이학석사학위논문

조절 모티프 조합을 동정하기 위한 진화 알고리즘

Evolutionary Algorithms for
Identifying Regulatory Motif Combinations

2004년 2월

서울대학교 대학원
협동과정 생물정보학
정 제 균

조절 모티프 조합을 동정하기 위한 진화알고리즘

Evolutionary Algorithms for
Identifying Regulatory Motif Combinations

지도 교수 장 병 탁

이 논문을 이학석사 학위논문으로 제출함

2003년 10월

서울대학교 대학원

협동과정 생물정보학

정 제 균

정제균의 이학석사 학위논문을 인준함

2003년 12월

위원장 김 주 한 印

부위원장 장 병 탁 印

위 원 김 규 원 印

초 록

전사관련 모듈을 밝히는 것은 유전자 조절네트웍의 규명에 있어서 중요한 요소이다. 마이크로어레이(microarray)에 의한 유전자 발현 분석이 가능해짐에 따라 유전체 범위의 전사 조절 연구가 용이해 졌다.

본 논문에서는 전사관련 모티프들과 유전자 발현데이터를 이용하여 전사에 있어서 공동으로 작용하는 모티프 조합을 탐색하는 방법에 관한 연구를 한다. 여기서 공동으로 작용하는 모티프 조합은 유전자 발현 패턴에 함께 영향을 미치는 모티프 집합을 말한다. 이들의 탐색을 위하여 전사관련 모티프 조합들의 탐색을 최적화 문제로 정의하고, 진화 알고리즘에 의한 해결방법을 제시한다. 진화알고리즘은 하나의 해가 아닌 해의 집단 단위의 탐색을 수행하기 때문에 탐색 공간이 커질 때 최적의 해를 찾을 확률이 높다. 따라서 진화알고리즘은 탐색 공간이 비교적 큰 전사관련 모티프 조합 탐색 문제의 해결에 유용한 알고리즘이라고 할 수 있다. 본 논문에서 제시하는 진화알고리즘은 최적화 성능을 향상시키기 위하여 진화 연산의 하나인 지역탐색 연산을 추가로 수행한다. 그리고 적합도는 공통 모티프를 가지는 유전자 발현패턴들의 군집화 정도를 측정한다.

실험으로써 진화알고리즘을 통하여 효모의 유전자 발현 프로파일 상에서 공동 작용한다고 추정되는 모티프 조합의 탐색 결과들을 제시한다. 이러한 결과들은 유전자 발현 양상에 따라서 공동 작용하는 모티프들의 유용한 생물학적 증거들을 제공해 준다.

주요어: 전사관련 모티프, 진화 알고리즘, 유전자 발현 프로파일
학번: 2002-20634

제 목 차 례

표 차 례

그 림 차 례

제1장 서 론

1.1 연구 배경

현대 생물학에 있어서 주요한 연구 주제 중에 하나는 어떻게 조절에 관한 정보가 유전체(genome)상에 인코딩되어 있는지를 밝히는 것이다. 지난 몇 년 동안 유전자 조절에 관한 연구가 집중적으로 연구가 되어왔지만 조절 기전(mechanism)에 관한 부분적인 지식만 밝혀지고 있다. 진핵생물(eukaryote)에 있어서 전사 조절(transcriptional regulation)은 전사인자(TFs; transcription factors)로 알려진 수 백개의 특정한 DNA 결합 단백질들에 의한 작용이라고 할 수 있다. 각각의 전사인자는 짧은 서열 요소들인 프로모터(promotor)의 특정한 영역을 인지한다. 이러한 영역 정보 및 관련 전사인자 하나 하나가 조절 기전의 기본 구성원이 된다.

최근 지놈 서열을 빠른 속도로 시퀀싱할 수 있는 기술이 개발됨에 따라 수많은 종들의 전체 지놈 서열들을 통한 연구가 가능하게 되었다. 연구자들은 이를 바탕으로 하여 특정 모티프의 탐색과 유전자 예측 알고리즘과 같은 서열 분석 알고리즘들을 개발하게 되었다. 또한, 현재 전사 조절에 관련된 요소들의 서열을 통하여 실험하기 전에 미리 프로모터를 탐색해 보는 것이 가능하게 되었다. 이와 더불어 대규모 DNA 마이크로어레이(microarray)의 사용은 수 천개의 유전자들의 발현을 동시에 관측할 수 있게 하였고, 전사 조절 기전의 연구에 좋은 리소스를 제공하게 되었다.

과거에는 단지 서열정보만을 가지고 특정 패턴을 인지함에 의해서 조절요소를 찾는 연구들이 많았다. 하지만 최근 현재 몇몇 연구들은 이러한 대량의 발현 데이터를 추가함으로써 조절요소(regulatory elements)들을 찾는 새로운 방법들을 시도하고 있다[Brazma et al., 1998; Tavazoie et al., 1999; Sinha et al., 2000; Fujibuchi et al., 2001]. 이러한 방법의 기본 가정은 상호-발현(co-expression)되는 유전자들은 공통적인 조절 기전을 수행한다는 것이다[Peter et al., 2002].

마이크로어레이 데이터를 통한 전사 조절 인자들을 탐색하는 기본적인 접근 방법은 마이크로어레이 데이터와 서열 데이터의 두 소스의 상호 관계를 근거로한 것이다. 모티프 탐색을 위한 일반적인 절차는 먼저 발현 프로파일을 통하여 군집화를 수행한 다음, 각 그룹에 속하는 유전자들의 상류(upstream) 서열을 모티프 탐색 알고리즘으로 탐색해 나가는 것이다. 전사 조절 인자들을 탐색하는 것과는 반대의 접근도 가능하다. 우리가 전사 조절 모티프를 알고 있다면 발현 패턴을 어느 정도 유추할 수 있다.

지금까지 많은 연구들은 개개의 조절 요소를 찾는데 초점을 두고 진행되어 왔다[Fujibuchi et al., 2001]. 조절 기작의 연구는 이러한 개개의 조절 요소를 동정하는 것도 물론 중요하지만, 전체적인 조절 기작을 이해하기 위해서 단지 개개의 조절 요소로서 따로 분리해서 생각하지 않고, 하나의 조절에 대하여 다른 전사인자들과의 상호작용도 고려해야 한다. 예를 들어, 효모(*S. cerevisiae*)의 세포주기(cell cycle) G1 후기에서 MBF와 SBF는 상호 협조적으로 활성화되어서, 기능적으로 관련되는 유전자들을 조절한다고 보고되고 있다[Iyer et al., 2001]. MBF는 Mbp1과 Swi6의 복합체이고 SBF는 Swi4와 Swi6 복합체이다. 이러한 조절인자 복합체 조합들이 G1 후기 유전자들의 상류 지역에 바인딩해서 유전자들을 조절하게 된다.

최근에 이러한 전사인자들 간의 상호작용은 효모의 Genome-wide location 분석 방법을 통해 검증되고 있다[Simon et al., 2001]. 하지만 아직도 수많은 실험 조건과 이러한 조건에 반응을 하는 수많은 전사인자들의 조합들에 관하여 알려진 것이 적다. 앞으로도 마이크로어레이와 같은 대규모 실험 데이터 뿐만 아니라 다양한 데이터로부터 조절 기작을 해명하는 계산적인 접근 방법이 나오리라 예상하며, 본 논문은 계산학적 관점에서 위에서 언급한 모티프들의 상호관계 문제를 해결하는 하나의 방법을 제시하고자 한다.

1.2 연구 목적

전체적인 조절 기전의 이해는 조절인자가 결합하는 모티프들의 탐색이 필수적이다. 이와 더불어 조절관련 모티프들이 공동으로 작용하는 효과를 밝히는 작업도 중요하다. 이를 위하여 특정 기능에 있어서 공동적인 조절 효과를 가지는 모티프 조합을 탐색할 수 있는 계산학적 방법이 필요하게 된다.

본 논문에서는 공동으로 작용하는 모티프 조합에 대한 탐색 문제를 최적화 문제로 보고, 계산적으로 쉽게 해결할 수 있는 방법을 제안하고자 한다. 최적화시키는 방법은 기존에 흔히 행해졌던 군집화 수행 후에 각각의 모티프 조합을 탐색하는 방법이 아니라, 군집화를 동시에 수행하면서 공동으로 작용하는 모티프 조합을 탐색한다.

1.3 연구 내용

본 연구에서는 진화 알고리즘을 통하여 조절 모티프 조합을 탐색하는 방법을 제시한다. 진화 알고리즘은 자연선택의 원리를 기반으로 한 최적화 방법으로서 탐색, 최적화 및 기계학습을 위한 도구로 널리 사용되고 있다. 진화 알고리즘을 통한 모티프 조합 탐색의 장점은 전체 조합을 모두 조사하지 않고도 비교적 빠른 시간 안에 최적의 해들은 찾아낸다는 것이다.

먼저 진화 알고리즘에 있어서 개체를 설계하고 적합도 함수(fitness function)를 정의한다. 개체는 해결하고자 하는 문제의 특성에 맞게 문자열이 인코딩되어야 하고 적합도는 해답을 유도하는 수식으로써 정의되어야 한다. 본 논문에서는 이러한 두 조건들에 만족하는 개체설계 및 적합도 함수를 제시한다. 먼저, 진화 알고리즘의 기본 개념에서부터 시작해서, 최적화 성능을 개선한 Memetic 알고리즘에 의한 탐색 방법으로 확장한다.

본 연구에 사용되는 모티프 데이터는 효모의 유전체를 AlignACE 프로그램을 통하여 탐색한 후, 전처리과정을 거친 것이다[Pilpel et al., 2001]. 다음, 유전자 발현 데이터는 조절 네트웍 연구에 주로 이용되는 효모의 데이터이며 Spellman의 세포주기를 포함하여 sporulation, heat-shock, diauxic shift에 대한 것이다[Cho et al., 1998; Spellman et al., 1998; Chu et al., 1998; Eisen et al., 1998]. 마지막으로 단백질 상호작용 데이터는 DIP (Database of Interacting Proteins) 데이터베이스에서 제공하고 있는 효모 단백질들의 상호작용 리스트이다.

실험 결과는 최적화를 통한 모티프 조합 탐색 결과와 알려진 단백질 상호작용 데이터를 비교하여 탐색 결과의 타당성을 검증한다. 또한, 알려진 모티프 집합을 가지고 세포주기의 각 단계에서 진화 알고리즘을 통하여 얻은 모티프 조합 결과를 기존 문헌들을 통하여 분석한다. 더 나아가, 지역 탐색의 효과를 알아보기 위하여 유전 알고리즘과 Memetic 알고리즘의 성능을 평가해 본다. 마지막으로 알려지지 않은 모티프 집합을 추가하여 가능성이 있는 모티프 조합들을 탐색한다.

