지역정렬을 이용한 유전자 발현 조절 프로그램 예측

이지연° 진희정 조환규
정보컴퓨터공학과, 부산대학교
{jylee°, hjjin, hgcho}@pusan.ac.kr

Inference of Gene Regulatory Program using Local Alignment

Ji-Yeon Lee° Hee-Jeong Jin Hwan-Gue Cho
Dept. of Computer Science, Pusan National University

요 약

세포의 활동은 단순히 하나의 유전자의 발현으로 설명되기보다 여러 유전자와 그로 인해 생성된 단백질의 상호 작용에 의해 나타난다. 또한 마이크로어레이 실험을 통해 세포 내의 유전자 발현에 대한 정보를 알 수 있게 되고, Chromatin IP 마이크로어레이 실험을 통해 신뢰도가 높은 유전자 발현 조절 관계 데이터를 얻을 수 있게 되면서, 유사한 기능과 유사한 발현 패턴을 보이는 유전자들을 그룹으로 묶어 유전자 모듈로 규정하고 이를 하나의 유전자 조절 네트워크로 구성하고, 분석하는 연구들이 진행되고 있다. 본 논문에서는 ChIP 실험 데이터와 유전자 발현 데이터를 이용하여 지역 정렬을 수행해 하나의 유전자 모듈을 조절하는 조절 프로그램을 예측하는 알고리즘에 대해 기술한다. 조절 프로그램은 유전자 조절 모듈을 조절하는 조절자들의 역할 및 발현 여부에 따른 유전자 조절 모듈 내 유전자들의 발현을 설명할 수 있는 것이다.

1. 서 론

마이크로어레이 실험 기술의 발전은 우리에게 세포 내에서 일어나는 여러 유전자들의 활동에 대한 대량의 정보를 얻을 수 있게 해 주었다. 이 실험을 통한 대량의 유전자 발현 데이터를 이용해 클러스터링, 유전자 발현 조절 관계 추론 등의 다양한 연구를 진행해왔다. 하지만 유전자 발현 조절 관계 추론에 있어 유전자 발현 데이터만 이용하는 것은 데이터의 크기가 방대하고, 데이터 노이즈 및 미싱 값 등으로 인해 false positive 결과가 많다는 단점이 존재한다. 그래서 최근의 연구는 유전자 발현 데이터에 PPI(Protein-Protein Interaction) 데이터, genome-wide location 데이터 및 MIPS[1] 데이터와 같은 부가적인 정보를 더해 발현 조절 관계를 추론하고 있다. 직접적으로 유전자 발현 조절 관계를 밝혀낼 수 있는 실험에 Chromatin IP 마이크로어레이 실험이 있다. 이 실험을 통해 DNA 프로모터와 상호작용하는 전사인자를 찾을 수 있다.

대량의 데이터에서 발현 조절 관계를 추론하는데 있어 최근의 연구는 좀 더 전체적인 관점에서 이루어지고 있다. 이러한 경향은 세포의 활동은 단순히 하나의 유전자의 발현으로 설명되기 보다는 여러 유전자와 그로 인해 생성된 단백질 및 물질 대사에 의한 상호 작용으로 나타나기 때문이다. 그래서 단순히 유전자 하나와 조절자 사이의 조절 관계를 추론하기 보다는 유사한 발현 패턴을 보이며 기능적으로 유사한 유전자들을 그룹으로 묶어 그들 사이의 상호작용을 연구하는 것이 요즘의 추세이다. 공통의 전사인자 집합에 의해 발현이 조절되고 특정 환경에서 함께 발현하는 유전자들의 집합을 유전자 모듈(gene module) 혹은 유전자 조절 모듈(gene regulatory module)이라고 한다.

본 논문에서는 하나의 유전자 모듈에 대하여 *cis*-regulators가 어떤 조건에서 조절 역할을 하는지를 나타내는 조절 프로그램을 예측하는 알고리즘을 제시한다. 조절 프로그램은 특정 환경 및 실험 조건에서 유전자들의 발현에 대한 가설을 제시할 수 있다는 의의가 있다. 유전자 모듈의 조절자 집합은 ChIP 마이크로어레이 실험을 통해 밝혀진 조절 관계 데이터를 사용하였다.

