 개수를 4개로 제한한 것은 하나의 예를 보여준 것이며 보다 의미 있는 클러스터링을 하기 위해서는 생물학적으로 유의한 클러스터의 개수를 먼저 선정해야 한다. 또한 유전자 발현데이터의 유사성에만 의존한 기계적인 클러스터링보다는 생물학적으로 의미가 밝혀진 정보가 있을 경우 그 정보를 추가적으로 이용하여 클러스터링 하는 것이 필요하다. 예를 들어 이미 기능이 유사한 것으로 밝혀진 유전자들은 유전자 발현 데이터의 양상을 비교하기 이전에 동일한 클러스터로 할당하는 등의 사전처리과정이 필요하다. 그림 3은 각각의 클러스터링 별로 데이터를 출력한 것이다. 4개의 클러스터 각각에 속한 유전자 발현 데이터를 평균하여 대표 시계열 데이터를 생성한 다음 회귀신경망 모델을 적용하기 위하여 다시 0과 1사이로 크기를 재조정(normalize)하였다. 이렇게 구해진 4x18 데이터는 4개의 유전자 그룹을 대표하는 값으로서 이것들을 이용하여 각 유전자 클러스터간의 네트워크를 구성한다. 우리는 이전 실험과 유사하게 GA와 진화프로그래밍을 사용하여 모델링하였다. GA를 사용하여 유전자 네트워크 파라미터를 구한 결과 중 하나의 유전자 네트워크를 이용하여 유전자 발현 데이터를 구해 보았다. 그림 4는 4개의 유전자 클러스터들의 평균값과 구해진 유전자 네트워크가 생성한 유전자 발현 데이터를 함께 그려 놓은 것이다 (실선은 유전자 클러스터의 평균값, 점선은 구해진 유전자 네트워크로부터 생성된 데이터를 각각 나타냄).

이 실험에서는 모두 100번의 실험을 수행하여 평균 및 표준편차를 구하였다. 각 파라미터별 허용 최소, 최대값은 W와 b를 최소 -5에서 +5로 τ를 최소 4에서 최대 21로 설정하였다. 표 6은 GA와 진화프로그래밍으로 구한 결과 값들을 정리한 것이다.

표 6으로부터 비교적 표준편차가 작은 파라미터가 많음을 알 수 있다. 표준편차가 작다는 것은 표준편차가 큰 것보다 상대적으로 신뢰도가 높은 결과임을 시사한다. GA와 진화프로그래밍에서 거의 동일하게 나타나는 파라미터들도 많이 보인다. 이것 역시 이전의 인위적인 데이터를 이용한 실험보다는 신뢰도가 높은 의미 있는 결과로 해석된다. 그러나 여전히 표준편차가 비교적 큰 부분이 많이 존재한다. 그렇다하더라도 준 근사적으로 구해진 유전자 네트워크를 이용하여 유전자별 상호영향을 분석할 수 있고 이를 이용하

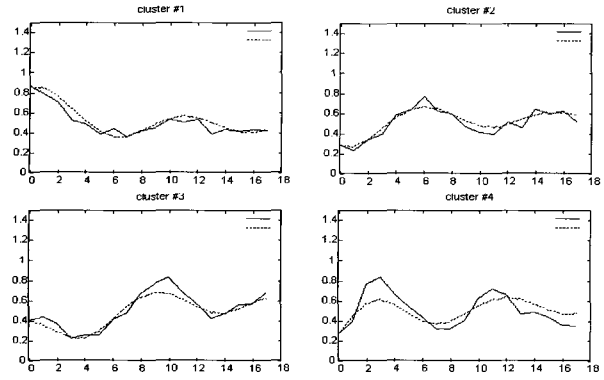


그림 4. 4개의 클러스터 데이터 프로파일 및 생성된 데이터.
Fig. 4. Four cluster data profiles and generated data.

표 6. 실험결과의 비교 : (a) GA (b) EP.
Table6. Comparison of experiments : (a) GA (b) EP.