이 논문에서는 먼저 2장에서 조절 모티프 및 모티프 조합 탐색에 대한 기본 개념을 설명한다. 그리고 진화 알고리즘에 대한 간략한 기술을 한다. 3장에서는 조절 모티프 조합을 탐색하기 위한 알고리즘의 설계 및 학습 방법에 대해서 살펴보기로 한다. 4장에서는 실험에 사용한 데이터의 설명과 실험 방법 및 실험 결과를 상술하며, 마지막 장에서는 결론을 언급하고 향후 연구 주제들을 제시하고자 한다.

제2장 조절 모티프 및 모티프 조합 탐색

2.1 조절 모티프

하나의 세포에는 무수히 많은 유전자가 존재하고 유전체 상의 특정한 유전자의 발현을 재프로그래밍함에 의해서 환경적 변화에 응답한다. 이렇게 많은 유전자는 각각의 고유 기능을 가지고 있으며, 다른 유전자들과의 직접적이거나 간접적인 단순한 관계를 가지고 복잡한 전체 유전자의 조절 시스템을 만들어 간다. 특정한 유전자의 발현율은 다양한 조절 단백질의 상호작용에 의해서 결정된다. 하지만 전체적으로 조절 단백질들의 집합들이 유전자들의 집합을 어떻게 조절하는지에 대한 관계는 조절 네트웍에 의해 표현될 수 있다.

그림 1은 각각의 조절에 관련된 단백질들과 목표유전자들과의 관계를 보여주는 조절네트웍의 간단한 개념도이다. 각각의 원은 전사적인 활성 상태를 나타내고, 사각형은 지놈상에서 잠재적인 목표유전자를 나타내며, 각각의 화살표는 활성의 관계를 나타내고 있다. 이처럼 유전자 조절 네트웍은 마치 회로도와 같이 유전자의 활성과 불활성을 조절하는 동적인 시스템의 관점으로 보여질 수 있다.

이러한 유전자 발현의 조절은 전사 시작 영역의 상류(upstream)에 위치해 있는 보존된 짧은 서열 요소(conserved sequence element) 또는 DNA 모티프가 관련되어 있다. 이러한 서열 대부분은 전사요소들의 결합 영역(binding site)들이다. DNA 조절 모티프들은 유전자들의 상류 지역에 대하여 지역 정렬(local alignment)을 하여 탐색할 수 있다[Roth et al., 1998]. 지금까지 다양한 모티프 탐색 알고리즘에 대한 샘플링(sampling)방법이 소개되었다.

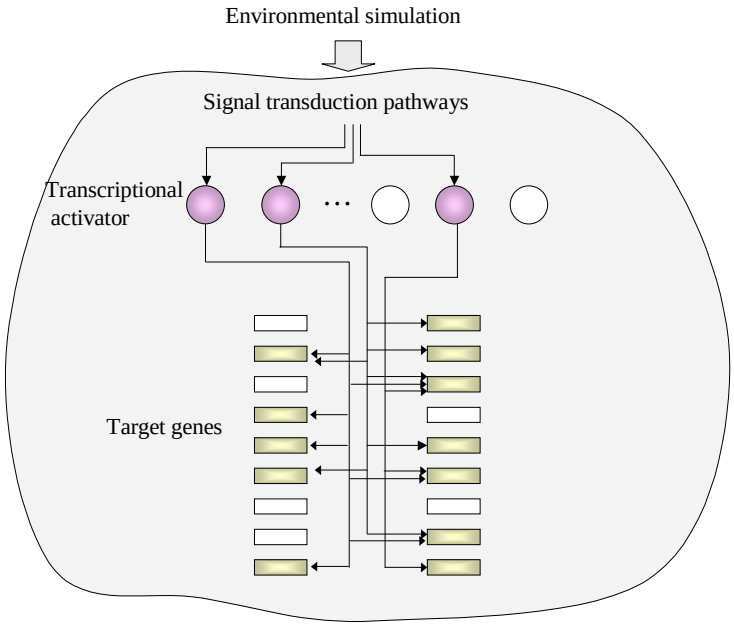

그림 1 조절네트웍의 개념도: 각각의 원은 전사적인 활성 상태를 나타낸다. 사각형은 지놈상에서 잠재적인 목표유전자를 나타낸다. 각각의 화살표는 활성의 관계를 나타내고 있다.

먼저 Gibbs 샘플링은 반복적인 샘플링에 의한 통계적 방법에 근거하고 있다[Liu et al., 1995]. 이러한 알고리즘은 박테리아 지놈과 *S. cerevisiae*의 모티프들을 탐색하기 위한 AlignACE 프로그램에서 적용된바 있다[Roth et al., 1998]. 다음 MEME (Multiple EM for Motif Elicitation)은 입력 서열 데이터에서 하나 또는 그 이상의 반복되는 패턴들을 발견하기 위하여 개발된 알고리즘이다[Bailey and Elkan, 1995]. 모티프들을 탐색하기 위한 여러 알고리즘들이 존재하고 있지만 비암호화(noncoding) 영역의 짧은 모티프를 탐색하기는 쉽지 않다.

그래서 최근에 DNA 칩 또는 마이크로어레이로 부터 얻어진 발현 데이터로부터 각각의 조절 모티프의 탐색을 시도하는 몇몇 연구가 있었다 [Brazma et al., 1998; Tavazoie et al., 1999; Sinha et al., 2000; Fujibuchi et al., 2001]. 대게 전사요소들은 세포내의 비슷한 기능을 수반하는 유전자 그룹들을 조절한다. 따라서 상류 지역에 같은 모티프를 가지고 있는 유전자들은 세포내의 같은 기능에 참여한다고 볼 수 있는 좋은 후보 집합이다. 바꾸어 말하면 유사한 패턴으로 발현되는 유전자 집합은 같은 모티프들을 포함하고 있을 확률이 많다. 이러한 지식들은 유전자 조절 네트웍에서 연결선들에 관한 강력한 가설들을 제공할 수 있다.

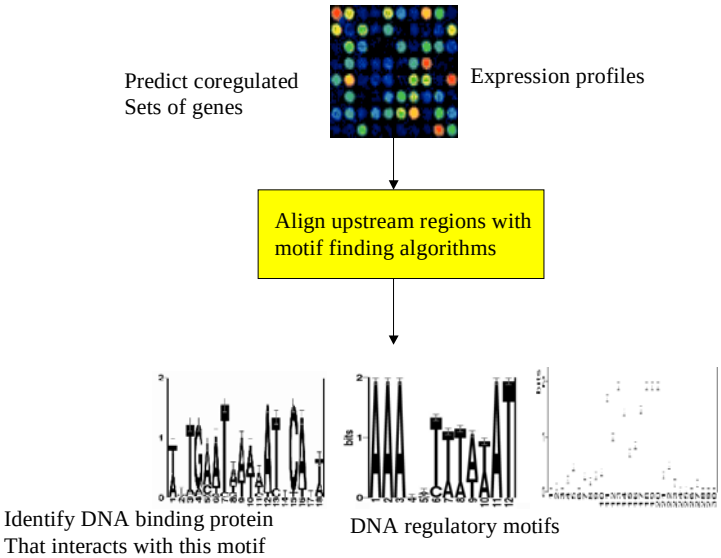

그림 2 발현 데이터를 통한 모티프 탐색 절차

그림 2는 발현 데이터로부터 모티프 탐색의 수행 절차를 보여주고 있다. 첫 번째 단계에서는 유전자들의 상호-조절되는 그룹들을 군집화 알고리즘 등에 의하여 군집화한다. 그런 다음, 각 그룹내에 속하는 유전자들의 상류 지역을 정렬알고리즘을 통하여 정렬하여 모티프들을 찾아낸다. DNA 조절 모티프들의 결과는 그림에서 보여지는 바와 같이 모티프 로고를 이용하여 설명될 수 있다. 여기서 문자들의 스택 높이는 정보 함유량(information content)을 나타낸다. 그리고 각 베이스의 상대적인 빈도는 상대적인 높이로 보여진다. 이러한 방법으로 발견되는 중요한 모티프의 존재는 상호-조절되는 유전자의 예측에 재사용될 수 있다.

2.2 모티프 조합 탐색

현재 여러 가지 모티프 탐색 알고리즘들이 개발되어 있지만 대부분 개개의 조절 요소를 찾는데 초점을 모으고 있다. 많은 진핵생물(eukaryote)들의 유전자는 전사를 조절하기 위하여 상호적으로 다중 전사 인자들이 결합된다. 따라서 개개의 조절 요소 뿐만 아니라 상호작용으로 조절하는 요소들의 연구가 더욱 중요해지고 있다. 최근 몇몇 연구들은 조절 모티프들의 조합을 정의하기 위하여 계산적 알고리즘을 제시하고 있다[Pilpel et al., 2001; Bussenmaker et al., 2001].

이러한 연구 중의 하나로써 모티프 모듈이라는 것을 정의하기 위한 모델이 제시된 바 있다[Segal et al., 2002; Segal et al., 2003]. 이 연구에서는 모티프들을 탐색함과 동시에 모티프 모듈에 관련된 발현 패턴을 분석하는 단일화된 확률모델을 제시하였다. 이러한 시도는 모티프 조합과 발현 패턴의 상호 관계를 분석하는데 있어서 유용한 툴을 제공할 수 있다.

일반적으로 모티프 조합을 탐색하는 것은 많은 계산 시간을 요구하게 된다. 예를 들어, 효모는 200개 이상의 전사를 조절하는 단백질이 있고 300개 이상의 결합 모티프가 존재하고 있다. 만약 5개 까지의 모티프 조합을 계산한다고 하더라도 상당한 조합의 수가 발생하게 된다. 이러한 조합 탐색은 목표 함수만 정의된다면 최적화 알고리즘으로 쉽게 해결할 수 있다.

2.3 진화 알고리즘

진화 알고리즘(evolutionary algorithms, EAs)은 자연계의 진화 과정을 컴퓨터 상에서 시뮬레이션함으로써 복잡한 실세계의 문제를 해결하고자 하는 계산 모델이다. 진화 알고리즘은 염색체를 표현하는 방법과 사용되는 유전 연산자의 종류 및 특성에 따라서 크게 네 가지의 모델로 구분된다. 유전 알고리즘(genetic algorithms, GAs)과 진화 전략(evolution strategy, ES)에서는 고정된 길이의 이진 스트링이나 실수의 값으로 구성된 벡터를 염색체로 사용하는 반면, 진화 프로그래밍(evolutionary programming, EP)은 염색체의 표현에 제약이 없으며, 유전자 프로그래밍(genetic programming, GP)에서는 트리로 염색체를 표현한다. 그리고, EP와 ES는 돌연변이(mutation), GAs 와 GP는 교차(crossover) 연산자와 돌연변이, 모두를 주로 사용한다.