2. 관련 연구

Cis-regulation이란 하나의 유전자를 조절하기 위한 하나의 프로모터에 전해지는 조절 신호(regulato2ry signals)의 조합을 말한다. 최근의 조절 관계 예측 알고리즘의 흐름은 유사한 발현 패턴을 가지는 유전자들을 하나의 클러스터로 하여 이들의 공통 *cis*-regulatory 요소들을 찾는 것이다. 이들 연구 중 Segal *et al.*[2]의 연구 결과는 환경 및 조건에 특화하여 조절자들의 조절이 분화되는 시점까지 예측한 조절 프로그램(regulation prog- ram)을 제시하고 있어 매우 흥미롭다. 이 논문에서는 유전자 발현 데이터, 후보 조절자 집합, GO 등과 같은 유전자 주석(annotation) 데이터, genome-wide location 데이터 등을 이용하여 유전자들을 유사한 발현 패턴이 나타나는 모듈로 분류하고, 이 모듈의 발현 특성을 반영할 수 있는 조절 프로그램을 학습, 주석 데이터와 프로모터 분석 등을 함께 제공하는 시스템을 소개했다. 이는 전체적인 유전자 조절 모듈 네트워크와 그 생물학적 의미를 쉽게 살펴볼 수 있도록 한 데 의의가 있다.

Segal의 시스템은 먼저 발현 패턴이 유사한 유전자들끼리 클러스터링하여 각각의 모듈에 맞는 조절 프로그램을 학습하여 조절자의 조절 분기 시점, 역할 등을 예측하고 있다. 조절 프로그램에 의해 분류된 조절 모듈들이 생물학적으로 어떠한 의미가 있는지 보여주기 위하여 GO 기

능 데이터 등을 이용하여 중요한 주석을 표시해주고 promotor 분석 정보를 보여준다.

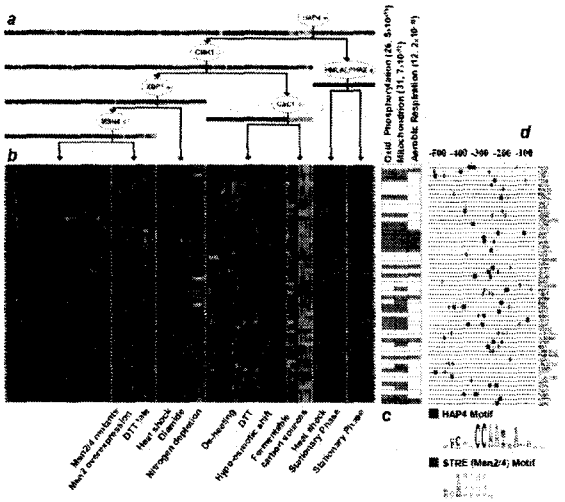

그림 1 조절 프로그램의 예시[2]

Segal의 조절 프로그램은 다른 논문에서와는 달리 조절자의 조절 분기를 이진 트리(binary tree) 형태로 표현하여 조절 프로그램을 직관적으로 이해하기 쉽도록 나타내었다. 그림 1은 조절 프로그램을 주석 데이터와 프로모터 분석 등의 데이터와 함께 보여주는 예이며, 조절 모듈과 이 모듈을 조절하는 조절자들의 조절 프로그램, 중요한 주석 정보, transcription factor binding site 분석으로 구성되어 있다. 그림 1의 상단 트리 부분이 모듈을 조절하는 조절 프로그램이고 하단은 이 조절 프로그램의 단말 노드(leaf node)들로 모듈 내 유전자의 발현 프로파일을 조절 프로그램에 따라 나눈 것이다. 유전자 발현 프로파일의 가로줄은 유전자를 나타내고, 세로줄은 특정 시간 포인트(time-point)와 어레이들을 나타낸다. 조절 프로그램은 의사결정 트리 형태로 결정 노드와 단말 노드로 구성된다. 각각의 결정 노드는 하나의 조절자와 조절자의 발현값에 대한 질의어로 구성된다. 질의어에는 up-regulation, no change, down-regulation이 있다. 결정 노드는 2개의 자식 노드들을 가지는데, 결정 노드의 질의어 값이 참일 때는 오른쪽 노드가 선택되고, 거짓일 때는 왼쪽 노드가 선택된다. 하지만, 결정 노드와 자식 결정 노드 사이의 발현 인과관계는 성립하지 않는다.

