마이크로어레이 기반 miRNA 모듈 분석을 위한 하이퍼망 분류 기법

(Hypernetwork Classifiers for Microarray-Based miRNA Module Analysis)

김 선 [†] 김 수 진 [††] 장 병 탁 [†††]

(Sun Kim) (Soo-Jin Kim) (Byoung-Tak Zhang)

요 약 마이크로어레이는 분자 생물학 실험에 있어 중요한 도구로 사용되고 있으며, 마이크로어레이 데이타 분석을 위한 다양한 계산학적 방법이 개발되어 왔다. 그러나, 기존 분석방법은 주어진 조건에 영향을 주는 개별 유전자를 추출하는 데 강한 반면, 유전자 간의 복합작용에 의한 영향을 분석하기 힘들다는 단점을 가지고 있다. 하이퍼망 모델은 생물학적인 네트워크 작용을 모방한 구조이며, 계산과정에서 요소간의 복합작용을 직접 고려하기 때문에 기존 방법에서 다루기 힘들었던 요소간 상호작용 분석이 가능하다는 장점을 가진다. 본 논문에서는 마이크로어레이 데이타를 기반으로 microRNA (miRNA) 프로파일 분석을 위한 하이퍼망 분류 기법을 소개한다. 하이퍼망 분류기는 miRNA 쌍을 기본 요소로 하여 진화 과정을 통해 miRNA 분류 데이타를 학습한다. 학습된 하이퍼망으로부터 유의하다고 판단되는 miRNA 모듈을 쉽게 추출할 수 있으며, 사용자는 추출된 모듈의 유의미성을 직접 판단할 수 있다. 하이퍼망 분류기는 암 관련 miRNA 발현 데이타 분류 실험을 통해 91.46%의 정확도를 보임으로써 기존 기계학습 방법에 비해 뛰어난 성능을 보여주었으며, 하이퍼망 분석을 통해 생물학적으로 유의한 miRNA 모듈을 찾을 수 있음을 확인하였다.

키워드 : 하이퍼망, miRNA 모듈 분석, 마이크로어레이, 데이타 분류

Abstract High-throughput microarray is one of the most popular tools in molecular biology, and various computational methods have been developed for the microarray data analysis. While the computational methods easily extract significant features, it suffers from inferring modules of multiple co-regulated genes. Hypernetworks are motivated by biological networks, which handle all elements based on their combinatorial processes. Hence, the hypernetworks can naturally analyze the biological effects of gene combinations. In this paper, we introduce a hypernetwork classifier for microRNA (miRNA) profile analysis based on microarray data. The hypernetwork classifier uses miRNA pairs as elements, and an evolutionary learning is performed to model the microarray profiles. miRNA modules are easily extracted from the hypernetworks, and users can directly evaluate if the miRNA modules are significant. For experimental results, the hypernetwork classifier showed 91.46% accuracy for miRNA expression profiles on multiple human cancers, which outperformed other machine learning methods. The hypernetwork-based analysis showed that our approach could find biologically significant miRNA modules.

Key words : hypernetworks, miRNA module analysis, microarrays, classification

· 본 연구는 과학기술부 국가지정연구실(NRL) 사업 및 산업자원부 차세대 신기술 개발 사업의 분자 진화 컴퓨팅(MEC) 과제에 의하여 일부 지원되었다.

† 학생회원 : 서울대학교 컴퓨터공학부
skim@bi.snu.ac.kr
†† 학생회원 : 서울대학교 생물정보학 협동과정
sjkim@bi.snu.ac.kr
††† 종신회원 : 서울대학교 컴퓨터공학부 교수
btzhang@bi.snu.ac.kr

논문접수 : 2008년 1월 21일
심사완료 : 2008년 4월 22일

정보과학회논문지 : 소프트웨어 및 응용 제35권 제6호(2008.6)

1. 서 론

마이크로어레이를 이용한 유전자 발현 프로파일 분석은 분자 생물학 분야에서 가장 중요한 접근 방법의 하나로 사용되고 있다[1]. 기존 기법은 한번에 한 개 또는 소수의 유전자들을 대상으로 한 측정 방법이었던데 반해, 마이크로어레이 기술은 동시에 수천 개의 유전자 발현을 측정할 수 있다. 생체 내에서 발생하는 생물학 메커니즘은 여러 다양한 유전자들이 서로 영향을 주고받으며 전체적으로 조절되므로, 전체 유전체 수준에서 발현 양상을 총체적으로 관찰할 수 있는 마이크로어레이 분석은 필수적이라 할 수 있다. 따라서 이는 암과 같은 특정 질병 메커니즘을 분자 레벨에서 분석하는 도구로서 널리 이용되고 있기도 하다. 한편, 최근 들어 개별 유전자의 전체적인 분석보다는 생물학적인 모듈 단위로의 분석이 암 조절 메커니즘을 밝히는데 중요하다는 사실이 밝혀졌다[2]. 이러한 생물학적 모듈 단위의 분석은 개별 유전자의 발현 정도는 의미가 없더라도 두 개 이상 조합하여 상호 조절 인자로서 발현에 영향을 주는 원인을 파악할 수 있다는 점에서 생물학적으로 중요한 의미를 가진다. 그러나, 상호 조절 유전자 모듈을 파악하기 위해 암과 관련한 생물학적 경로(pathway)를 유추하는 것은 결코 쉽지 않은 문제다[3].

암 관련 유전자를 분석하기 위한 대표적인 마이크로어레이 분석 방법은 개별 유전자와 특정 샘플 등 사이의 상관관계를 측정하는 것이다. 상관관계가 높은 유전자는 마이크로어레이 발현 패턴에서 암 또는 일반 조직(tissue)를 구별하는 척도가 된다. 따라서 상관관계를 측정하는 방법은 질병과 관련된 특정 발현 패턴을 분석하기 위한 방법으로써 널리 사용되고 있다. 그렇지만 이 방법은 유전자간의 상호관계가 아닌 개별 유전자 단위의 분석을 수행하기 때문에 앞에 설명한 바와 같은 유전자 모듈 단위로의 분석에는 적절하지 못하다.

기계학습을 이용한 마이크로어레이 분석 방법이 최근들어 각광받고 있다. 유전자 발현데이타의 클러스터링 방법으로는 베이지안망이 대표적이며, 특히 발현 데이타의 분류를 위해서는 최대 마진 분류를 기반으로 한 지지벡터머신(support vector machines) 및 부스팅 기법이 다수 사용되고 있다[4,5]. 여기에서 마진 분류 기법은 유전자 발현 데이타의 샘플들이 마진을 기준으로 분리되도록 하는 경계를 찾는 방법이다. 그러나, 이러한 통계기법에 기반한 기계학습 방법의 한계는 비선형 분류 문제에서의 최적해를 찾는 그 자체에 있다. 기본적으로 개별 유전자를 기준으로 하여 고차원에 사상시키는 방법에 의한 분류방법이기 때문에 유전자들 사이에 존재하는 관계를 파악하기 쉽지 않고, 그들이 복합작용에 의해 일어나는 역할 역시 쉽게 분석하기 힘들다. 최근에 유전자를 개별단위가 아닌 생물학적 모듈로서 간주하여 유전자 발현 데이타 분석을 시도한 사례가 보고되었지만[6,7], 마이크로어레이로부터 직접 여러 유전자 모듈의 상호 유도 작용을 분석하거나 유추하는 작업은 여전히 어려운 문제로 남아 있다.