진화 알고리즘 중에서 유전 알고리즘은 1975년에 John Holland가 저서 “Adaptation on Natural and Artificial Systems” 에서 처음 소개하여 이론적 기반을 다졌으며 자연도태의 원리를 기초로 한 최적화 방법으로서 탐색, 최적화 및 기계학습을 위한 도구로 많이 사용한다[Holland, 1975]. 유전 알고리즘에서는 풀고자 하는 문제에 대한 가능한 해들을 정해진 형태의 자료구조로 표현하는데 이를 염색체, 혹은 개체(individual)라고 부른다. 그리고 정해진 수의 염색체 집단을 운영하는데 이 집단을 개체군(population)이라고

한다. 염색체 상의 각 인자는 유전자라고 부른다. 여기서의 유전자를 본 논문에서 생물학 관련 유전자와 혼동해서는 안된다. 생물학에서는 많은 수의 염기가 모여서 유전자를 형성하지만, 유전 알고리즘에서는 유전자가 최소 단위가 된다. 여기서는 이러한 혼동을 피하기 위하여 문자열(string)이라고 표현하기로 한다.

```
begin
  t := 0;
  Initialize population P(t);
  Evaluate population P(t);
  while (not termination condition) do
     t := t + 1;
     Select P(t) from P(t-1);
     Crossover P(t)
     Mutate P(t)
     Evaluate  P(t)
  end
end
```

그림 3 유전 알고리즘의 Pseudo 코드

유전 알고리즘은 임의의 값으로 초기화된 개체들의 집합으로 시작한다. 그림 3은 유전 알고리즘의 Pseudo 코드를 보여주고 있다. 각각의 개체는 상대적인 문제해결 능력에 따라 적합도(fitness)가 평가되며 적합도 및 연산자에 따라 다음 세대(generation)에 문자열이 복제(reproduce)되는 가능성이 조절된다. 유전 알고리즘에서 사용되는 기본적인 연산자는 선택(selection), 교차(crossover), 변이(mutation)의 3가지이다. 선택 연산자는 교차를 할 해

를 개체군에서 선택하는 연산자로서, 잘 적응한 해는 살아남아서 다음 세대에 전달되지만 적응을 잘 못한 해들은 도태되도록 유도한다. 이 때 선택된 해를 부모해(parent)라고 하고 다음 세대에 생성되는 해를 자식해(offspring)라고 한다. 따라서 만약 개체군에서 적합도가 높은 개체들은 다음 세대에 복제될 가능성이 많게 된다. 교차는 두 개의 부모해로부터 자식해를 만들어내는 연산자로서, 부모해에 있는 우수한 속성을 자식해에 전달하고자 하는 것이 목표이다. 반면에 변이는 해를 임의로 변형시키는 연산자로서 부모해에 없는 속성을 도입하여 해의 다양성을 높이는 것이 목표이다.

제3장 진화 알고리즘에 의한 모티프 조합 탐색

공동 작용하는 모티프 조합 탐색을 위한 진화 알고리즘에 있어서 개체의 염색체는 가변길이의 인코딩이 가능하도록 하였다. 그림 4는 전사 조절 모티프 조합을 탐색하기 위한 진화 알고리즘의 개념도를 보여주고 있다. 개체집단은 잠재적인 모티프 조합들의 집합이고 세대가 감에 따라 유전 연산자들에 의하여 적합하게 된다. 알고리즘의 진행 후, 마지막 세대에서 최적해가 선택되고 단백질 복합체와 같은 다양한 소스로부터 탐색 결과가 검증된다.

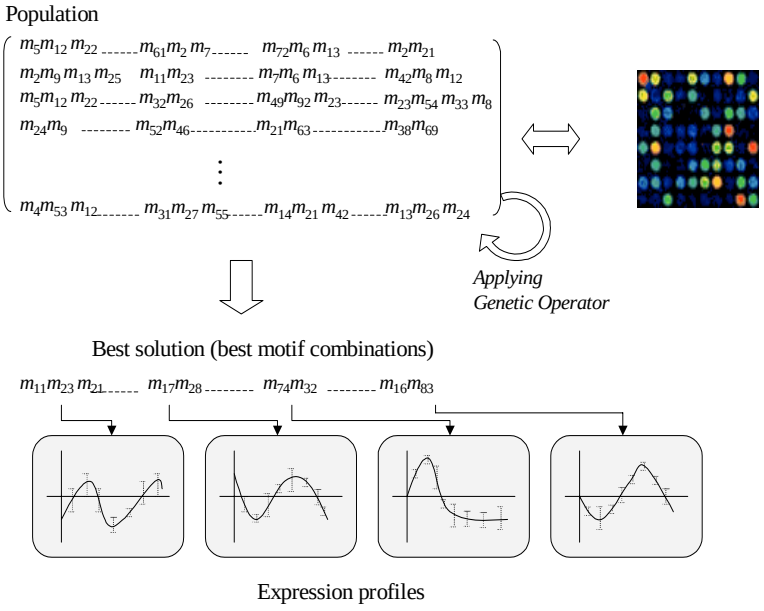

그림 4 진화 알고리즘에 의한 모티프 탐색의 개념도: 각 개체의 m_i는 모티프 인덱스를 나타낸다. 최적의 해는 각 모티프 조합을 가지는 유전자군 내의 발현패턴들 간의 유사도가 높으며, 군 간의 발현 패턴 유사도는 적어야 한다.

3.1 개체 표현 및 학습

각 개체에 대한 염색체는 K개의 문자열들의 집합 $\{S_1, \ldots, S_K\}$으로 표현된다. 각각의 문자열 S_i는 $[L\ m_1\ m_2\ \ldots\ m_{Lmax}]$이다. 여기서 m_i는 모티프의 인덱스를 나타낸다. L은 몇 개의 모티프들이 실제 표현형으로 사용되는지를 명시한 값으로써 모티프 조합에 대하여 가변길이를 가지도록 한다. 따라서 첫 번째 문자 m_1부터 m_L개 까지만 사용되고 나머지는 사용되지 않는다. 이러한 L은 진화가 됨에 따라 최적화될 것이다. 본 논문에서는 L을 이진으로 인코딩한다. 그림 5에서 보이는 바와 같이 각각의 M_i가 i번째 문자열 $\{m_1, \ldots, m_L\}$의 모티프 조합이라고 할 때, 진화는 K개의 모티프 조합들 $[M_1, \ldots, M_K]$를 최적화하는 것이다.

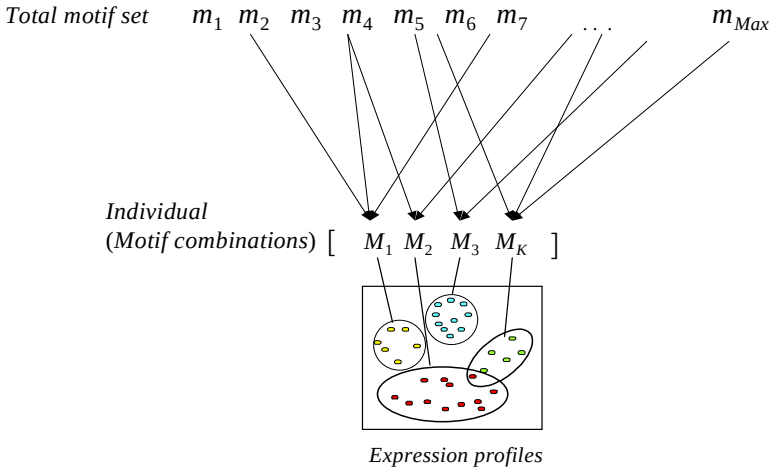

그림 5 개체의 표현 및 학습: 하나의 개체는 모티프 조합들이고 학습은 최적의 모티프 조합들을 탐색하는 것이다.

선택 방법은 roulette wheel selection (RWS)을 취하였다. 유전 알고리즘은 다양한 선택 계획이 존재하며 RWS는 확률에 근거한 방법에 속한다 [Miller et al., 1995]. RWS는 개체군 내의 각각의 개체들은 적합도에 근거한 확률에 해당하는 roulette wheel 슬롯의 크기를 가지고서 슬롯을 돌려서 선택하는 것과 비슷한 방식으로 동작한다. RWS는 각 개체당 하나의 섹터에 해당하도록 한다. 여기서 하나의 섹터는 다음식과 같이 $P_{sel}(i)$의 확률로 선택될 수 있다.

$$P_{sel}(i) = \frac{f(i)}{\sum_{i=1}^{n} f(i)}$$

RWS는 먼저 균일하게 분포한 랜덤 수 r이 생성된다. 만약 r이 i 와 $i+1$ 번째의 누적확률 사이에 있다면 i번째 개체가 선택된다. 이러한 과정은 다음 세대에 교체될 개체 수 만큼 반복된다.

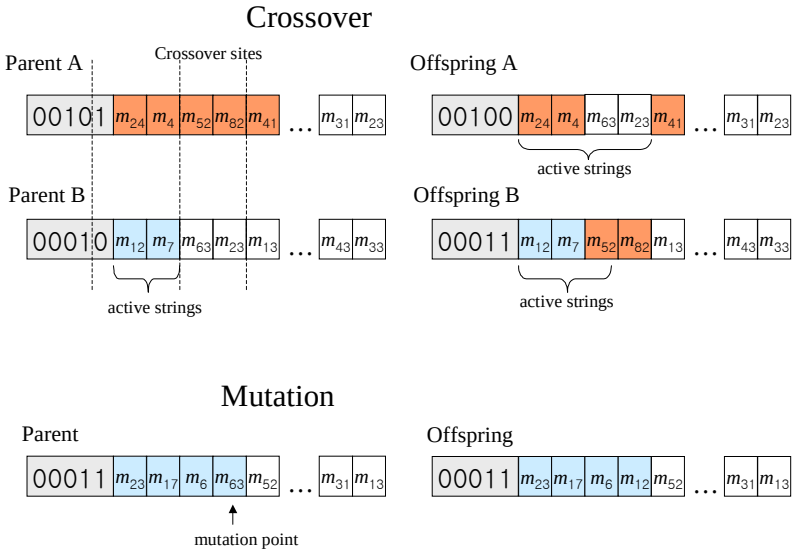

그림 6 교차연산과 돌연변이연산 방법

그림 6은 교차연산과 돌연변이연산 방법을 보여주고 있다. 교차연산은 두 부모로부터 스트링들의 쌍을 랜덤하게 선택한다. 그런 다음 선택된 스트링들이 확률 p_c에 의해서 교차연산에 적용받게 될지를 결정한다. 교차연산은 균일(uniform), 단일 포인트(single-point) 그리고 다중 포인트(multi-point)연산이 있을 수 있고 본 논문에서는 단일 포인트에 의한 교차연산을 사용하였다. 돌연변이연산은 각 부모 스트링에서 각 포인트마다 확률 p_m의 확률로 새로운 값으로 바뀌게 된다. 따라서 돌연변이연산에 적용을 받게 되면 모티프 인덱스가 다른 인덱스로 교체된다.