Segal이 제시한 발현 데이터를 이용하여 조절 프로그램을 학습하는 알고리즘은 베이지안 점수(bayesian score)를 유사도 점수로 사용하여 EM(xpectation Maximum) algorithm을 통해 조절 프로그램을 최적화시킨다. 베이지안 네트워크는 잘 설립된 이론적 기반과 통계적 견고성로 인해 조절 관계 추론에 매우 널리 쓰이고 있다. EM 알고리즘은 유사도 점수가 local maximum에 이를 때까지 증가하는 것을 보장해 준다는 중요한 특성이 있다.

조절 프로그램 학습은 E-step과 M-step이 반복적으로 수행하여 local maximum에 도달할 때까지 이루어진다. M-step에서 모듈의 발현 패턴을 잘 반영하는 조절 프로그램을 학습하고, E-step에서는 모듈의 특성을 가장 잘 반영하는 조절 프로그램과 관계된 모듈을 결정해 모듈에 다시 유전자를 대입하는 작업을 한다. 의사 결정 트리는 루트 노드로부터 단말 노드들로 확장되는데, 더 이상 모듈이 나누어지지 않을 때까지 생성된다. 이와 같은 일련의 과정들을 통해 각 조절 모듈에 대한 조절 프로그램을 학습하게 된다.

본 논문에서는 Segal이 제안한 조절 프로그램에 비하여, 실험 조건에 따른 발현 시점 분석이 더욱 쉽고, 빠른 시간 내에 구성할 수 있는 알고리즘을 제안한다.

3. 조절 프로그램의 입력 데이터

본 연구에 사용된 데이터는 공개된 효모(yeast) 유전자 발현 데이터와 ChIP 마이크로어레이 실험 데이터이다. 사용된 유전자 발현 데이터는 Gasch *et al.*이 실험한 효모의 환경 및 자극에 대한 실험 데이터이다[4]. Gasch의 데이터는 전체 6,153개의 유전자와 173개의 array로 구성되어 있으며, 여러 가지 실험 환경에 따라 작업한 것으로 일정한 시간 간격으로 측정된 것도 있고 그렇지 않은 것도 있다. 이 데이터는 온도 변화, 아미노산 제거, 질소 공급 차단 등 여러 가지 환경 변화에 따른 유전자 발현 정도를 측정한 것으로 연속적으로 일정한 시간차를 두고 실험한 것은 아니다. 조건에 따라 짧은 것은 5분에서 긴 것은 하루 정도의 시간차를 두고 발현량을 측정하였다.

유전자 조절 관계를 얻기 위한 ChIP 마이크로 어레이 실험 데이터는 Lee *et al.*[3]이 2002년 발표한 것으로 106개의 전사인자에 대하여 실험한 것이다. 데이터의 신뢰도를 위해 P-value를 측정하였는데, P-value가 0.001 이하에서 3985개의 조절 관계를 찾을 수 있었다. P-value를 낮추게 되면 false negative 데이터가 증가하는 단점이 있지만 false positive의 수를 줄이기 위하여 낮은 P-value cut-off를 사용한 데이터를 이용한다.

4. 유전자 모듈 조절 프로그램 예측 알고리즘

그림 2는 본 논문에서 제안하는 조절 프로그램 예측 알고리즘의 전체적인 구성을 나타낸다. DNA-바인딩 데이터에서 모듈의 후보 조절자 집합을 추출하고, 조절자와 모듈의 평균 발현 데이터를 상진화하여 지역 정렬(local alignment)을 수행해 조절 프로그램을 학습한다. 유전자 조절 모듈의 유전자들을 구하는 것은 조절 프로그램을 예측하는 범위 밖의 문제이므로 본 논문에서는 다루지 않는다.