microRNA (miRNA)는 약 21~25 nucleotide길이의 small RNA의 한 종류로서 유전자 발현을 제어하는 중요 조절인자 중 하나로 알려져 있다. miRNA는 목적 mRNA의 3'-untranslated region (3'-UTR)에 불완전한 상보 결합을 통해 유전자의 발현을 억제, 전사 후 번역을 방해하게 된다(그림 1). 이처럼 miRNA는 다양한 생물학적 과정에 관여하며 유전자 조절 네트워의 구성요소로 주목을 받고 있다. 최근에는 마이크로어레이, bead assay, quantitative PCR, serial analysis gene expression (SAGE) 등과 같은 분석 방법을 이용하여 miRNA의 기능을 연구하기 위한 대량의 발현 데이타가 산출되고 있으며, 이를 이용하여 암과 miRNA의 상관성을 분석하기 위한 노력이 최근에 이루어지고 있다. 특정 miRNA에 의해 암 유전자의 발현 패턴이 변화되어 암 발생을 유도한다고 보고되고 있으며, 따라서 miRNA의 이상 발현 패턴은 암 진단의 주요 잣대로 인식되고 있다[8]. 한편, 생체 내에서 발생하는 생물학적 메커니즘은 매우 복잡하고 다양한 과정에 의해 발생되기 때문에 miRNA간의 복합작용을 분석하는 것은 의미 있는 일이라고 할 수 있다.

본 논문에서는 miRNA 발현 프로파일 데이타로부터 암과 연관된 유전자 모듈을 찾아내기 위한 하이퍼망 분류방법을 제시하고자 한다. 하이퍼망[9,10]은 랜덤 하이퍼그래프 모델의 한 종류로서 하이퍼그래프에 가중치를 가진 간선을 사용한다. 생체분자 네트웍은 세포내의 환

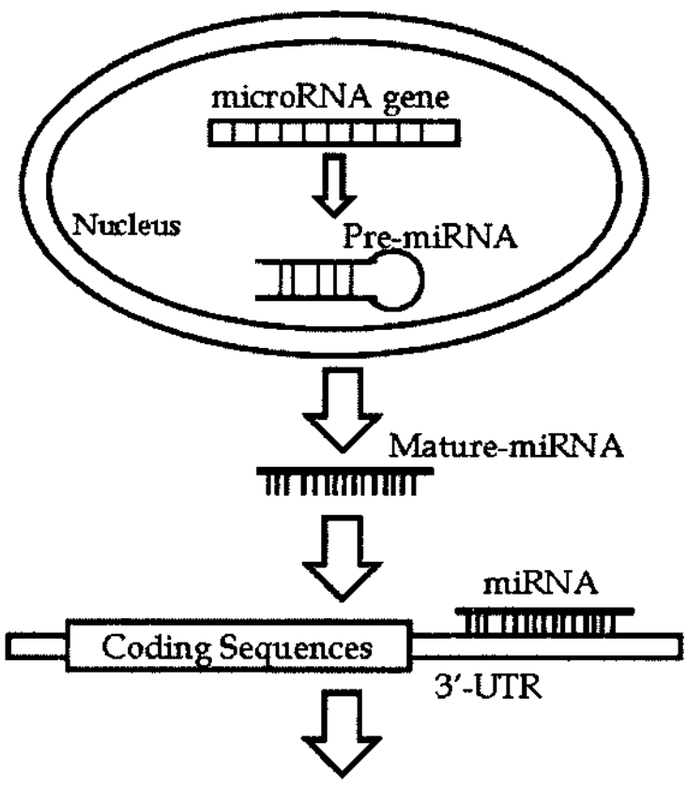


그림 1 miRNA 조절 기작

경이 변화함에 따라 이에 빨리 적응하면서도 안정성을 유지한다는 특징을 가지는데, 하이퍼망은 이러한 생체분자 네트워의 특징을 모방하여 만들어진 모델이다. 환경에 대한 적응력 및 안정성과 같은 속성은 암 조절 메커니즘과 같은 복잡하고 큰 규모의 생물학 문제를 분석하는데 유용할 수 있다. 하이퍼망의 또 다른 특징은 예측한 결과에 대해 그 이유를 사람이 이해할 수 있는 형태로 쉽게 분석할 수 있다는 점이다. 이러한 하이퍼망 구조는 고차원 자질(feature)을 이용, 최적의 조합 구성 및 가중치를 구하는데 사용되며, 이를 위한 기법으로 본 논문에서는 진화 학습 알고리즘을 제시한다.

높은 분류 정확도와 함께 암 발현 패턴과 관련된 miRNA 페어를 찾기 위한 실험 수행결과, 하이퍼망 기반 분류 기법은 기존 기계학습 기법인 신경망 또는 지지벡터머신과 대등한 분류 성능을 보여주었으며, 결정트리와 나이브베이즈 보다 뛰어난 성능을 보여주었다. 그리고 학습된 하이퍼망 분류기에서 암과 연관이 있다고 예측된 miRNA 모듈을 추출하였으며, 목적 유전자에 대한 통계적 분석 및 온톨로지 분석 등을 통해 추출된 miRNA 모듈이 유의미한 결과가 될 수 있음을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 마이크로어레이 데이타의 분류 및 분석을 위해 사용한 하이퍼망 모델의 구조 및 이론적 배경을 설명한다. 3장에서는 하이퍼망 모델을 이용해 마이크로어레이 데이타를 분류하기 위한 진화 학습 과정을 설명한다. 4장은 miRNA 발현 프로파일을 하이퍼망 분류기에 적용한 실험 결과 및 분석 내용을 다루며, 마지막으로 5장에서 본 논문의 결론을 내리고자 한다.

2. 하이퍼망 분류기

본 장에서는 마이크로어레이 데이타를 분류하고 그 결과를 분석하기 위한 구조인 하이퍼망 모델을 설명한다. 생체분자 네트워 구조에 기반한 하이퍼망은 그래피컬 모델의 일종으로서 하이퍼그래프의 간선에 가중치가 붙은 형태를 취한다. 하이퍼그래프는 널 값을 가지지 않는 노드들을 연결하는 간선들로 구성되는 무방향성 그래프(G)이다[11]. 즉, $G = \{X,E\}$에서 $X = \{X_1, X_2, ..., X_n\}$, $E=\{E_1, E_2, ..., E_m\}$이며, $E_i = \{x_{i1}, x_{i2}, ..., x_{ik}\}$이다. 여기에서 E_i를 하이퍼간선이라고 부른다. 수학적으로, E_i는 집합이며 그 크기(차수)는 1이상이다. 다시 말해 일반적인 그래프는 간선이 최대 2개까지의 정점을 연결할 수 있는데 반해, 하이퍼간선은 2개 이상의 간선들간의 연결이 가능하다. 이제 차수 k의 하이퍼간선은 k-하이퍼간선이라 하겠다. 이러한 하이퍼간선의 정의는 그래프 이론에서 정의되는 수학적인 기법을 그대로 사용할 수 있으면서 네트워을 표현하는데 있어 더 많은 자유를 허용하는