$W_{i,j}$				b_i	τ_i
-3.3(1.7)	2.7(2.4)	4.4(1.2)	-3.3(2.0)	-1.2(1.8)	6.6(1.4)
1.5(2.8)	-3.5(1.6)	-4.5(1.1)	4.0(1.7)	2.0(1.6)	7.3(2.1)
-4.6(1.2)	3.9(1.2)	-1.1(1.4)	0.1(2.8)	0.9(2.1)	4.3(0.7)
4.7(0.7)	-4.4(1.1)	-0.4(2.0)	-4.1(1.1)	2.3(1.4)	4.0(0.0)

(a)

$W_{i,j}$				b_i	τ_i
-3.3(1.9)	-0.8(3.6)	4.0(1.1)	-4.6(0.9)	1.8(2.6)	4.6(0.8)
-3.6(2.3)	-3.4(1.9)	-4.8(0.5)	4.8(0.4)	3.7(2.0)	4.2(0.5)
-4.7(0.6)	4.7(0.5)	-0.7(1.9)	-1.8(3.0)	1.2(2.4)	4.2(0.4)
4.7(0.6)	-4.8(0.4)	-2.0(2.2)	-4.5(0.8)	3.7(1.4)	4.0(0.1)

(b)

여 보다 낮은 실험 설계를 할 수 있으며 또한 새로운 실험을 통하여 그 결과를 검증해 볼 수 있다는 측면에서 유용한 결과로 판단된다. 보다 정확한 네트워크 구성을 하기 위해서는 다음과 같은 작업이 추가적으로 필요할 것으로 생각된다. 첫 번째로, 동일한 조직세포에 대하여 독립적인 유전자 발현 데이터를 많이 확보하는 것이다. 그렇게 되면 파라미터 공간 검색에 있어서 보다 효과적인 해의 평가가 가능해지고 이로 인해 표준편차가 보다 작은 결과를 도출할 수 있을 것으로 기대된다. 두 번째로 다른 조직의 동일세포를 이용하여 유전자 발현 데이터를 얻고 이를 실험에 이용하는 것이다. 그러나 이러한 작업의 결과로도 만족할 만한 표준편차로 줄어들지 않는 경우도 있을 것이다. 이는 알고리즘적으로 보완할 수 있을 것으로 예견된다. 그러한 방법으로 표준편차가 0에 가까운 (예로, 0.8보다 작은 것) 파라미터를 그 파라미터의 평균으로 고정하고 다시 100번의 실험을 통하여 평균과 표준편차를 정하는 방법이다. 새로운 실험의 결과로 얻어진 표준편차가 0에 가까운 것이 생성되면 위 과정을 반복할 수 있을 것이다. 이를 반복하면 모든 파라미터에서 표준편차가 0인 값을 얻을 수도 있을 것이다. 또한 최종 결과에서 0에 가까운 평균을 0으로 놓고 위 실험을 반복할 수도 있다. 그러나 다른 측면으로는 이러한 알고리즘이 오류의 누적을 가져올 수도 있다. 그러므로 보다 확실한 실험적, 이론적 근거를 정립한 뒤 접근해야만 한다.

V. 결론

본 논문에서는 진화연산 알고리즘으로 유전자 발현 시계열 데이터로부터 유전자 네트워크를 구성하는 방법을 제안하고 실험결과를 제시하였다. 먼저 방법의 정당성을 확보하기 위하여 인위적으로 주어진 가상의 유전자 네트워크로부터 만들어진 유전자 발현 시계열 데이터를 이용하여 유전자 네트워크 구성을 확인해 보았다. GA를 사용한 방법이나 진화프로그래밍을 사용한 방법이 대부분 실험별로 큰 표준편차를 보여 제한된 데이터로는 근본적으로 정확하게 추정해 내는 것이 어렵다는 것을 확인하였다. 그러나 각 유전자 사이의 관계가 양의 관계인지 음의 관계인지 그 크기가 대소인지 정도는 분석 할 수 있는 결과를 보였다 (그리고 이러한 정보가 생물학적으로도 상호작용을 판별하는데 중요하다). 실제의 실험데이터를 클러스터링 하여 적용한 결과 이전 실험보다는 작은 표준편차를 보였다. 또한 GA와 진화프로그래밍의 결과가 많은 부분에서 거의 동일하게 산출됨을 확인하였다. 이러한 결과는 본 논문에서 제안된 방법이 어느 정도 의미 있는 것으로서 유전자 발현데이터를 추가하거나 보완하면 더욱 좋은 결과를 도출해 낼 수 있는 가능성을 보였다고 할 수 있다. 향후 이러한 실험을 통하여 본 방법의 유용성을 확인하고 더불어 보다 정교한 알고리즘으로 확대 발전시키는 것이 필요하다.

참고문헌

[1] J. Reinitz and D. H. Sharp, "Mechanism of eve stripe formation," *Mechanisms of Development,* no. 49, pp. 133-158, 1995.

[2] M. B. Eisen, P. T. Spellman, P. O. Brown, and D.~Botstein, "Cluster analysis and display of genome-wide expression patterns," *in Proceedings of the National Academy of Sciences,* pp. 14863-14868, Dec. 1998.

[3] T. Akutsu, S. Miyano, and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model," *in Pacific Symposium on Biocomputing,* pp. 17-28, Jan. 1999.