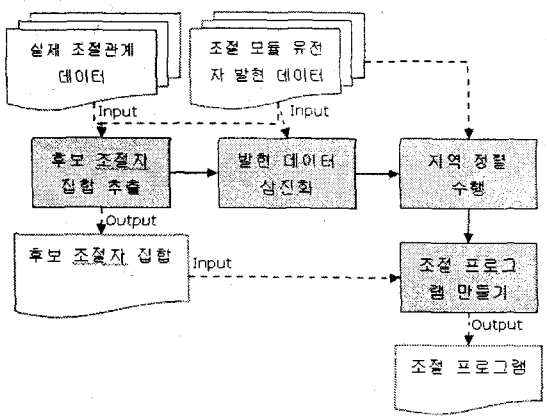

그림 2 조절 프로그램 예측 알고리즘 흐름도

조절 프로그램의 입력은 실제 조절 관계 쌍의 집합과 유전자 발현 데이터이다. 먼저 유전자 조절 모듈의 후보 조절자 집합을 실제 조절 관계 데이터로부터 찾고 후보 조절자와 유전자 모듈의 발현 데이터를 정렬을 수행하기 위해 삼진화한 후 지역 정렬을 수행한다. 정렬 수행 결과 점수에 따라 조절 프로그램을 구성하게 되는데, 정렬 점수가 높은 조절자가 유전자 모듈의 발현과 유사도가 높은 것이므로, 이를 기준으로 조절 프로그램을 구성하게 된다. 알고리즘 1은 조절 프로그램의 수도 코드이다.

```
Input:
  E          //조절 모듈의 유전자 발현 데이터 집합
  R          //조절자 집합
Output:
  T          //Tree형태의 조절 프로그램
PNURegProgram
  //조절 모듈 유전자 발현 데이터의 평균값 구하기
  meanOfModule = GetMeanOfModule(E)
  //모듈 유전자 발현 데이터 삼진화
  expStringOfModule =
                    GetExpString(meanOfModule)
  //후보 조절자 집합 구하기
  CR = SelectCandidateRegulators(R)

  //후보 조절자 집합 조절자들과 조절 모듈의 평균
  // 데이터와 정렬 수행
  For every regulator in CR
    expStringOfRegulator = getExpString(regulator)
    For every experiment //실험 조건에 따라 정렬 수행
      Align(expStringOfRegulator, expStringOfModule)
    end For
  end For
  //
  MakeRootNode(T)
```

```
FindChildNodes(root of T, E, T)
```

알고리즘 1. 조절 프로그램의 알고리즘

4.1 후보 조절자 집합의 선택

후보 조절자 집합을 선택하기 위해 DNA-바인딩 데이터를 살펴보았다. Segal이 제시한 유전자 조절 모듈에는 평균적으로 50개의 유전자가 존재한다. 일단 유전자 조절 모듈을 조절하는 후보 조절자 집합을 선택하기 위해 유전자 조절 모듈 내 유전자들과 조절 관계가 존재하는 조절자 집합을 구해보았다. 평균적으로 조절 모듈 하나당 3~40개의 조절자들이 검색되었다. 이 중에서 유전자 조절 모듈 내 유전자들 중 2개 이상의 유전자들과 상호작용하는 조절자의 수는 적게는 4개에서 많게는 20여개가 존재하였다. 표 1은 Segal이 제시한 유전자 조절 모듈 중 2가지 모듈을 나타낸 것이다.

유전자 모듈 정보	조절자 후보 집합
Module 1. · 조건- Respiration and carbon regulation · 유전자 수- 55개	YKL109W(26), YGL237C(11), YBL021C(11), YOR372C(3), YKL112W(3), YGL209W(2), YDR259C(2), YDR421W(2), YDR501W(2), YGL073W(2), YGL035C(2), YBR112C(2), YBR049C(2), YOR028C(1), YML027W(1), YKL043W(1), YBR297W(1), YMR016C(1), YNL068C(1), YDR216W(1), YOR358W(1), YMR043W(1), YBR289W(1), YCR084C(1), YDR176W(1), YDR448W(1), YHL025W(1), YJL176C(1), YMR070W(1), YOR140W(1), YOR290C(1), YPL016W(1), YLR403W(1), YHR206W(1), YDR310C(1), YEL009C(1), YGL192W(1), YPR065W(1), YCR065W(1), YGL096W(1), YKR099W(1), YKL062W(1), YLR256W(1)
Modlue14 · 조건- Ribisonal and phosphate metabolism · 유전자수 - 32개	YKL112W(5), YOR372C(3), YPR104C(3), YER111C(3), YBR289W(2), YDL106C(2), YFR034C(2), YHL025W(2), YOR290C(2), YMR016C(2), YMR043W(2), YNL068C(2), YOL108C(1), YIL131C(1), YLR098C(1), YEL009C(1), YKL109W(1), YDR123C(1), YJL056C(1), YIR023W(1), YHR084W(1), YGL025C(1), YLR014C(1), YOL004W(1), YCR065W(1), YDR501W(1), YDR310C(1), YOR028C(1), YKL032C(1), YDR043C(1), YPL089C(1), YPR065W(1), YHR206W(1), YDR259C(1), YDR451C(1), YML027W(1), YLR182W(1)