3.2 적합도 함수

각 개체에 대한 적합도 함수는

$$Fitness = \alpha\ MECS + \beta\ SSC$$

으로써 정의된다. MECS (mean of EC score)는 각 유전자 그룹 내에서 발현 패턴들의 밀집정도를 측정한 것이고 SSC (sum-of-squares for centers)는 서로 다른 유전자 그룹의 평균 발현 패턴들의 퍼짐 정도를 측정한 것이다. 여기서 각 그룹 내의 유전자들은 특정 모티프조합 M_k를 가지고 있다. 그림 7에서 보이는 바와 같이 MECS는 한 그룹 내에서 발현패턴들의 밀집도를 나타내고 SSC는 그룹들 간의 분리 정도를 나타낸다. 일반적인 k-Means 또는 계층적 군집화와 같은 군집화 알고리즘과는 달리 여기서는 모티프 조합에 의존하여 하나의 유전자가 여러 그룹에 속할 수 있다. 식에서 α 와 β는 각 항목에 대한 가중치이다.

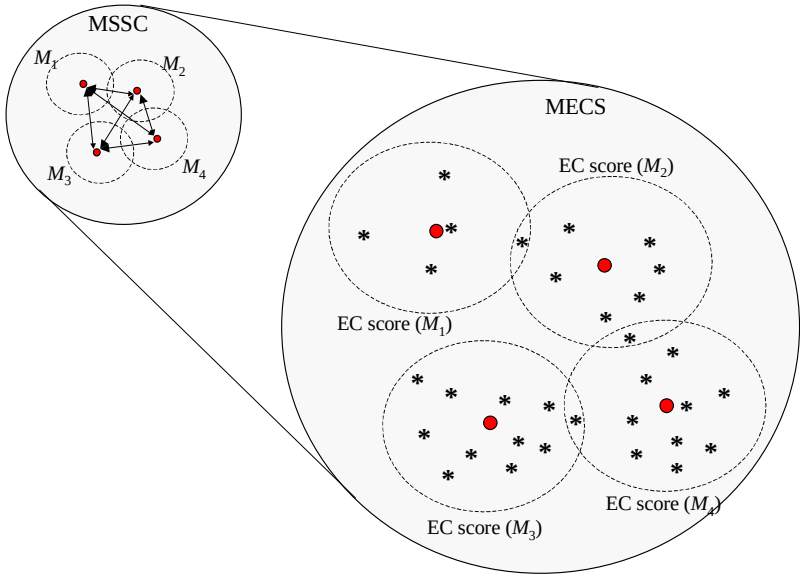

그림 7 적합도 함수의 개념도 (K=4): MECS는 한 그룹 내에서 발현패턴들의 밀집도를 나타내고 SSC는 그룹들 간의 분리 정도를 나타낸다. *은 유전자 발현 프로파일을 의미한다. ●은 그룹 내의 평균 발현 프로파일을 의미한다.

MECS는 K개의 서로 다른 모티프 조합들이 주어질 때

$$MECS = \frac{1}{K} \sum_{k=0}^{K} EC\ score_k$$

으로써 발현 응집력 점수(EC score)들의 평균이다. 발현 응집력 점수는 Pilpel이 제안했으며 모티프 조합을 가지고 있는 유전자들의 발현 패턴들의 유사도를 측정한 것이다[Pilpel et al., 2001]. 여기서 각각의 발현 응집력 점수는

$$EC\ Score = \frac{the\ number\ of\ gene\ pairs\ (d_{ij} < T)}{the\ number\ of\ total\ gene\ pairs}$$

으로써 계산된다. 발현 응집력 점수에서 분모는 그룹 내에 있는 총 유전자 쌍의 수를 말한다. 만약 그룹에 J개의 유전자가 존재한다고 하면 가능한 유전자 쌍은 $0.5\ J\ (J-1)$가 된다. 그리고 분자는 모든 유전자 쌍 i, j에 대하여 발현 패턴들 간의 거리(distance) d_{ij}가 역치 T 이하의 값을 가지는 개수이다. 여기서 역치 T는 전체 유전자들에서 랜덤 추출된 모든 유전자 쌍의 발현 패턴에 대한 거리들을 오름차순으로 정렬했을 때 상위 5 % 지점에 해당하는 d 값을 말한다. 즉 임계치 T는 5 백분위수(percentile)를 의미한다.

유전자 쌍의 발현패턴에 대한 거리는 유클리디안 거리(Euclidian distance) $d_{ij} = ED(v(g^i_{M_k}), v(g^j_{M_k}))$를 계산하여 구하게 된다. 여기서 $v(g^i_{M_k})$, $v(g^j_{M_k})$은 특정 모티프 조합 M_k를 가지는 유전자들로 이루어진 그룹 G_{M_k}에 속하는 유전자들의 발현패턴들이다.

SSC는

$$\frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{K} ED^2(v(\widehat{G_{M_i}}), v(\widehat{G_{M_j}})), \quad j \neq i$$

와 같이 계산된다. 여기서 $ED(v(\widehat{G_{M_i}}), v(\widehat{G_{M_j}}))$는 i그룹의 평균 발현 패턴과 j그룹의 평균 발현 패턴 간의 유클리디안 거리이다.

3.3 Memetic 알고리즘에 의한 모티프 조합 탐색

Memetic 알고리즘은 유전 알고리즘과 마찬가지로 조합 최적화 문제를 풀기 위한 개체 기반의 탐색을 수행한다[Moscato et al., 1989]. Memetic 알고리즘은 진화 알고리즘에 지역적 근접(local neighborhood) 탐색 또는 시뮬레이티드 어닐링(simulated annealing)과 같은 지역 탐색 기법을 결합시킨 방법이다. 그림 8과 같이 Memetic 알고리즘의 기본 전략은 지역 최적화에 대한 탐색을 수행하여 자식해에 대한 개선을 유도하는 것이다. Memetic 알고리즘은 TSP, GBP, QAP, NK-Landscapes, binary quadratic programming등의 조합 최적화 문제에 매우 좋은 성능을 보였다[Freisleben and Merz, 1996].

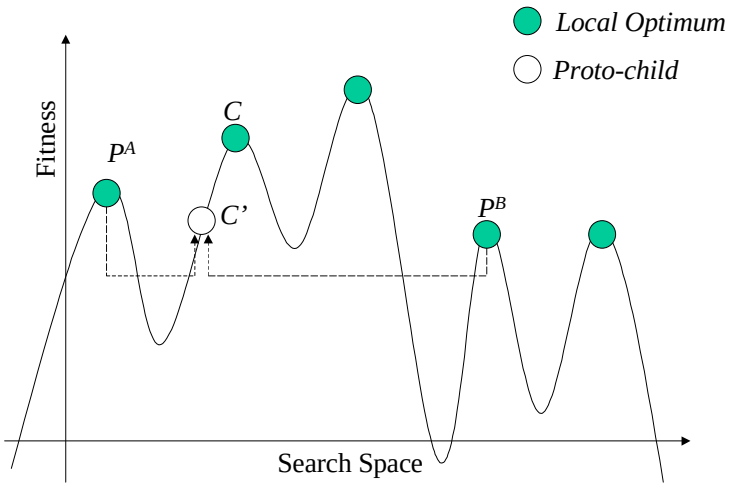

그림 8 부분 공간 탐색과정을 수행하는 Memetic 알고리즘

그림 9는 Memetic 알고리즘에 대한 Pseudo 코드를 보여주고 있다. Memetic 알고리즘은 보통 교차연산과 돌연변이 후에 지역탐색을 수행하게 된다. 본 연구에서는 자식에 대한 문자열들이 주어졌을 때, 3.1절에 언급한 각 문자열 S_i에 대하여 언덕오르기(hill-climbing)기법으로 지역 탐색을 수행하였다. Memetic 알고리즘에 대한 형식적 기술을 하면 다음과 같다.

만약 탐색공간(search space)을 Z 라고 하고 표현공간(representation space)을 V 라고 한다면, 표현 함수는

$$\rho : Z \rightarrow V$$

과 같이 표현형에서 인자형의 매핑이라고 할 수 있다. Z 상에 어떤 해답이 있을 때, 알고리즘은 표현형을 나타내는 V를 반환한다. 표현공간을 실수값으로 매핑하는 적합도 함수

$$f : V \rightarrow R^{+}$$

가 주어진다면 알고리즘은 표현 공간에서의 전역 최적화 $V^{\#}$ 탐색하기 위하여 적합도를 극대화할 것이다. $V^{\#}$의 문자열은 특정 이동 연산자(move operator) Q에 의하여 성취된다. 유전 알고리즘에 있어서 이동 연산자들은 교차와 돌연변이 연산자들이다. K_Q는 Q에 의하여 가능한 공간상의 이동이 실제로 발생할 것인가의 결정을 이끄는 매개변수를 제어하는 제어집합이다. 표현공간 V와 이동 연산자 Q와 Q에 의한 지역 최적화의 부분공간 V_Q가 주어진다면, 언덕오르기는 어떤 함수

```
procedure MA;
  begin
    initialize population P of size #Popsize;
    for each individual I∈P do I := Local-Search(I);
    t := 0
    while (not termination condition) do
       for i := 1 to #recombinations do
          select two parents I^a, I^b ∈P randomly;
          I' = Recombine(I^a, I^b)
          I' = Local-Search(I')
          add individual I' to P;
       end
       for i := 1 to #mutations do
          select a parent I∈P randomly;
          I' := Mutation(I)
          I' := Local-Search(I' )
          add individual I' to P;
       end
       t := t + 1
    end
end
```

그림 9 Memetic 알고리즘의 Pseudo 코드

$$H: V \times K_H \rightarrow V_Q$$

로 나타낼 수 있다. 전형적인 유전 알고리즘은

$$X : V \times V \times K_X \rightarrow V_Q$$
$$Y : V \times K_Y \rightarrow V_Q$$

와 같이 교차 X와 돌연변이 Y에 의하여 표현공간에 해당하는 새로운 문자열들을 생성한다. 교차와 돌연변이가 결합한 유전적 재생산 함수는 $R_{genetic} = XY$로 나타낼 수 있다. 따라서 여기서 제시하는 Memetic 유전적 함수는

$$R_{memetic} : V_Q \times V_Q \times K_H \times K_M \times K_X \rightarrow V_Q$$

과 같이 언덕오르기를 추가한 $R_{memetic} = XMH$가 된다.