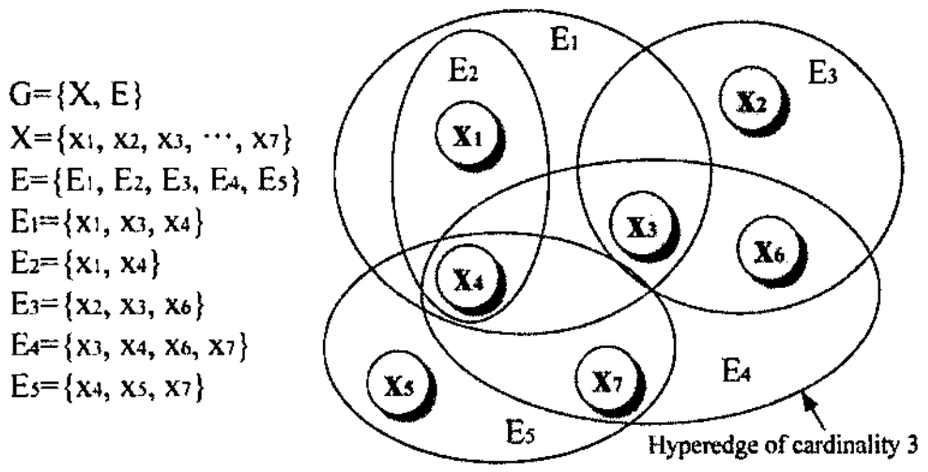


그림 2 하이퍼그래프의 예

이점을 가지게 된다. 그림 2는 7개의 정점 $X = \{X_1, X_2, ..., X_7\}$과 각기 다른 차수를 가지는 5개의 하이퍼간선 $E = \{E_1, E_2, ..., E_5\}$로 구성되는 하이퍼그래프의 예를 보인 것이다.

하이퍼망은 앞에서 설명한 하이퍼그래프의 각 하이퍼간선에 가중치를 할당하여 일반화한 형태로써, 가중치에 의해 각 정점 집합이 얼마나 강하게 연결되어 있는지를 표현할 수 있게 되었다. 하이퍼망은 하이퍼그래프에 가중치 W가 추가된 $H = (X, E, W)$로 정의되며, 여기에서 $X = \{X_1, X_2, ..., X_n\}$, $E = \{E_1, E_2, ..., E_m\}$, $W = \{w_1, w_2, ..., w_m\}$이다. 따라서, 하나의 k-하이퍼망은 정점의 집합 X와 $X[k]$의 부분집합인 E, 그리고 하이퍼간선의 가중치 집합인 W로 구성되며, 여기에서 $E = X[k]$는 구성요소가 정확히 k개인 X의 부분집합을 말한다. 만약 모든 하이퍼간선 E_i의 차수가 k라면, 이를 k-uniform 하이퍼망이라고 한다. 예를 들어, 일반 그래프는 $w_i = 1$인 2-uniform 하이퍼망이다.

생체분자 네트워 관점에서 보면 하이퍼망을 구성하는 하이퍼간선은 모듈, 모티프(motif), 회로(circuit)와 같은 빌딩 블록들에 해당한다고 볼 수 있다[12]. 특히, 높은 가중치를 가지는 하이퍼간선은 생물학 문제에서 중요한 역할을 수행한다고 말할 수 있을 것이다. 이러한 관점에서 볼 때, 하이퍼망 구조는 복합 또는 상호 작용을 하는 생물학 모듈을 분석하는 틀로 사용될 수 있다.

학습 모델에서의 학습 과정은 주어진 데이타 셋을 저장하는 과정으로 볼 수 있으며, 테스트 과정은 특정 예제에 의해 저장된 데이타를 검색하는 과정이라고 말 할 수 있다. 이런 관점에서 하이퍼망은 이론적으로 확률 메모리로써 사용될 수 있다. 하이퍼망의 에너지를 $E(x^{(n)};W)$라고 하자. 여기에서 $x^{(n)} \in D$는 하이퍼망에 저장된 n번째 데이타이며, W는 하이퍼망의 파라미터, 즉, 하이퍼간선의 가중치를 말한다. 이 때, 하이퍼망으로부터 데이타가 생성될 확률은 Gibbs 분포 가정 하에 다음과 같이 주어진다.

$$P(\mathbf{x}^{(n)} \mid w) = \frac{1}{Z(W)} \exp\left\{-\mathrm{E}(\mathbf{x}^{(n)};W)\right\},$$

여기에서 exp{-E(x$^{(n)}$;W}는 Boltzmann 요소이며 $Z(W)$는 정규화 요소이다.

데이타를 분류하는 문제에 있어서 데이타는 자질의 집합 x_i 및 클래스 y로 구성된다. 즉, $(\mathbf{x},y)\in D$이다. 하이퍼망 분류기는 위에 정의한 하이퍼망 구조, 즉 정점의 집합 X에 정점 y를 추가한 형태이기 때문에 결합 확률 $P(\mathbf{x},y)$를 다음과 같이 나타낼 수 있다.

$$P(\mathbf{x},y)=\frac{1}{Z(W)}\exp\{-\mathrm{E}(\mathbf{x},y;W)\}$$

분류기는 주어진 입력에 대하여 각 클래스에 대한 조건부 확률을 계산하여 가장 높은 확률 값을 갖는 클래스를 예측 값으로 리턴해 주는 장치이다. 이는 다음과 같은 수식으로 형식화 할 수 있다.

$$y^*=\arg\max_y P(y\mid\mathbf{x})=\arg\max_y\frac{P(\mathbf{x},y)}{P(\mathbf{x})}, \quad (1)$$

여기에서 $P(\mathbf{x},y)=P(y|\mathbf{x})P(\mathbf{x})$이고 y는 클래스를 의미한다. 분류기는 정확한 확률 값을 구하는 것이 중요하기 보다는, 각 클래스 간의 차이를 보고 결과를 예측하는 형태이기 때문에, $P(\mathbf{x})$는 생략이 가능하며, 따라서 수식 (1)은 다음과 같이 정리할 수 있다.

$$\begin{aligned}y^*&=\arg\max_y\frac{P(\mathbf{x},y)}{P(\mathbf{x})}=\arg\max_y P(\mathbf{x},y)\\&=\arg\max_y\frac{1}{Z(W)}\exp\{-\mathrm{E}(\mathbf{x};W)\}\\&=\arg\max_y\exp\{-\mathrm{E}(\mathbf{x};W)\}\\&=\arg\max_y-\mathrm{E}(\mathbf{x};W)=\arg\min_y\mathrm{E}(\mathbf{x};W).\end{aligned}$$

에너지 함수 $E(\mathbf{x};W)$는 선형 함수, sigmoid 함수, Gaussian 함수 등 여러가지 형태로 정의될 수 있다. 종합해 보면, 하이퍼망은 하이퍼간선 및 하이퍼간선의 가중치로 구성되는 일련의 규칙 집단(population)을 사용하여 특정 데이타 집합을 표현하는 확률 모델이라고 할 수 있다.

3. 하이퍼망 학습 알고리즘

2장에서 정의한 하이퍼망 분류기는 에너지 함수 E를 최소화하는 클래스 y를 선택하는 것이다. 여기에서 학습 과정은 주어진 데이타를 표현할 수 있도록 하이퍼간선의 가중치를 조절하는 것이다. 본 장에서는 앞에서 정의한 하이퍼망 분류기를 기반으로 분류 성능을 최대화하는 하이퍼망 분류기를 생성하기 위한 진화 학습 기법을 제시한다.