표 1 후보 조절자 집합

표 1에서 보는 것과 같이 유전자 조절 모듈의 유전자들과 조절 관계가 존재하는 조절자의 수가 매우 많다. 이 중에서 후보 조절자를 선택하는 것도 중요한 문제이다. 현재는 모듈 내 유전자들 중 적어도 셋 이상의 유전자들과 조절 관계가 존재하는 유전자들로 그 조건을 제약하였다.

4.2 발현 데이터의 삼진화

후보 조절자들이 정해지면 실험 조건에 따라 조절 모듈의 유전자 발현 데이터와 후보 조절자들의 발현 데이터를 삼진화하여 정렬을 수행한다. 정렬을 수행하기 위해 발현 데이터를 삼진화한 기준은 발현값의 정도이다. 발현이 활성화되어 있으면 U로 표시하고 발현이 나타나지 않은 경우는 C, 발현이 억제된 경우는 D로 표시한다. 발현이 활성화되었다고 판단하기 위한 임계값은 1.0으로 설정하였다. 현재는 상수 값으로 두고 실험을 진행하고 있지만, 이 값은 실험에 따라 달라질 수 있을 것이다.

결과를 좀 더 민감하게 보기 위해서는 임계값을 낮출 필요가 있다. 그리고 발현이 특별히 많이 일어나거나 특별히 억제된 경우만을 보기 위해서는 임계값을 높일 수도 있다. 다만 임계값을 낮추게 되면 조절 프로그램의 트리 구조가 복잡해질 수 있다. 반대로 임계값을 높이게 되면 트리 구조는 단순해지게 되지만 조절 가능성이 있는 조절자들이 조절 프로그램에서 제외되는 경우가 발생할 가능성 또한 높아지게 된다.

기호	의미
U	Upregulated. 발현이 활성화된 경우
C	Constant. 발현량의 변화가 없는 경우
D	Downregulated 발현이 억제된 경우
M	Missing value. 발현값이 존재하지 않는 경우

표 2 유전자 발현 데이터의 삼진화

4.3 조절자와 유전자 모듈 발현 데이터 지역 정렬 수행

발현 데이터를 삼진화하고 나면 실험 조건별로 정렬을 수행하여 그 결과 점수를 모두 합한다. 정렬에 사용되는 scoring matrix는 다음 표 3과 같다. Scoring matrix의 α와 β는 모두 양수이고, D-D의 경우 유전자의 발현이 억제된 것이 전적으로 활성자가 발현 억제된 것에 영향을 받았다고 보기 힘들므로 R-R보다 적은 가중치를 부여한다(0<a<1). 정렬 점수가 높을수록 조절자는 유전자 모듈의 활성자일 확률이 높아진다. 억제자의 경우, 유전자 모듈의 발현 데이터 삼진화 결과를 역으로 바꾸어 정렬을 수행하여 예측한다. 여기서 사용하는 정렬은 갭을 허용하지 않는다. 이는 시간 간격으로 측정되지 않은 데이터가 다수 존재고, 시간 간격을 두고 측정한 데이터들도 그 간격이 일정하지 않기 때문이다.