제4장 실험 및 평가

4.1 실험 데이터

제안된 방법을 테스트를 위하여 모티프 데이터, 유전자 발현 프로파일 데이터, 단백질 상호작용 데이터의 세 가지 데이터 소스를 사용했다.

먼저 모티프 데이터는 조절 네트웍을 연구하기 위해 사용했던 Pilpel의 모티프 데이터를 사용했다[Pilpel et al., 2001]. 이 데이터에는 효모의 알려진 37개의 모티프들과 각 유전자의 상류지역을 AlignACE 툴을 사용하여 얻은 329개의 모티프들이 수록되어 있다. 이 데이터에서 하나의 파일 목록은 각 모티프를 가지고 있는 유전자들의 인덱스와 모티프 위치 등이 수록되어 있다. 실제로 학습에 사용하기 위하여 전처리 작업을 하였으며, 전처리된 데이터는 유전자 각각에 대하여 모티프들의 존재 여부를 표현한 바이너리 벡터들을 리스트한 메트릭스이다.

다음, 마이크로어레이 프로파일 데이터는 Spellman의 세포주기 실험 데이터를 포함한 네 가지 데이터이다. 세포주기 데이터는 효모의 6179개의 유전자 중에서 세포주기에 관련된 800개의 유전자 발현 패턴을 정리한 것으로서 알고리즘의 검증을 파악하기 위하여 사용하였다. 이 데이터는 세포주기의 각 단계에 대한 태그(tag)가 명시되어 있다. 세포주기를 비롯한 나머지 sporulation, heat-shock, diauxic shift 데이터는 알려지지 않은 주요 모티프 탐색을 위하여 사용되었다.

마지막으로, 단백질 상호작용 데이터베이스는 DIP이며, 모티프 조합 탐색 결과의 검증을 위하여 사용되었다. DIP 데이터베이스는 실험적으로 밝혀진 단백질 상호작용 목록들을 담고 있다.

그림 10에서와 같이 DIP에는 상호작용에 참여하는 단백질들이 리스트되어 있다. 모티프 조합에 대한 검증은 아래 리스트에서 단지 ORF 이름들 간의 상호 매치를 확인함으로써 수행되었다.

```
DIP:2551N   AAC1   YMR056C DIP:1189N   APG12   YBR217W DIP:11374E    P
DIP:2551N   AAC1   YMR056C DIP:1330N   LSM1    YJL124C  DIP:3267E
DIP:2551N   AAC1   YMR056C DIP:4449N   PUF3    YLL013C  DIP:6745E
DIP:2551N   AAC1   YMR056C DIP:2425N   RAD3    YER171W DIP:13079E    P
DIP:6289N   AAC3   YBR085W DIP:1189N   APG12   YBR217W DIP:11375E    P
DIP:6289N   AAC3   YBR085W DIP:5008N   BUD32   YGR262C DIP:11546E
```

그림 10 DIP 데이터 리스트의 예

알고리즘에 대한 파라미터 설정은 다음과 같다. 먼저 개체군의 크기는 100으로 설정했으며 세대수 역시 100으로 하였다. 교차확률 p_c은 0.9, 돌연변이 확률 p_m은 0.01로 설정하였다. 다음 세대에 대한 복제에 있어서 엘리트(elitist) 선택이 수행되었다. 엘리트(elitist) 선택은 전 세대에서 가장 우수한 개체를 다음 세대에 전달하는 것을 말한다. 개체 표현에 있어서 최대 모티프 개수 L_{max}는 6으로 설정하였다. 그리고 적합도 함수에서 α 와 β에 대한 각 가중치는 각각 10.0과 1.0으로 설정하였으며, 모티프 조합들의 수 K는 4로 설정하였다.

4.2 실험 결과 및 평가

실험은 먼저 5개의 cell-cycle phase별로 군집화된 Spellman의 데이터에서 각 그룹별로 진화 알고리즘을 적용해 보았다. 그런 다음, 발현 응집도 점수를 측정하여 진화를 통해 얻어진 모티프 조합이 공동 작용하는지를 검증하는 방법으로 PI율(protein interaction ratio)을 측정해 보았다.

PI율은

$$PI\ ratio = \log((\sum IR^{M}_{i,j}/P^{M})/(\sum IR^{R}_{i,k})/P^{R}), \quad i \neq j,\ i \neq k$$

으로써 계산된다. 여기서 $\sum IR^{M}_{i,j}$는 모티프 조합 M을 가지는 유전자에 해당하는 단백질 i, j 쌍이 상호작용을 할 경우만을 카운트하게 된다. $\sum IR^{M}_{i,k}$은 랜덤하게 선택된 단백질 쌍에 대하여 계산한다. 그리고 P^{M}과 P^{R}은 각각 모티프 조합의 목표 유전자들 간의 쌍에 대한 총 개수, 랜덤하게 선택된 유전자들 간의 쌍에 대한 총 개수를 의미한다. 그래서 PI율은 랜덤하게 선택된 단백질들이 서로 상호작용 하는 확률과 공통 모티프를 가지는 유전자들에 해당하는 단백질들 간에 상호작용을 하는 확률의 상대적인 비율로써 계산된다. 이러한 식은 공통 모티프를 포함하는 유전자들은 서로 상호작용을 할 확률이 많다는 근거로 계산된다.

Group	Motifs	EC Score	PI Score
M/G1 phase	SFF MCM1'	0.037	0.434
	SWI5 MCM1'	0.076	-
G1 phase	MCM1' STE12	0.333	2.957
	MCB SCB	0.164	2.420
S phase	SFF MCM1' CCA	0.321	5.224
S/G2 phase	BAS1 SFF	0.178	-
G2/M phase	MCM1 SFF SFF'	0.060	1.056
	SFF SCB	0.089	1.673

표 1 유전 알고리즘에 의해 탐색된 모티프 조합: 세포주기의 각 단계별로 공동 작용한다고 추정되는 주요한 모티프 조합 결과

표 1은 유전 알고리즘을 이용하여 효모 세포주기의 각 단계에서 중요하게 공동 작용하는 모티프 조합들을 탐색한 결과이다. 여기 리스트된 모티프들은 BAS1, STE12를 제외하고 모두 세포주기에서 동작한다고 알려져 있다. 특히 기존 문헌에 SCB와 MCB는 G1 후기에 작용하여 세포주기제어(cell cycle control)에 관련하는 모티프 쌍으로 알려져 있다[Simon et al., 2001].

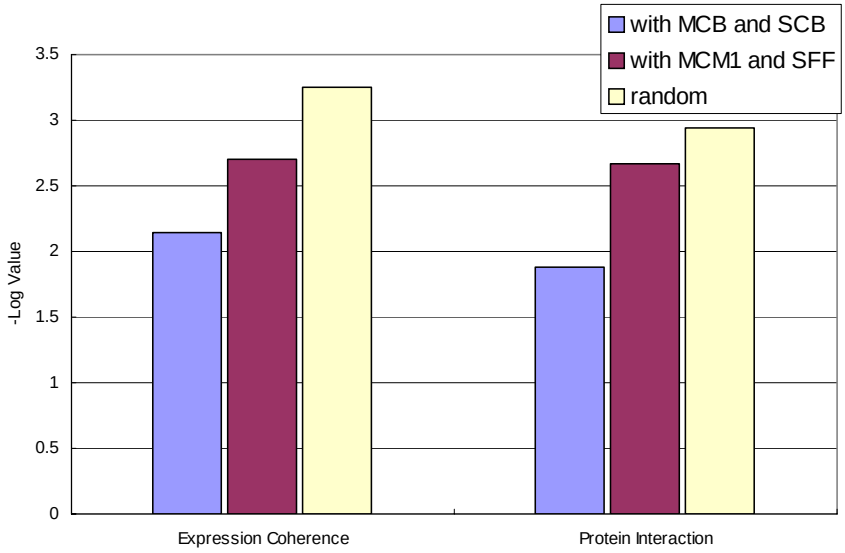

그림 11 모티프 조합과 랜덤 샘플링에 의한 EC 및 PI율 비교

다음은 공동 작용한다고 알려진 모티프 조합을 가진 유전자들과 랜덤하게 샘플링된 것과의 상대적 비교를 위하여 EC와 PI의 관한 측정을 해보았다. 그림 11은 공동 작용한다고 알려진 모티프 조합을 가진 유전자들에 대해서 모든 쌍의 유클리디안 거리의 평균과 PI확률을 랜덤하게 샘플링하여 얻은 값과 비교한 그래프이다. 상대적 비교를 위하여 각각의 막대는 EC 와 PI대해서 음의 로그값을 취한 것이다. 그래프에서 보이는 바와 같이

유전자를 랜덤하게 추출하여 얻어진 것에 비해서 모티프 조합을 가진 유전자에 의해 얻어진 EC와 PI값이 낮은 값을 가짐을 알 수 있다. 이러한 결과로 공통적인 모티프는 발현 프로파일과 단백질 상호작용에 비례적인 연관관계를 가지고 있음을 확인할 수 있었다.

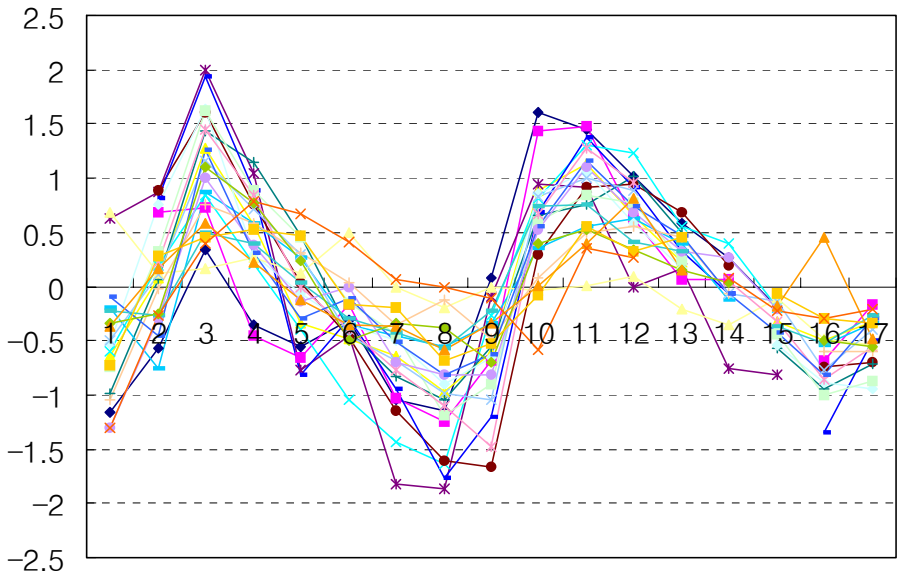

그림 12 MCB와 SCB 모티프를 가지는 유전자들의 발현 프로파일

그림 12는 G1 후기에 발현에 관련한다고 알려진 MCB와 SCB모티프 조합을 가지는 유전자들의 발현 프로파일을 보여주고 있다. 두 모티프들을 가지는 유전자들의 발현 패턴들이 매우 유사하다는 것을 알 수 있고 G1 후기 부분에서 발현됨을 확인할 수 있다.