진화연산 기반의 하이퍼망 학습을 수행하기 위한 가정은 다음과 같다. 하나의 하이퍼망 분류기는 한 해집단을 나타내며, 한 하이퍼망 분류기를 구성하는 하이퍼간선은 개체이다. 여기에서 개체는 자질과 클래스의 조합으로써 주어진 자질의 조건이 만족하면 해당 클래스가 된다는 규칙을 의미하게 된다. 따라서 한 해집단은 각 개체에 의해 정의되는 규칙의 앙상블로 분류 예측을 하는 시스템이 된다. 이는 유전자 프로그래밍과 유사한 개념이다. 한편, 하이퍼간선의 가중치는 한 집단 안에서 똑같은 모양의 개체를 가질 수 있도록 함으로써 표현될 수 있으며, 따라서 하이퍼망의 진화학습은 분류 성능을 최대한으로 하는 방향으로 개체 수를 조절하는 문제가 된다. 그림 3은 하이퍼간선이 해집합의 개체로 표현되는 형태를 예로 보인 것이다.

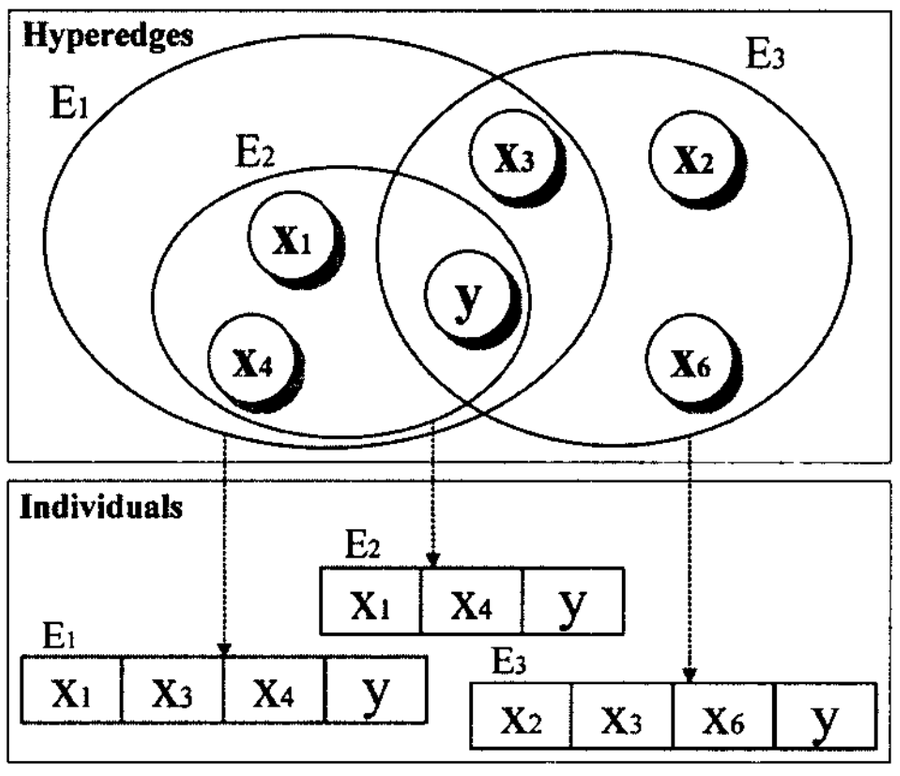


그림 3 진화연산 기반 하이퍼망 학습에서의 개체 표현의 예

초기 집단을 구성하기 위한 방법으로 본 논문에서는 랜덤 그래프 모델을 활용한다. 랜덤그래프는 랜덤 과정에 의해서 생성된 그래프를 말한다[9]. 초기해 구성을 위해 랜덤 그래프 모델을 사용하는 이유는 다음과 같다.

k-하이퍼그래프에서 가능한 하이퍼간선의 개수는 다음과 같이 정의된다.

$$|E|=C(n,k)=\frac{n!}{k!(n-k)!},$$

여기에서 $n=|X|$이다. 만약 가능한 모든 그래프의 집합을 Ω라고 한다면, 그 크기는 $|\Omega|=2^{C(n,k)}$가 된다. 따라서 k 및 n이 증가함에 따라서 $|\Omega|$은 기하급수적으로 증가한다. 따라서 조합의 경우의 수가 폭발적으로 증가하는 문제를 해결하기 위해 랜덤 그래프에 기반한 방법을 해집합 구성을 위해 사용한다. 기존 하이퍼망과 구별하기 위해 랜덤 그래프에 의해 만들어진 하이퍼망을 이제부터 랜덤 하이퍼망이라고 하겠다.

랜덤 그래프는 가능한 모든 그래프 집합으로부터 동등한 확률로 그래프를 랜덤하게 추출한다. 여기에서 확

률 공간을 (Ω, F, P)이라 정의할 수 있으며, Ω는 모든 그래프의 집합, F는 Ω의 모든 부분집합의 패밀리, 그리고 Ω에 속하는 모든 원소 a에 대한 확률은 다음과 같이 할당된다.

$$P(\omega) = 2^{-C(n,k)}.$$

정의된 확률공간은 $C(n,k)$개 바이너리 공간의 곱으로써 표현될 수 있으며, 한 코인을 $C(n,k)$번만큼 던져서 얻은 결과(베르누이 시행)로 해석될 수 있다. 랜덤 하이퍼망은 이항 랜덤 그래프 과정에 의해 생성될 수 있으며, 실수 $p(0 \le p \le 1)$에 대한 이항 랜덤그래프 $\Psi(n,p)$는 앞서 설명한 확률공간 및 다음 확률에 의해 정의된다.

$$P(\Psi) = p^{|E(\Psi)|}(1-p)^{C(n,k)-|E(\Psi)|},$$

여기에서 $|E(\psi)|$는 ψ의 간선의 개수를 말하며, 랜덤 하이퍼망은 이와 같은 랜덤 하이퍼그래프 과정을 반복함으로써 만들어진다.

1. 하이퍼망을 초기화한다. 즉, $H = (X, E, W) = (\phi, \phi, \phi)$.
2. 확률 p로 학습 데이터 $\mathbf{x}$를 선택한다.
3. 다음 과정을 이용해 하이퍼망 $H' = (X', E', W')$를 생성한다.
   - 랜덤 하이퍼그래프 과정에 의해 $\mathbf{x}$로부터 차수 k의 개체(하이퍼간선) E_i를 생성한다.
   - $E' \leftarrow E' \cup \{E'\}$.
   - $W' \leftarrow W' \cup \{w_i \mid w_i = w_{init}\}$, w_{init}는 초기 가중치.
   - $X' \leftarrow X' \cup \{x_j \mid x_j \in E_i\}$
4. $H \leftarrow H \cup H'$.
5. 해집합이 채워질 때까지 2~4번 과정을 반복한다.