	U	C	D	M
U	α	0	-β	0
C	0	0	0	0
D	-β	0	a · α	0
M	0	0	0	0

표 3 Scoring matrix

다음은 Segal의 유전자 모듈 1의 발현 데이터의 평균과 조절자 HAP4의 발현 데이터를 지역 정렬을 수행한 결과를 실험 조건 별로 나열한 것이다. 실험 조건에서 서로 매칭되는 것이 임계값 이상일 때 그 실험 조건 하에서는 조절자에게 조절 영향을 받는다고 가정한다.표 4는 정렬 결과 70% 이상이 매치되었을 때 조절된다고 판단하였다.

실험	매칭 결과
Heat Shock 1	UUUUmmcm
Heat Shock 2	DDDcccd
Temperature shift from 37 to 25	ccccc
Heat Shock 3	UUUUc
Heat Shock 4	cccc
Heat Shock at variable osmolarity	ccdccc
Hydrogen peroxide treatment	ccccccccc
Menadione Exposure	cccccccuc
DTT exposure 1	ccccccc
DTT exposure 2	cccdcc
Diamide treatment	ucccccdcc
Hyper-osmotic shock	ccducc
Hypo-osmotic shock	ccccc
Amino acid starvation	ccccc
Nitrogen source depletion	dcccdUcccc
Diauxic shift	ccucdUU
Stationary phase 1	cuUUUUUUUU
Stationary phase 2	UuuUUUUUUUUU
HM	ccccccccU
PM	ccc
Over expression	UcDDcD
CS	UDcDDUD
Steady-state growth ct-1	ccccc
Steady-state growth ct-2	cccccccc

표 4 조절자 HAP4와 유전자 조절 모듈의 정렬 결과

4.4 조절 프로그램 만들기

조절 프로그램의 형태는 바이너리 트리형태의 의사 결정 트리이다. 루트 노드는 정렬 점수가 가장 높은 것이 선택된다. 조절 프로그램의 트리는 의사 노드와 단말 노드로 구성되는데, 의사 노드는 조절 프로그램을 조절하는 조절자와 조절자의 역할 및 발현 여부에 대한 질의어로 나타난다. 그림 4에서 루트노드를 살펴보면 YKL109W 조절자가 대입되어 있다. 노드에서 조절자의 역할이 활성자이면 빨간색으로, 역할이 억제자이면 초록색으로 이름을 표시하였다. 조절자 이름 뒤에 위쪽으로 향하는 화살표가 있는데 이는 조절자의 발현 여부에 대한 질의어이다. 즉 루트 노드가 갖고 있는 의미는 '활성자 YKL109W가 발현 되었는가'이다. 이 노드의 오른쪽 자식 트리에 속하는 단말 노드들에 있는 유전자 모듈의 발현은 YKL109W가 발현 활성화된 결과이다.

유전자 모듈의 발현 데이터는 조절 프로그램의 트리

모양에 따라 재정렬 되어있다. 트리의 노드의 질의어에 따라 의사 노드에 영향을 받는 실험 조건들은 의사 노드 하부에 대입되며 유전자 모듈의 데이터는 재정렬 된다. 이 때 같은 실험 조건하에서 측정된 데이터는 하나로 묶여서 이동되어 서로 떨어지지 않도록 한다.

5. 실험 결과

그림 3, 4는 하나의 실험 조건 하에서 모듈의 발현 데이터와 조절자의 발현 데이터의 발현 유사도가 0.7이상일 때 조절자가 조절역할을 한다는 가정에서 조절 프로그램을 구성한 것이다. 조절 프로그램의 의사 노드는 조절자의 역할 및 조절자의 발현 상태에 대한 질의어로 설명된다. 의사 노드의 왼쪽 노드는 질의어에 대한 답이 false일 때, 오른쪽 노드는 true일 때 선택된다.