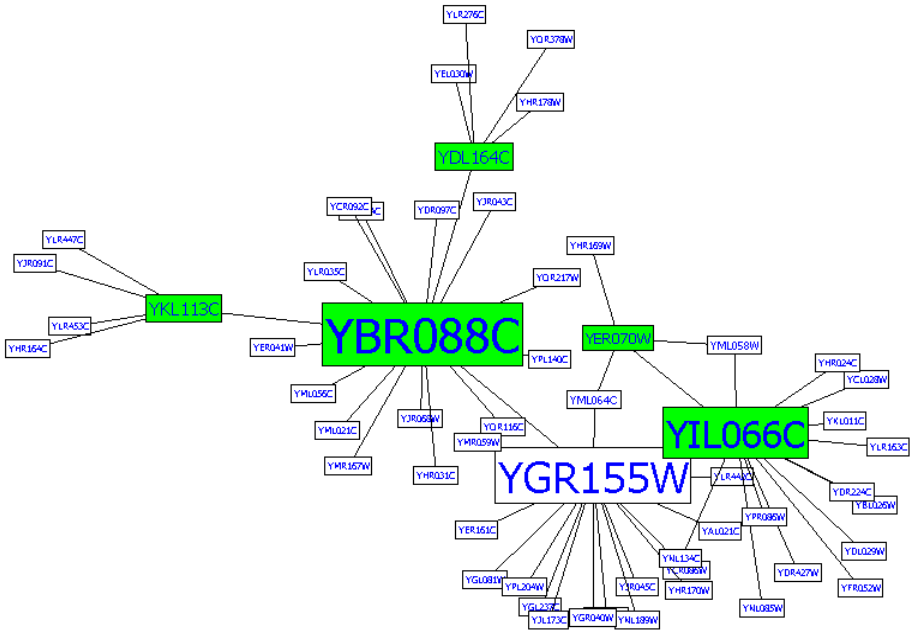

그림 13 상호작용 단백질 쌍을 기준으로 확장한 단백질 네트웍: MCB와 SCB 모티프를 가지는 유전자들에 있어서 상호 작용하는 단백질 쌍을 추출하여 얻은 네트웍

다음은 MCB와 SCB를 가지는 유전자의 상호 작용 증거를 단백질 상호 작용 네트웍을 통하여 알아보기로 한다. 그림 13은 G1 후기에 공동 작용한다고 알려져 있는 MCB와 SCB를 가지는 유전자 중에 상호작용을 하는 단백질을 기준으로 확장한 네트웍을 보여주고 있다. 각각의 사각형내의 이름은 ORF명을 나타내고 있다. 여기서 음영으로 채워진 ORF는 MCB와 SCB 목표 유전자중에 단백질 상호작용을 하는 ORF들이다. 이들은 모두 물리적으로 상호작용을 하는 것으로 YER070W (RNR1)은 실제로 MCB와 SCB의 상호적으로 조절 받는다는 증거를 Localization analysis 실험 결과를 통해서 알 수 있다[Simon et al., 2001]. RNR1은 Ribonuclotide reductase large

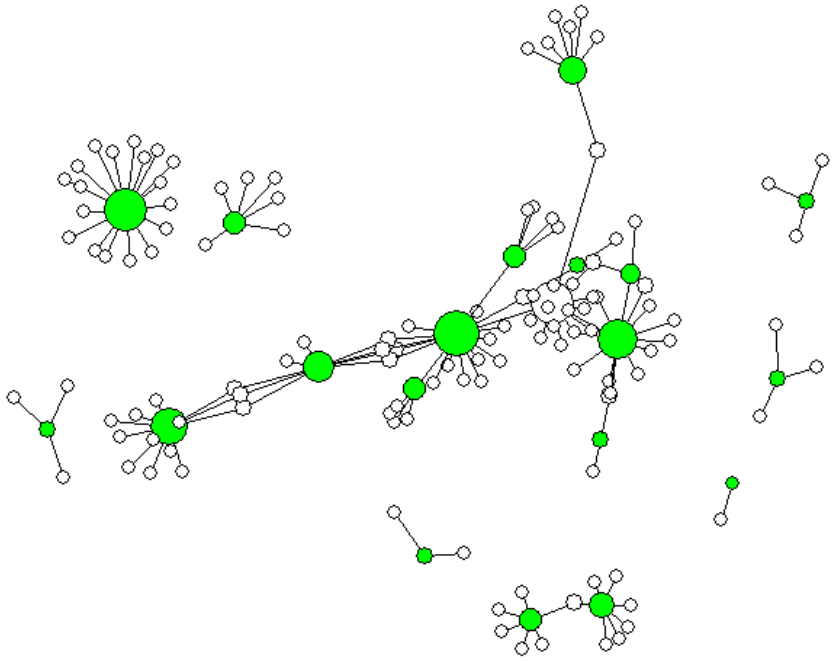

그림 14 MCB 와 SCB 모티프를 가지는 유전자들의 단백질 네트웍

subunit으로써 DNA 복제(replication)에 관련하는 유전자이다.

그림 14는 MCB와 SCB를 가지는 유전자 모두에 대하여 확장한 네트웍을 보여주고 있다. 전체적으로 보면 약 50%의 노드들이 하나의 네트웍 상의 복합체(complex)에 밀집되어 있음을 보여주고 있다. 이러한 결과는 공통 모티프를 포함하고 있고 발현 패턴이 비슷한 유전자들은 비슷한 기능을 수행할 수 있으며 그들끼리 서로 상호작용을 할 확률이 많다는 증거를 제시해 주고 있다. 또한 공통 모티프를 포함하고 있고 발현 패턴이 비슷한 유전자들이지만 단백질 상호작용 데이터에는 부재한 유전자 쌍들은 실제로 네트웍의 가운데 밀집되어 연결될 확률이 많을 것이다.

다음은 유전알고리즘과 Memetic 알고리즘의 학습의 성능을 적합도 그래프를 통해서 알아보기로 한다. 그림 15에서 위의 두 곡선은 최적해의 적합도를 나타내고 아래의 두 곡선은 적합도의 평균을 나타내고 있다. 이 그래프는 10번 수행하여 평균을 낸 그래프이다. Memetic 알고리즘에서 각각의 개체에 대한 지역탐색의 반복횟수는 20번이다. 이 결과, 그래프에서 보이는 바와 같이 유전 알고리즘에 의한 적합도 보다 Memetic 알고리즘에 의한 적합도가 최적해에 있어서 높음을 알 수 있다. 이는 지역탐색에 의한 효과가 반영된 것이라고 볼 수 있다.

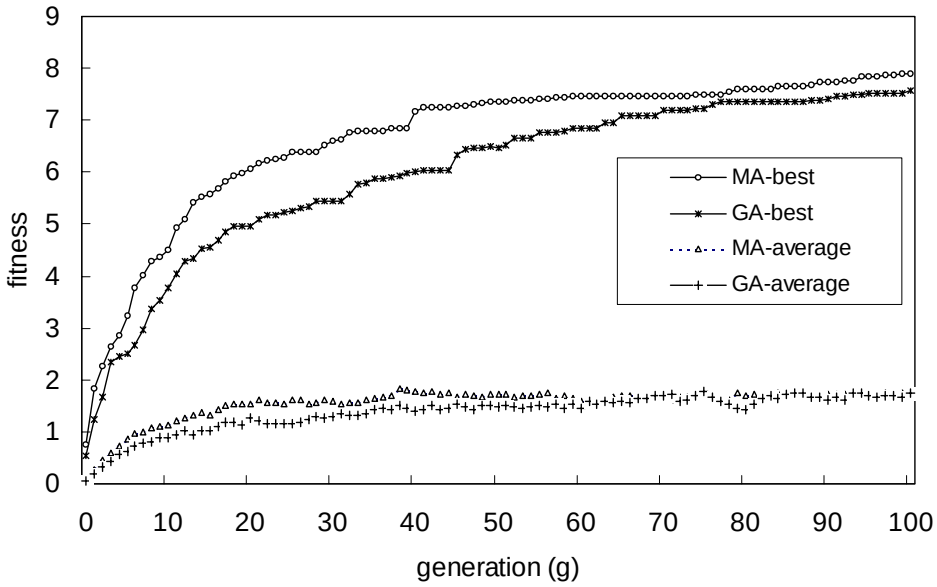

그림 15 유전알고리즘과 Memetic 알고리즘의 적합도 비교

그림 16은 MECS와 SSC의 각 항목이 적합도에 어느 정도의 영향을 미치는지 보여 주고 있다. 이 그래프에서 보면, EC 점수에 대해서 Memetic 알고리즘이 유전 알고리즘에 비해 높다는 것을 알 수 있다. 뿐만 아니라, SSC 항목 역시 높다는 것을 알 수 있는데, 이는 지역탐색에 의해 찾아진 모티프 조합들을 가지는 유전자들의 발현패턴들이 그렇지 않은 것에 비해서 군집화 될 확률이 높고 그룹 간에 분리가 더 잘된다는 것을 말해준다.

여기서 각 적합도 항목에 대한 가중치의 조정은 trade-off 인자 α 와 β 를 설정함으로써 가능하고 α와 β의 설정값에 따라 모티프 조합의 결과가 달라질 것이다. 만약 α에 치중하면 그룹 내의 발현패턴들이 좀 더 촘촘하게 밀집될 것이고, 반대로 β에 치중하면 그룹간의 발현패턴들이 서로 떨어져 있게 될 것이다.