그림 4 초기 해집합 구성을 위한 랜덤 하이퍼망 생성 알고리즘

그림 4는 랜덤 하이퍼 그래프 과정에 의해 초기 해집합, 즉 랜덤 하이퍼망을 만드는 알고리즘을 설명한 것이다. 비어있는 하이퍼망으로부터 시작해, p의 확률로 선택된 학습 데이타 x로부터 H'가 만들어지며 이 과정이 반복 수행된다. H' 생성을 위해 $(1-p)$의 확률로 학습 데이타를 이용하지 않은 랜덤 H'을 생성할 수도 있으며, 이는 해집합의 다양성에 영향을 준다. 생성된 H'에 대해 하이퍼간선 E'의 개체들, 즉 가중치 w_{init}만큼이 초기 해집합에 추가된다. 랜덤 하이퍼망 H의 업데이트는 미리 정의된 해집합의 크기가 채워질 때까지 반복된다. 랜덤 하이퍼망의 생성은 문제에 대한 분류 성능을 떨어뜨리지 않으면서, 가능한 조합의 기하급수적인 증가로 야기되는 계산상의 복잡도, 즉 해집합의 크기를 효과적으로 줄여주는 역할을 수행하게 된다.

그림 5는 본 논문에서 최적 하이퍼망을 구하기 위해

1. 그림 4에 의해 랜덤 하이퍼망을 생성한다.
2. 데이터 $(\mathbf{x}, y)$를 학습데이터로부터 선택한다.
3. 현재 해집합을 이용해 데이터 $\mathbf{x}$를 분류하고, 예측된 클래스 y^*를 얻는다.
4. 만약 $y \neq y^*$일 경우, 다음 과정을 이용해 해집합을 업데이트한다.
   - $c_{E_i} \leftarrow c_{E_i} + \Delta c_{E_i}$, c_{E_i}는 하이퍼간선 $E_i \in E(\mathbf{x}, y)$에 해당하는 개체들의 수.
   - 확률분포를 유지하기 위해 모든 하이퍼간선을 정규화한다.
5. 정지 조건이 만족할 때까지 2~4과정을 반복 수행한다.

그림 5 하이퍼망 분류기를 위한 진화 학습 알고리즘

사용한 진화 학습 알고리즘을 보인 것이다. 랜덤 하이퍼망으로부터 시작해 학습 데이타 (x,y)가 관찰되면 현재 해집합을 기준으로 분류작업을 수행하며, 이에 대한 결과로 클래스 y^*가 얻어진다. 만약, y^*가 정답이라면, 현재 해집합은 최적의 상태라고 간주하여 수정되지 않는다. 만약 y^*가 오답이라면, $E_i \in E(\mathrm{x},y)$에 대해 일정한 비율만큼 해당 개체를 증가시킨다. 한편, 앞에 정의한 하이퍼망은 학습 데이타에 대한 확률 메모리 모델이기 때문에 전체 확률 분포를 유지하기 위해 업데이트가 일어날 때마다 정규화 과정을 거치게 된다.

여기에 제시한 하이퍼망의 학습 알고리즘은 지역 최적해로 수렴하며, 진화학습 과정은 최적해를 찾기 위해 분류 오류를 줄이는 방향으로 경사도 탐색(gradient search)을 하는 것과 유사하다. 데이타 (x,y)는 $\mathrm{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, $y \in \{0,1\}$이고, 에너지 함수 $\mathrm{E}(\mathrm{x}^{(n)}; W)$는 sigmoid 함수라고 가정하자.

$$\mathrm{E}(\mathbf{x}; W) = \frac{1}{1 + \exp(-f(\mathbf{x}, W))},$$

여기에서 $f(\mathrm{x}, W)$는 다음과 같다.

$$f(\mathbf{x}, W) = \sum_{i=1}^{|E|} w_{i1} w_{i2} \ldots w_{i|E_i|} x_{i1} x_{i2} \ldots x_{i|E_i|}.$$

참고로 $x_{i1} x_{i2} \ldots x_{i|E_i|}$는 하이퍼망의 k-하이퍼간선를 표현하는 조합이다. 이 때, 에러 함수는 다음과 같이 정리된다[13].

$$G(W) = -\sum_{n=1}^{N} (y^{(n)} \ln E(\mathbf{x}^{(n)}; W) + (1 - y^{(n)}) \ln(1 - E(\mathbf{x}^{(n)}; W))).$$

이를 편미분한 $\mathbf{g} = \partial G / \partial W$는 다음과 같다[13].

$$g_i = \frac{\partial G}{\partial w_i} = \sum_{n=1}^{N} -(y^{(n)} - y^{*(n)}) \mathbf{x}^{(n)}. \quad (2)$$

여기에서 $\partial G / \partial W$는 모든 데이타에 대한 $\mathbf{g}^{(n)}$의 합이므로, 매 입력 데이타마다 $\mathbf{g}^{(n)}(y^{(n)} - y^{*(n)})$의 반대방

향으로 가중치 W를 조절하는 온라인 형태의 학습을 생각할 수 있으며, 이 때 가중치 W는 시스템의 분류가 틀렸을 경우에만 수정된다. 결국 그림 5에 주어진 학습 알고리즘은 수식 (2)의 온라인 경사도 탐색과정을 단순화한 과정이 된다[14].

4. 실험 결과 및 분석

실험을 위해 암에 관련된 miRNA 모듈 분석을 위한 마이크로어레이 데이타[15]를 사용하여 분류 실험을 수행하였다. 실험에 사용한 miRNA 마이크로어레이 데이타는 89개의 샘플에서 151개의 miRNA의 발현을 측정한 것이다. 89개의 샘플은 68개의 암 조직과 21개의 정상 조직으로 구성되어 있다. 표 1은 데이타 샘플의 구성 조직을 보인 것이다.

학습 데이타는 실험을 위해 각 샘플의 중간값에 기반하여 miRNA의 발현 수준을 0 또는 1로 나누어 바이너리 변환하여 하이퍼망 분류기에 적용하였다. 이러한 변환 과정을 거친 이유는 하이퍼망 구현의 편의성 및 miRNA 모듈의 분석을 쉽게 하는데 있다. 하이퍼망의

표 1 실험에 사용한 miRNA 데이타 샘플 조직 정보

Tissue type	Cancer	Normal
Bladder	1	6
Breast	3	6
Colon	4	7
Kidney	3	4
Lung	2	5
Pancreas	1	8
Prostate	6	6
Uterus	1	10
Melanoma	0	3
Mesothelioma	0	8
Ovary	0	5

구조는 2-*uniform* 하이퍼망을 구성하였으며, 이는 miRNA 모듈의 차수가 증가함에 따른 희소성(sparseness) 문제 및 주어진 데이타에 대한 결과 모듈의 의미 분석의 난이도를 고려함에 따른 것이다. 그림 6은 miRNA 발현 데이타가 하이퍼망 분류기에 적용되는 과정을 도식화한 것이다.