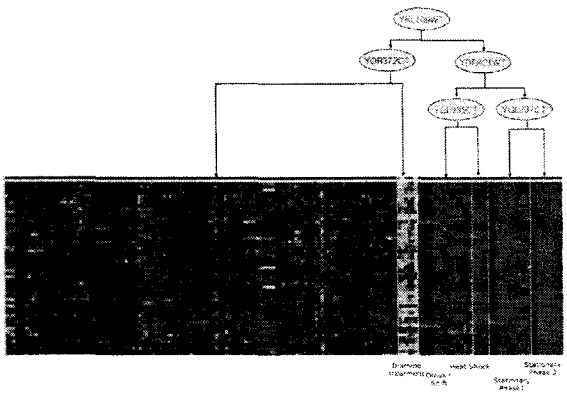

그림 3 Segal의 유전자 모듈 1에 대한 조절 프로그램 예측 결과 - 실험 조건에서 70%이상 유사도를 보일 때 조절 받는다고 가정한 경우

그림 3은 조절자와 유전자 모듈의 정렬 결과 하나의 실험 조건에서 발현 유사도가 70%이상일 때 조절자의 조절 영향을 받는다고 가정했을 때이다. 그림 4는 그림 3보다 좀 더 낮은 임계값을 준 경우이다. 여기서는 발현 유사도가 50%이상일 때, 그리고 정렬을 수행하기 이전의 발현 데이터 3진화 과정에서 유전자 모듈의 평균 발현 데이터와 편차가 많이 나는 데이터를 제외하고 평균을 구하여 발현 데이터를 구한 것이다. 유전자 모듈의 발현 데이터가 전체적으로 발현값이 높더라도 편차가 많이 나는 데이터들이 존재하게 되면 평균값을 떨어뜨리는 역할을 해서 발현 평균값이 떨어지게 되고, 이 결과로 정렬을 수행하게 되면, 조절받는다고 할지라도 조절 구간이 나타나지 않게 된다.

Segal의 결과와 본 논문의 결과에서 가장 두드러지는 차이점은 조절 프로그램의 조절자들이다. Segal은 유전자 발현 데이터 및 annotation 데이터 등을 이용하여 466개의 알려진 조절자 리스트 중에서 조절 모듈을 조절하는 조절자를 직접 예측하였다. 하지만 본 논문에서는 DNA-바인딩 데이터를 통해 실제 알려진 조절 관계가 존재하는 조절자들을 선택해 조절 프로그램을 구성하고 있다. 위의 실험 결과에도 나타나 공통적으로 나타나는 조절자는 HAP4 외에는 존재하지 않았다. 그리고 현재 조절 프로그램을 구성하는 데 여러 가지 제약을 두어 Segal의 결과보다는 좀 더 단순화된 형태의 조절 프로그

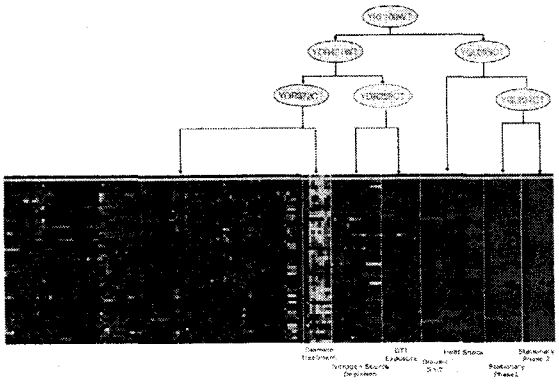

그림 4 Segal의 유전자 모듈 1에 대한 조절 프로그램 예측 결과 - 발현 데이터 3진화 과정에서 편차가 큰 데이터를 제외하고 삼진화하고, 실험 조건에서 50% 이상 유사도를 보일 때 조절 받는다고 가정한 경우

램을 얻을 수 있다. 같은 실험 조건 하에 실험된 array들은 재정렬되지 않으므로 트리 생성에 제약이 많이 된다. 하지만 그로 인해 특정 실험 조건에 대한 유전자 조절 모듈의 발현에 대한 가설을 제시할 수 있다는 장점이 있다.

6. 결론

본 논문에서는 지역 정렬을 이용해 조절 프로그램을 추론하는 알고리즘을 구현해 보고, 실험한 결과를 나타내었다. 본 논문에서 제시한 조절 프로그램은 Segal의 것과 같은 형태인 binary tree 형태를 취하고 있다. 본 논문의 제안 방법과 Segal이 제안한 방법의 차이점은 다음과 같다.

① EM 알고리즘을 사용한 Segal과는 달리 본 논문에서는 서열 정렬 기법을 응용하여 알고리즘을 구성하였다.