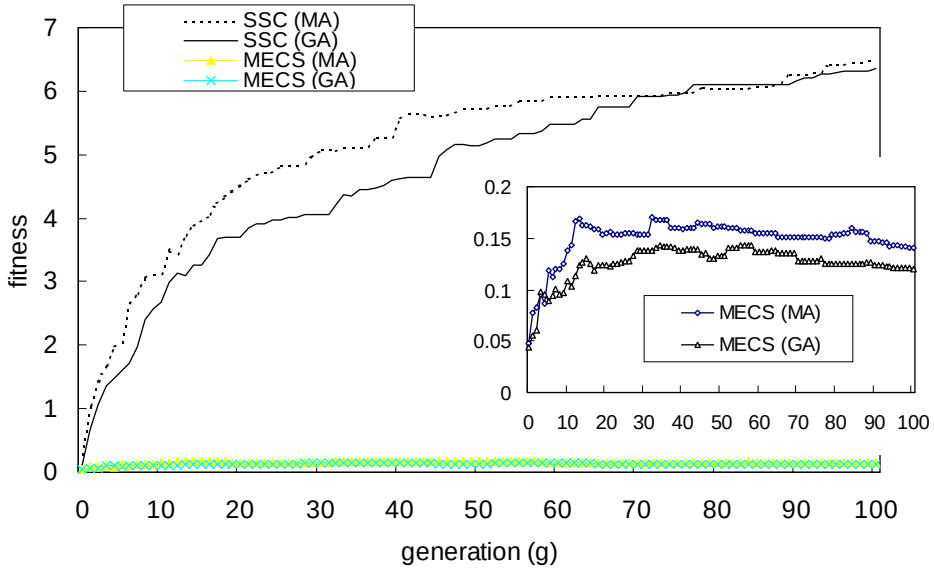

그림 16 유전알고리즘과 Memetic 알고리즘의 적합도 항목 비교

그림 17과 18은 PCA (principle component analysis)를 통하여 발현데이터를 이차원 공간상에 투영시켜 얻은 결과이다. 그림 17은 유전 알고리즘을 수행하여 얻어진 모티프 조합을 가진 유전자들의 발현 패턴들의 공간상의 투영결과를 보여주고 있다. 각각의 그룹은 특정 모티프 조합을 가지는 유전자들의 발현패턴을 포함하고 있다. 여기서 유전자 발현 패턴들은 비교적 그룹별로 나뉘어 있지만 그룹 3과 4는 서로 분리되어 있지 않고 또한, 그룹 1의 경우는 발현패턴이 비교적 분산되어 있다.

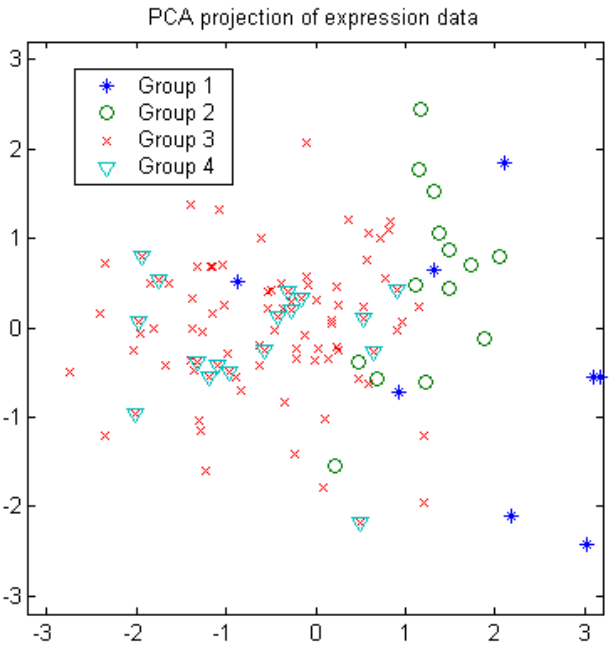

그림 17 PCA를 통하여 발현데이터를 이차원 공간상에 투영(a): 유전 알고리즘을 이용하여 얻어진 모티프 조합들을 가진 유전자들의 발현 패턴들

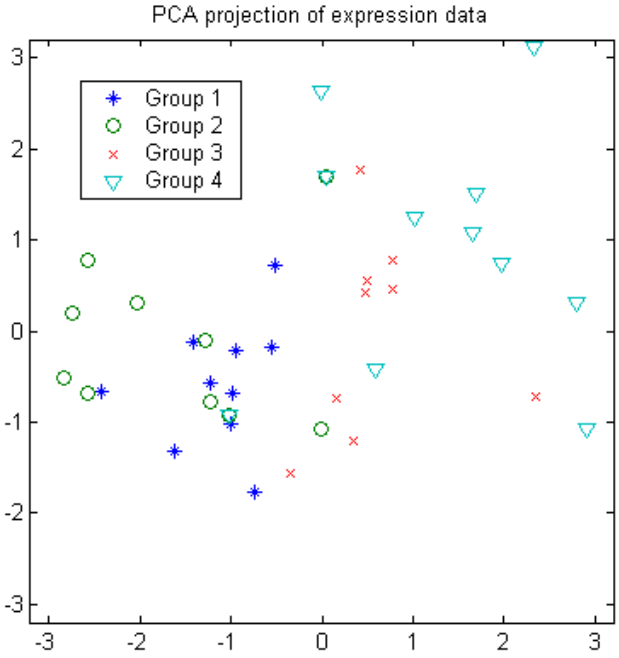

그림 18 PCA를 통하여 발현데이터를 이차원 공간상에 투영(b): Memetic 알고리즘을 이용하여 얻어진 모티프 조합들을 가진 유전자들의 발현 패턴들

그림 18은 Memetic 알고리즘에 의한 모티프 조합간의 목표 유전자 발현 패턴들의 분포를 보여주고 있다. 유전 알고리즘에 의해 얻어진 결과에 비해 그림 18의 결과는 비교적 그룹별로 서로 나뉘어 있으며, 그룹 내의 유전자 발현패턴들 역시 밀집되어 있다. 이러한 결과는 Memetic 알고리즘에 의한 지역탐색 기법을 통하여 그룹 내의 유전자들의 발현패턴들이 비교적 비슷하면서 서로 상이한 모티프 조합들을 찾아낸 것이라고 할 수 있다.

표 2는 4개의 마이크로어레이 실험을 통하여 얻어진 발현프로파일 데이터와 현재 알려지지 않은 모티프를 추가한 모티프 데이터를 이용하여 Memetic 알고리즘을 실행하여서 얻은 모티프 조합들의 결과이다. 여기서 모티프 조합 m74, m89, m59는 표 3의 인덱스에 나와 있듯이 MCB, SFF, MCM1에 해당하며 세포주기에서 중요한 모티프로 이미 알려져 있다.

Experiment	Motif Combination	EC Score
Cell-cycle	m187, m353	0.088
	m74, m68, m89	0.066
	m3, m89	0.190
	m59, m247	0.065
Heat-shock	m132, m209	0.045
	m308, m306, m133	0.088
	m185, m326	0.066
	m149, m175	0.018
Sporulation	m312, m301	0.177
	m268, m112, m28	0.038
	m155, m152	0.090
	m218, m109	0.110
Diauxic shift	m243, m29, m104	0.345
	m273, m49	0.466
	m314, m112	0.653
	m31, m151	0.377

표 2 각 마이크로어레이 실험에 대한 주요 모티프 조합 분석

Exp.	Motif Index	Motif Description
Cell-cycle	m187	m. organization of chromosome structure orfnum2SD n17
	m353	m. allantoin and allantoate transporters orfnum2SD n13
	m74	MCB: dna synthesis and replication orfnum2SD n2
	m68	m. deoxyribonucleotide metabolism orfnum2SD n10
	m89	SFF: FKH1_orfnumA2SD_n1
	m3	MET31-32 aminoacid biosynthesis orfnumA2SD n7
	m59	MCM1: CLB2 M Cluster orfnumA2SD n3
	m247	m pentose-phosphate pathway orfnum2SD n14
Heat-shock	m132	ALPHA2': Matalpha1 orfnum2SD n7
	m209	m. other energy generation activities orfnum2SD n12
	m308	m_RPE32: ribosomal proteins orfnum2SD n32
	m306	m_RPE17: ribosomal proteins orfnum2SD n17
	m133	ALPHA1': Matalpha1Spr orfnum2SD n1
	m185	CCA: organization of chromosome structure orfnum2SD n3
	m326	m. sugar and carbohydrate transporters orfnum2SD n6
	m149	m. metal ion transporters orfnum2SD n14
	m175	m. nutritional response pathway orfnum2SD n7
Sporulation	m312	m_RPE57: ribosomal proteins orfnum2SD n57
	m301	RAP1: ribosomal proteins orfnum2SD n1
	m268	m. phosphate transport orfnum2SD n8
	m112	m. ion transporters orfnum2SD n3
	m28	m. anion transporters orfnum2SD n20
	m155	ndt80(MSE): ndt80 orfnumA2SD n1
	m152	m. metal ion transporters orfnum2SD n26
	m218	m. lipid transporters orfnum2SD n10
	m109	m. homeostasis of other ions orfnum2SD n30

Diauxic shift	m243	m. other transport facilitators orfnum2SD n15
	m29	m. anion transporters orfnum2SD n22
	m104	m. g-proteins orfnum2SD n12
	m273	m. nucleotide transport orfnum2SD n9
	m49	m. c-compound carbohydrate transport orfnum2SD n18
	m314	m_RPE68: ribosomal proteins orfnum2SD n68
	m112	m. ion transporters orfnum2SD n3
	m31	m. anion transporters orfnum2SD n27
	m151	m. metal ion transporters orfnum2SD n25

표 3 주요 모티프들의 기술

표 3의 모티프 주석은 MIPS 기능별 분류에서 얻어진 것이고 m. 또는 m_으로 시작하는 명칭은 AlignACE에 의하여 추정되는 모티프임을 지칭한다. Pilpel의 결과에 의하면 Heat-chock 데이터에서는 RPE (ribosomal protein element)들이 관여할 것으로 추정되며, 이 결과와 어느 정도 일치함을 보여준다[Pilpel et al., 2001]. 그리고 sporulation에서는 sporulation 특이적 전사인자 Ndt80가 바인딩하는 MSE가 관여하는 것으로 알려져 있다[Jason et al., 2002]. MSE는 감수분열(meiosis)을 통하여 약 150여개의 유전자에 관련하고 있다고 보고 되고 있다. Diauxic shift에서 탐색된 모티프 조합은 모두 알려지지 않은 모티프들을 포함하고 있다. 이러한 조합들이 실제로 상호작용을 하는지는 실험에 의해서 검증되어야 할 것이다.

제5장 결론

본 논문에서는 유전자 발현 프로파일을 통하여 유전자 조절에 있어서 공동으로 작용하는 모티프 조합을 탐색하는 알고리즘을 제시하였다. 이러한 알고리즘으로서 최적화에 좋은 성능을 보이는 진화 알고리즘을 도입하였다.