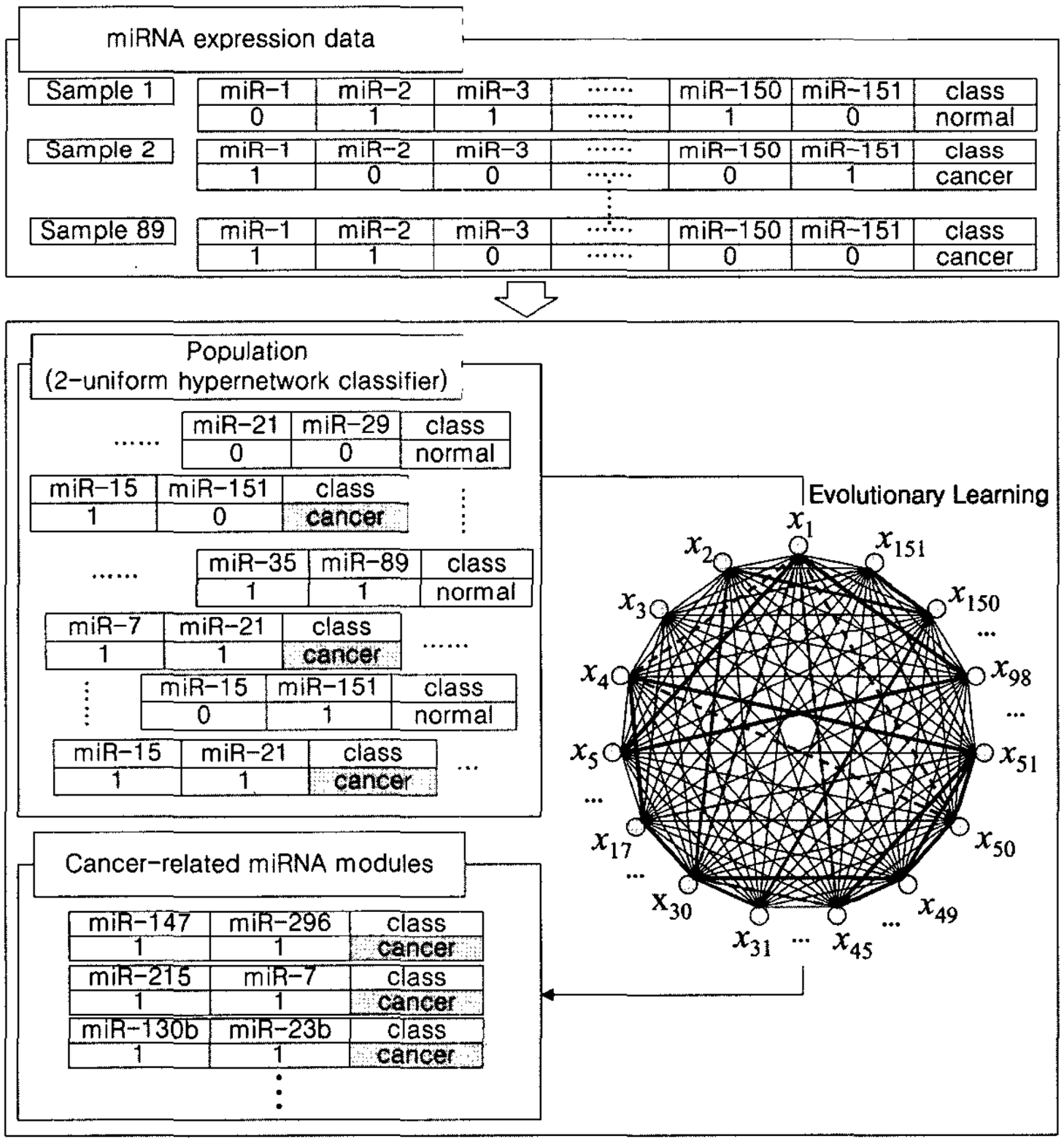


그림 6 miRNA 발현 데이타를 이용한 하이퍼망 분류기 구성

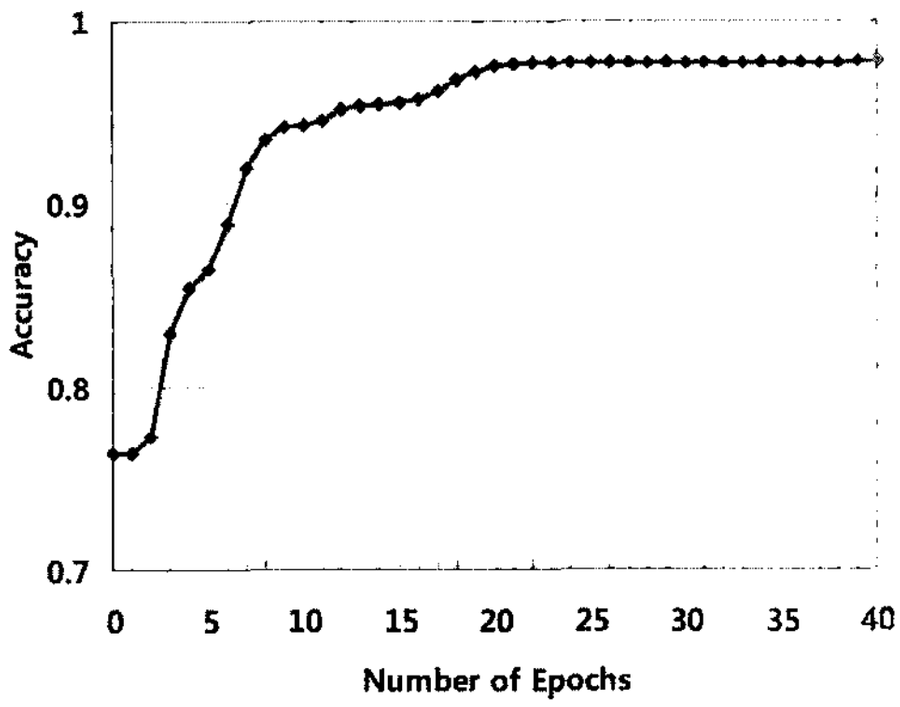


그림 7 miRNA 발현 데이타에 대한 하이퍼망 분류기의 학습에 따른 성능 변화 그래프

본문에 설명한 랜덤 하이퍼망을 이용하여 초기 해집합을 구성하였으며, 개체의 절반은 확률 $p = 0.5$에 의해 학습 데이타로부터 선택하였고, 나머지 부분은 랜덤하게 선택하였다. 이는 해집합의 다양성 및 클래스간 확률차이가 적을 경우 임의성을 반영하기 위함이다. 전체 해집합의 개체 크기는 50,000개로 설정하였으며, 각 개체의 초기 가중치(w_{init})는 동일하게 1,000으로 설정하였다. 하이퍼망 분류기의 에너지 함수로는 sigmoid 함수를 사용하였다. 해집합이 학습되는 정도, 즉 하이퍼간선의 가중치가 변화되는 정도 $\eta = \Delta c_{Ei} / c_{Ei}$는 데이타에 대한 적응성과 안정적인 학습 사이의 균형을 맞추는데 중요한 요소이다. 따라서 실험에서는 η값을 0.01부터 시작하여 매번 epoch의 전체 정확도가 이전 보다 떨어질 경우, $0.75 \cdot \eta$만큼 학습 비율을 감소시켰다. 하이퍼망 학습의 정지 조건은 40회의 epoch을 기준으로 하였다.

그림 7은 마이크로어레이 데이타에 대한 하이퍼망 학습의 각 세대별 분류 정확도를 보인 것이다. 초기 해집합의 구성에서 학습이 진행 됨에 따라 하이퍼망 분류기의 분류 성능은 증가하며, 데이타에 대한 학습이 20회 정도 반복된 후에 수렴이 되고 있음을 알 수 있다. 하이퍼망 학습이 수렴하기 전의 과정은 최적의 miRNA 분류를 위해 후보 하이퍼망을 탐색하는 과정이며, 최적의 하이퍼망을 찾게 되면 그만큼 성능의 증가 곡선이 떨어지는 형태를 취하게 된다.