[4] D. Weaver, C. Workman, and G. Stormo, "Modeling regulatory networks with weight matrices," *in Pacific Symposium on Biocomputing,* pp. 112-123, Jan. 1999.

[5] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear modelling of mRNA expression levels during CNS development and injury," *in Pacific Symposium on Biocomputing,* pp. 41-52, Jan. 1999.

[6] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Journal of Computational Biology,* no. 7, pp. 601-620, 2000.

[7] P. D'haeseleer, "Reconstructing gene networks from large scale gene expression data". PhD thesis, University of New Mexico, Albuquerque, 2000.

[8] M. Wahde and J. Hertz, "Coarse-grained reverse engineering of genetic regulatory networks," *BioSystems,* no. 55, pp. 129-136, 2000.

[9] M. Wahde and J. Hertz, "Modeling genetic regulatory dynamics in neural development," *Journal of Computational Biology,* no. 4, pp. 429-442, 2001.

[10] E. van Someren, L. Wessels, and M. Reinders, "Linear modelling of genetic networks from experimental data," *in Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology,* pp. 355-366, Aug. 2000.

[11] E. van Someren, L. Wessels, and M. Reinders, "Genetic network models: A comparative study," *in Proceedings of SPIE, Micro-arrays: Optical Technologies and Informatics,* pp. 236-247, Jan. 2001.

[12] L. Wessels, E. van Someren, and M. Reinders, "A comparison of genetic network models," *in Pacific Symposium on Biocomputing,* pp. 508-519, Jan. 2001.

[13] M. Wahde and J. Hertz, "Modeling genetic regulatory dynamics in neural development," *Journal of Computational Biology,* vol. 8, pp. 429-442, Sep 2001.

[14] M. Takane, "Inference of Gene Regulatory Networks from Large Scale Gene Expression Data," Master's thesis, McGill University, 2003.

[15] T. Chen and G. C. H. He, "Modeling gene expression with differential equations," *in Pacific Symposium on Biocomputing,* pp. 29-40, Jan. 1999.

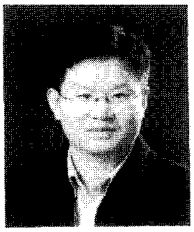
[16] J. Holland, Adaptation in natural and artificial systems. 1st ed.: University of Michigan Press, Ann Arbor; 2nd ed.: 1992, MIT Press, Cambridge, 1975.

[17] D. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison-Wesley, 1989.

[18] L. Davis, Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold, 1991. L. Davis (Ed.).

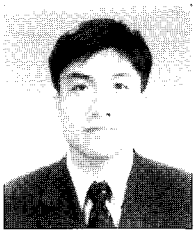
[19] D. B. Fogel, "An Introduction to Simulated Evolutionary Optimization," *IEEE Trans. on Neural Networks,* vol. 5, pp. 3-14, Jan. 1994.

[20] 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 Hibridization}," *Molecular Biology of the Cell,* vol. 9, pp. 3273-3297, 1998.



정 성 훈

1988년 한양대 전자공학과 졸업. 한국과학기술원 석사(1991), 동대학원 박사(1995) 및 박사후 연구원 (1996). 1996년~현재 한성대 정보공학부 전임강사, 조교수 및 부교수. 관심분야는 진화연산, 퍼지, 신경망 및 지능시스템 분야로 특히 지능시스템의 컴퓨터 게임에의 응용에 관심이 있으며 최근 생물정보학 및 시스템생물학에 많은 관심을 기울임.



조 광 현

1993년 KAIST 전기및전자공학과 졸업. 동대학원 석사(1995), 동대학원 박사(1998) 및 박사후 연구원(1998). 1999년~2004년 8월 울산대학교 전기전자정보시스템공학부 전임강사 및 조교수. 2002년 6월~2003년 8월 영국 UMIST 제어시스템센터 초빙연구원. 2003년 12월~2004년 3월 스웨덴 Royal Institute of Technology 자동제어그룹 초빙교수. 2004년 6월~2004년 8월 아일랜드 Hamilton Institute 초빙교수. 2004년 9월~현재 서울대학교 의과대학 의학과 (생명공학전공) 조교수 및 서울대학교 생명공학공동연구원(Korea Bio-MAX Institute, 관악캠퍼스 소재) 시스템생물학 연구실 책임연구원 겸임. 2004년~2005년 국제저널 'Systems Biology (IEE, 영국런던)'의 Editor-in-Chief 위임. 관심분야는 제어공학의 생명과학 응용 (시스템생물학) 및 생명과학에서 발견되는 자연현상의 공학적 응용.