② Segal의 조절 프로그램은 하나의 조건 하에 행해진 실험들이 여러 노드들 아래로 흩어진다는 단점이 존재하지만, 본 논문의 결과는 같은 실험 조건들을 가진 데이터들은 하나의 데이터로 취급되므로 여러 노드들의 결과로 흩어지지 않는다.

③ Segal의 조절 프로그램은 조절자들의 역할 및 발현 정도에 따라 조절 모듈 유전자들의 발현 정도가 잘 구분되어 설명되고 있지만, 본 논문의 결과는 그렇지 않은 결과들이 나타난다. 이는 같은 실험 조건에

서 실험된 어레이들을 재정렬하는 제약에 따른 것이다. 하지만 같은 실험 조건이라도 언제 발현이 나타나는지에 대해 좀 더 명확히 볼 수 있다는 장점이 있다.

④ Segal의 경우 말단 노드가 유전자 조절 모듈의 발현 상태는 잘 설명하고 있지만 하나의 말단 노드를 구성하는 실험 조건의 수가 많게는 10가지가 넘을 경우도 있다. 이러한 경우는 특정 말단 노드에 대해서 발현 조건에 대한 가설을 세우기 힘들다. 하지만 본 논문의 실험결과는 조건 및 환경에 영향을 받는 유전자 모듈의 발현에 관한 가설을 세우기 쉽다는 장점이 있다.

앞으로 생각해볼 문제는 다음과 같다.

① 트리 노드에 다수 조절자 허용 문제 - 특정 실험 조건에서 또는 특정 말단 노드에 영향을 미치는 조절자가 둘 이상이 존재할 가능성이 있다. 현재는 가장 영향을 많이 미치는 하나의 조절자만을 결정 노드에 대입하고 있지만, 이를 좀 더 유연하게 처리할 수 있는 방법이 필요하다.

② 정렬 결과 제약에 따른 트리의 확장 - 현재 조절 프로그램에서 특정 실험 하의 어레이들이 발현이 일어났는지 여부는 유전자 조절 모듈의 발현 프로파일과 조절자의 발현 프로파일의 특정 임계값 이상의 매치가 있을 때로 판단한다. 따라서 임계값에 따라 여러 가지 다른 결과들을 얻을 수 있다. 앞으로 정렬 결과 매치되는 부분이 50%이하인 구간, 30% 이하인 구간 등을 나누어 보아 이들을 조절 프로그램의 노드로 구성할 수 있을 것이다.

③ 후보 조절자의 선택 - 현재 후보 조절자를 선택하는 과정은 조절 모듈 내 유전자와 상호작용하는 조절자들을 DNA-바인딩 데이터에서 추출하는 데부터 시작한다. 앞에 서술한 바와 같이 평균적으로 50여개의 유전자들로 구성된 조절 모듈과 상호작용하는 조절자의 수는 20~50여개로 그 수가 매우 많다. 그러므로 이 중에서 의미 있는 조절자를 선택할 필요가 있다. 현재는 조절 모듈의 유전자들 중 2개 이상의 유전자와 조절 관계가 있는 조절자들을 선택하도록 제약을 두고 있다. 하지만 단 하나의 유전자와 조절 관계가 존재하는 경우라도 조절 모듈의 발현과 유사도가 높은 조절자가 존재한다. 이는 유전자 모듈의 다른 유전자들과는 조절 관계가 밝혀지지 않았을 수도 있고 또는 조절자이지만 해당 유전자 조절 모듈에 포함이 되어야하나 누락이 된 경우라 볼 수도 있다. 앞으로 후보 조절자를 선택하는 명확한 기준에 대해 정립할 필요가 있다.

6. 참고 문헌

[1] HW Mewes, K Heumann, A Kaps et al., MIPS: a database for genomes and protein sequences, Nucleic Acids Res. 28, 37-40, (2000).

[2] Eran Segal, Michael Shapira et al., Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature. Genet.*, 34, 166(2003)

[3] T.I. Lee, Nicola J. Rinaldi et al., Transcriptional Regulatory Networks in Saccharomyces cerevisiae, Science, 298, 799-804(2002)

[4] Gasch, A.P. *et al.*, Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11, 4241-4257(2000)