4.1 miRNA 발현 데이타 분류 성능

표 2는 랜덤 하이퍼망과 기존 기계 학습 기법의 분류 성능을 비교한 것이다. 기존 기계학습 기법으로는 신경망, 지지벡터머신, 결정트리 및 나이브베이즈가 사용되었으며, leave one-out cross validation을 이용하여 성능을 측정하였다. 그 결과, 랜덤 하이퍼망 분류기는 0.9146의 정확도로서 결정트리 및 나이브베이즈보다 높은 성능을 보여주었으며, 신경망과 지지벡터머신과는 대등한 분류 성능을 보여주었다. 신경망과 지지벡터머신은 일반 문제에서 안정적으로 비교적 높은 성능을 보여주는 기계학습 기법으로 알려져 있다. 그러나, 두가지 방법 모두 학습된 결과에 대한 원인 분석이 쉽게 가시화 될 수 없다는 단점을 가진다. 한편 랜덤 하이퍼망 분류기는 분류 성능 외에 유전자 모듈 분석을 위한 도구로서의 장점을 가지고 있기 때문에, 이런 점에서 신경망 및 지지벡터머신과의 차별성이 존재한다.

표 2 분류 기법에 따른 정확도 비교

분류 기법	정확도
신경망	0.9213
랜덤 하이퍼망	0.9146
지지벡터머신	0.9101
결정트리	0.8876
나이브베이즈	0.8314

4.2 miRNA 모듈 분석

하이퍼망을 구성하는 개체, 즉 하이퍼간선은 특정 자질의 조합으로 표현되기 때문에 특정 클래스를 생성하기 위한 하나의 규칙으로서 해석될 수 있다. 따라서 마이크로어레이 데이타를 학습한 하이퍼망 분류기의 개체는 그 가중치에 따라서 주어진 문제를 푸는데 중요한 규칙이 된다고 볼 수 있다. 표 3은 10번의 실험을 반복하여 얻어진 하이퍼망들을 분석하여 높은 가중치를 보이고 중복 발생된 miRNA 모듈을 나열한 것이다. *hsa-miR-147*은 이형접합소실(LOH, loss of heterozygosity)의 발생률이 높은 대표 유전자(marker gene)의 위치에서 2Mb이내의 지역에 있다[16]. 이형접합소실은 정상세포에서 조절 되지 않는 암세포로 변환되는 메커니즘에서 나타나는 유전적 변화의 대표적인 현상 중 하나이다. *has-miR-215*는 자궁암과 유방암에서 DNA copy수가 정상인 경우보다 더 많이 나타나는 지역에 위치하고 있다[17]. DNA copy 수의 변화는 암 조직에서 miRNA 발현에 영향을 미치는 주요 요인이 될 수 있으므로, 특정 암에서 정상에서와 달리 DNA copy 수의 변화가 나타나는 지역에 위치한 miRNA는 암과 관련이 있다고 말할 수 있다. 또한, *has-miR-23b*는 *9q* 염색체 유전체 결손이 많이 일어나는 두 지역 중 한 지역에 위치한다.

표 3 학습된 하이퍼망에서 높은 가중치를 가지는 miRNA 모듈

miRNA modules	miRNA (a)	miRNA (b)
I	***has-miR-147***	*has-miR-296*
II	***has-miR-215***	*has-miR-7*
III	*has-miR-130b*	***has-miR-23b***
IV	*has-miR-105*	*has-miR-133a*
V	*has-miR-147*	*has-miR-206*

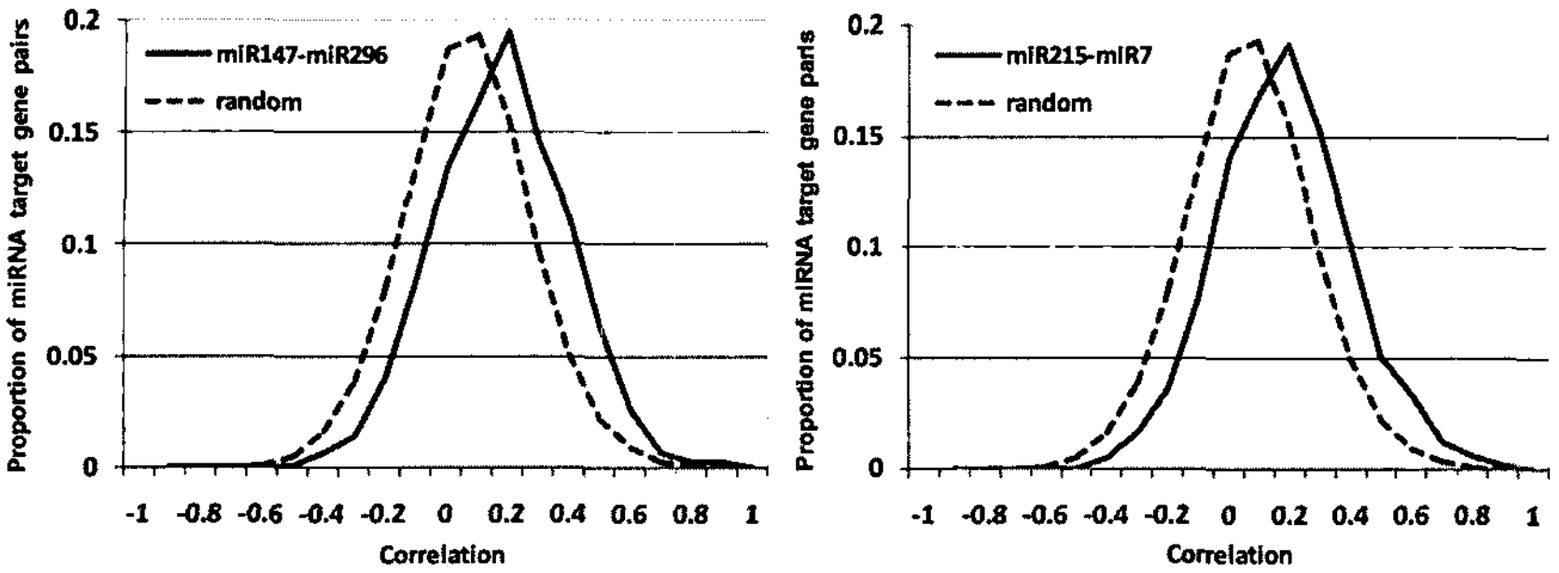


그림 8 miRNA 모듈에 존재하는 miRNA 목적 유전자들의 발현 양상

염색체의 일정 지역에서의 유전체 결손 등 변화가 일어나는 것은 암 발생과 관련이 있다고 알려져 있다[16].