진화 알고리즘의 적합도는 하나의 마이크로어레이 실험에 대하여 여러 개의 최적 모티프 조합들을 찾아내도록 설계되었다. 이를 위하여 하나의 개체는 여러 개의 모티프 조합들로 인코딩되었다. 따라서 알고리즘은 유전자 발현 프로파일들의 군집화를 수행하면서 동시에 최적 모티프 조합들을 탐색하였다. 이와 더불어, 알고리즘의 성능 개선을 위하여 진화 알고리즘 중 하나인 Memetic 알고리즘을 제안하였다. Memetic 알고리즘은 지역탐색이 수행되지 않는 유전 알고리즘에 비해서 높은 적합도 성능을 보였다. PCA를 수행하여 분석한 결과, Memetic 알고리즘에 의한 목표 유전자 프로파일들이 비교적 군집화가 잘 이루어졌음을 확인할 수 있었다.

실험 결과로써 먼저, 세포주기에 관한 마이크로어레이 데이터를 적용하였으며, 각 단계별로 어떤 모티프 조합들이 주도적인 영향을 미치는지 파악할 수 있었고, 기존의 문헌과 많은 일치를 보였다. 또한 제안된 알고리즘을 통하여 여러 마이크로어레이 데이터에 대하여 알려지지 않은 모티프 조합들을 제시하였다.

본 연구와 관련하여 앞으로 더 연구가 필요한 사항들을 살펴보면 다음과 같다. 첫째, 효모뿐만 아니라 다른 종에 대한 주요 모티프 조합들을 탐색하는 것이다. 효모는 이미 많은 연구가 되어 있기 때문에 제안한 알고리즘의 검증에 적합하였다. 현재 효모와는 별도로 초파리(Drosophila)에 대한 모티프 연구가 많이 진행되고 있는 추세이다. 따라서 초파리에 대하여 제안된 알고리즘으로 공동 작용하는 모티프 조합 탐색을 시도한다면 이 도메인에 있어서 다양한 생물학적 정보를 제공해 줄 수 있을 것이다.

둘째, 제안된 알고리즘은 다양한 지역탐색 방법과 다양한 파라미터 변환을 통하여 최적화 성능이 개선될 수 있다. 그 외에도 근사치를 구함으로써 지역탐색의 시간을 단축하는 방법과 적합도 측정에 있어서 발현 응집도 점수 외에 군집도를 측정하는 다른 방법을 적용하여 알고리즘의 성능을 개선할 수 있다.

본 연구에서 제시하고 있는 진화 알고리즘에 의한 모티프 조합 탐색 방법은 조절 네트웍 및 조절 모티프 연구에 유용하게 활용될 수 있으며, 진화 알고리즘 연구에 있어서 하나의 연구 도메인을 제시해 줄 것으로 기대한다.

참 고 문 헌

[Bailey and Elkan, 1995] Bailey, T. L. and Elkan, C., The value of prior knowledge in discovering motifs with MEME, *ISMB*, pp. 21-29, 1995.

[Brazma et al., 1998] Brazma, A., Jonassen, I., Vilo, J. and Ukkonen, E., Predicting gene regulatory elements in silico on a genomic scale, *Genome Research*, Vol. 8, No. 11, pp. 1202-1215, 1998.

[Bussenmaker et al., 2002] Bussemaker, H., Li, H. and Siggia, E., Regulatory element detection using correlation with expression, *Nat. Genet.*, Vol. 27, pp. 167-171, 2002.

[Cho et al., 1998] Cho, R. J., et al., A genome-wide transcriptional analysis of the mitotic cell cycle, *Mol. Cell*, Vol 2, pp. 65-73, 1998.

[Chu et al., 1998] Chu, S., et al., The transcriptional program of sporulation in budding yeast, *Science*, Vol. 282, pp. 699-705, 1998.

[Eisen et al., 1998] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci.*, USA 95, pp. 14863-14868, 1998.

[Freisleben and Merz, 1996] Freisleben, B. and Merz, P., New genetic local search operators for the Traveling Salesman Problem, *PPSN IV*, Vol. 141, pp. 890-900, 1996.

[Fujibuchi et al., 2001] Fujibuchi, W., Anderson J. S. J. and Landsman, D., PROSPECT improves *cis*-acting regulatory element prediction by integrating expression profile data with consensus pattern searches, *Nucleic Acids Research*, Vol. 29, No. 19, 2001.

[Goldberg, 1989] Goldberg, D. E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, New York, 1989.

[Holland, 1995] Holland, J. H., Adaptation in Natural and Artificial

Systems, The University of Michigan Press, Ann Arbor, 1975.

[Jason et al., 2002] Jason, S. L., David, S., Roger, T., Cynthia, W. and J. N. Mark G., Structure of the sporulation-specific transcription factor Ndt80 bound to DNA, *EMBO*, Vol. 21, No. 21, pp. 5721-5732, 2002.

[Lee et al., 2002] Lee, T., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. N., Harbison, C. T., Thompson, C. M., Simon, I., et al., Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, Vol. 298, pp. 824-827, 2002.

[Liu et al., 1995] Liu, S. J., Neuwald, A. F. and Lawrence, C. E., Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, *J. Am. Stat. Assoc.*, Vol. 90, pp. 1156-1170, 1995.

[Miller et al., 1995] Miller, B. L. and Goldberg, D. E., Genetic Algorithms, Selection Schemes and the Varying Effect of Noise, IlliGAL report, No. 95009. 1995.

[Moscato, 1989] Moscato, P., On Evolution, Search, Optimization, Genetic Algorithm and Martial Arts: Towards Memetic Algorithms, Technical Report C3P Report 826, Caltech Concurrent Computation Program, California Institute of Technology, 1998.

[Peter et al., 2002] Peter, J. P., Atul, J. B. and Isaac S. K., Comparing expression profiles of genes with similar promoter regions, *Bioinformatics*, Vol. 18, No. 12, pp. 1576-1584, 2002.

[Pilpel, 2001] Pilpel, Y., Sudarsanam, P. and Church, G., Identifying regulatory networks by combinatorial analysis of promoter elements, *Nat. Genet.*, Vol. 29, pp. 153-159, 2001.

[Roth et al., 1998] Roth, F., Hughes, P., Estep, J. D. and Church, G., Finding DNA regulatory motif within unaligned noncoding sequences

clustered by whole-genome mRNA quantitation, *Nat. Biotechnol.*, Vol. 16, pp. 939-945, 1998.

[Segal et al., 2003] Segal, E., Yelensky, R. and Koller, D., Genome-wide discovery of transcriptional modules from DNA sequence and gene expression, *Bioinformatics*, Vol. 19, Suppl., pp. i273-i282, 2003.

[Segal et al., 2001] Segal, E., Barash, Y., Simon, I., Friedman, N. and Koller, D., From promoter sequence to expression: a probabilistic framework, *RECOMB*, pp. 263-272, 2001.

[Simon et al., 2001] Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola T. S. and Young, R. A., Serial regulation of transcriptional regulators in the yeast cell cycle, *Cell*, Vol. 106, pp. 697-708, 2001.

[Sinha et al., 2000] Sinha, S. and Tompa, M., A statistical method for finding transcription factor binding sites, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, Vol. 8, pp. 344-354, 2000.

[Spellman et al., 1998] Spellman, P. T., et al., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell 9*, pp. 3273-3297, 1998.

[Tavazoie et al., 1999] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M., Systematic determination of genetic network architecture, *Nature Genetics*, Vol. 22, pp. 281-285, 1999.

Abstract

Transcriptional modules are essential to understand genetic regulatory networks. Developing a DNA microarray technology made it possible to simultaneously measure expression patterns of thousands of genes and made it easy to study regulatory mechanism.

In this paper, the method for search of synergistic motif combinations is presented and it uses both transcriptional motifs and gene expression profile. The synergistic motif combination is the set of motifs that affect expression patterns together. The search for putative regulatory motif combination is defined as a combinatorial optimization problem and an evolutionary algorithm is introduced to solve the problem effectively. Since evolutionary algorithms are based on not individual but population based search, it is highly possible to find an optimal solution of the problems which have the huge search space. Therefore, evolutionary algorithms are useful for searching synergistic motif combinations. In this paper, the additional evolutionary operator for local search is used to improve searching ability. The fitness function is the measure of clustering for gene expression profiles containing common motifs.

The presented method searched putative synergistic motif combinations over *Saccharomyces cerevisiae* gene expression. The experimental results showed valuable biological evidence for synergistic motifs corresponding to gene expressions.

Student Number: 2002-20634

감사의 글

이 논문을 정리하면서 그동안 격려해 주시고 도움을 주신 많은 분들께 감사의 인사를 드리고자 합니다.

먼저 지금까지 학문적인 열의와 열린 마음으로 저를 지도해 주신 장병탁 교수님께 감사드리며, 생물정보학 협동과정 학생들을 위해서 많은 힘을 써주신 이병재 교수님, 생물정보학 연구의 모범을 보여주신 김규원 교수님, 어설픈 제 논문을 세세하게 검토하시고 많은 조언과 충고를 아끼지 않으신 김주한 교수님, 그리고 수업을 통하여 많은 가르침을 주신 여러 교수님들께 감사드립니다.

그동안 여러모로 저에게 따뜻한 관심과 함께 든든한 버팀목이 되어주신 오석준 박사님께 감사드리며, 통계학에 관심을 가지게 해주신 양진산 박사님, 성실하게 여러 일을 도맡아 하시는 지선씨, 곁에 있기만 해도 믿음직한 성배, 연구에 대한 많은 조언을 해주었던 정호씨, 사소한 현상이라도 수식으로 표현할 것 같은 준식씨, 영균, 그리고 학문적 호기심이 많은 수동씨, 은석, 한 식구처럼 많은 이야기를 나누었던 하영, 연구실 살림살이를 묵묵히 꾸려온 선, 연구실의 진정한 만담가, 규백, 장민, 오랫동안 같은 연구 주제로 도움을 많이 준 동연, 수용, 항상 연구실 생활의 활력을 심어준 진우, 엑셀의 귀재 병희, 입학 동기 본웅씨, 연구실 홍일점 이화씨, 시스템생물학 연구에 입문한 상근, 어떤 일이든 딱부러지게 처리하는 재홍, 다재다능한 승준, 여러 학술행사를 무난히 처리한 호진, 취향이 비슷한 인희, 그리고 연구실 일이라면 솔선수범하는 민호, 친동생처럼 잘 따라준 화진, 시스템과 프로그램을 착실하게 맡아서 해준 기루, 그리고 연구실 선후배, 호규를 비롯한 대학원 동기들 모두에게 감사의 뜻을 전하고 싶습니다.

끝으로 무엇보다도 저를 믿고 따뜻한 사랑과 헌신적인 뒷바라지를 해준 가족에게 감사드립니다. 이 논문을 사랑하는 부모님, 그리고 아내 은경에게 바칩니다.