하이퍼망 분류기에 의해 찾아진 miRNA 모듈을 검증하기 위해 표 3의 상위 2개 모듈인 모듈 I 및 II를 구성하는 miRNA들의 목적 유전자들의 발현 양상을 분석해 보았다. 그림 8은 랜덤하게 추출한 유전자와 모듈 I 및 II 각각의 miRNA 목적 유전자 발현 정도를 이용하여 계산한 상관계수 분포 그래프이다. 즉, 모듈 I의 *has-miR-147, has-miR-296*의 목적 유전자와 모듈 II의 *has-miR-215, has-miR-7*의 목적 유전자를 추출하여 각 모든 쌍에 대해 계산한 상관 계수와 랜덤하게 추출한 유전자들간의 상관 계수들과 분포를 비교한 것이다. 목적 유전자들의 상관관계 곡선이 랜덤한 유전자들의 곡선보다 오른쪽으로 치우쳐 있는 것은 상대적으로 더 높은 상관관계가 있다는 것을 의미하며, 곧 하이퍼망에 의해 찾아진 miRNA 모듈의 목적 유전자들이 더 높은 공동발현(co-expression)을 한다는 것을 말한다.

4.3 miRNA 모듈의 생물학적 유의성 분석

추출된 miRNA 모듈의 생물학적 유의성을 검증하기 위해서 모듈을 구성하고 있는 miRNA의 목적 유전자들을 추출하여 Gene Ontology (GO) 분석을 통해 기능적인 연관관계를 알아보았다. GO는 유전자들간 기능적 긴밀성을 검증하는데 표준적으로 쓰이는 분석방법이다. 이 GO 프로젝트의 목적은 세가지 구조로 유전자들을 biological process (BP), cellular component (CC), molecular function (MF) 세가지로 분류하여, 종에 의존적이지 않은 독립적인 관점에서 각 유전자에 대해 생물학적 분석을 제공하는 것이며, 전형적으로 검증은 통계적으로 유의한지 여부에 따라 판단된다. 만약 찾아진 모듈에서 각 miRNA가 생물학적으로 밀접한 관련이 있다면, 그 모듈의 miRNA 목적 유전자들간에도 기능적으로 상관성이 존재할 것이다. miRNA가 특정 생물학적 환경에서 목적 유전자의 기능에 결정적인 영향을 줄 수 있기 때문에, miRNA의 목적 유전자를 이용한 분석은 생물학적으로 유의하다.

표 4는 GOstat[18]을 이용한 모듈 I(*has-mi-147* 및 *has-miR-296*)의 분석 결과이며, p - value<0.01인 유의한 텀(term) 목록을 나타낸 것이다. 표에 나타난 바와 같이, 두 miRNA가 공통으로 타겟하고 있는 13개의 목적 유전자(*BCL3, BCL6, CCND1, CCND2, CDH1, DDX6, ETV6, FGFR1, MYCL1, IRF4, NF2, NRAS, PDGFB*)가 유의한 수준으로 나타났다. 전체적으로, 모듈 I에서의 목적 유전자들은 전사, 단백질 결합, 세포 조절, 생리학적 또는 생물학적 과정에 연관된 특정 기능의 카테고리에 속해 있다.

표 5는 모듈 I을 구성하고 있는 miRNA와 GO 텀이 유의한 수준으로 나타난 공통 목적 유전자 13개에 대한 정보를 보여주고 있다. *has-miR-147*과 *has-miR-296*의 염색체 내 위치 정보와 두 miRNA의 공통 목적 유전자 기능에 대한 설명을 정리한 것이다. 앞서 언급한

표 4 모듈 I의 miRNA 목적 유전자에 대한 GO 텀 분석

GO ID	Term	Ontology	*p-value	Genes
GO:0050794	Regulation of cellular physiological process	BP	2.63E-18	BCL3, BCL6, CCND1, CCND2, CDH1, DDX6, ETV6, FGFR1, MYCL1, IRF1, NF2, NRAS, PDGFB
GO:0050789	Regulation of physiological process	BP	6.43E-18	
GO:0005634	Nucleus	CC	1.52E-17	
GO:0065007	Biological regulation	BP	1.60E-16	
GO:0031323	Regulation of cellular metabolic process	BP	3.73E-16	
GO:0045449	Regulation of transcription	BP	3.91E-16	
GO:0005515	Protein binding	MF	4.36E-16	
GO:0019219	Nucleobase, nucleotide and nucleic acid metabolism	BP	7.22E-16	

표 5 모듈 I의 miRNA의 위치 및 목적 유전자에 대한 설명

miRNA	Chromosome	Start-End Position
has-miR-147	Chr9	122047078-122047149
has-miR-296	Chr20	56826065-56826144

Target	Description
BCL3	B-Cell Leukemia/Lymphoma-3
BCL6	B-Cell Lymphoma-6 (zinc finger protein 51)
CCND1	Cyclin D1
CCND2	G1/S-specific cyclin D2
CDH1	cadherin 1, type 1, E-cadherin (epithelial)
DDX6	DEAD (Asp-Glu-Ala-Asp) box polypeptide 6
ETV6	ets variant gene 6 (TEL oncogene)
FGFR1	fibroblast growth factor receptor 1, fms-related tyrosine kinase 2, Pfeiffer syndrome
IRF4	interferon regulatory factor 4
MYCL1	v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived
NF2	neurofibromin 2 (bilateral acoustic neuroma)
NRAS	neuroblastoma RAS viral oncogene homolog
PDGFB	platelet-derived growth factor beta polypeptide, (simian sarcoma viral (v-sis) oncogene homolog)

바와 같이 *has-miR-147*은 *9q.22* 염색체에서 이형접합 소실 발생률이 높은 지역에 위치하고 있으며, *has-miR-296*은 염색체 20번에 존재한다. 또한, 표 5에 나타난 13개의 유전자는 모두 종양 형성 메커니즘 과정에 적극적으로 포함되어 있다. 예를 들어, *BCL3*는 DNA 손상을 유도할 수 있을 뿐만 아니라, 세포 분열 주기를 조절하여 암 억제 유전자로서의 기능을 하는 *p53*의 활동을 제어하는데 필요한 유전자이다[19]. 또, *BCL6* 역시 *p53* 암 억제 유전자의 발현을 낮추고, B 세포에서 DNA 손상을 유발하는 기능이 있다[20]. 따라서, *BCL3*와 *BCL6* 두 유전자 발현의 변화는 암을 형성하는 메커니즘에 영향을 줄 수 있으며 더 나아가 암의 성장과 존속을 결정하는데 주요한 역할을 할 수 있는 유전자임을 알 수 있다. 모듈 II에 대한 GO 분석도 모듈 I과 같은 결과를 보여주었으며, 따라서 하이퍼망 분류기를 이용해 추출된 miRNA 모듈은 복합작용이 있는 암 관련 모듈이라고 할 수 있다.

5. 결 론

본 논문에서는 마이크로어레이 데이타를 분석하기 위한 방법으로 하이퍼망 분류기를 사용하여 miRNA 모듈을 인식하는 기법을 제안하였다. 학습 데이타를 저장하고 예측하기 위해 확률 모델의 일종인 하이퍼망 분류기를 정의하였으며, 제한된 컴퓨팅 환경 하에서 효과적으로 최적 해를 찾기 위해 랜덤 그래프 모델과 진화 학습 알고리즘을 소개하였다. 하이퍼망은 생체분자 네트웍에 기반해 만들어진 모델로서 지지벡터머신 또는 신경망과 같은 기존 기계학습 기법과는 다르게 학습 결과를 사람이 이해할 수 있는 형태로 직접 분석 가능하다는 장점을 가진다. 암과 관련된 miRNA 발현 프로파일을 이용한 분류 실험에서 하이퍼망 분류기는 결정트리 및 나이브베이즈보다 뛰어난 분류 성능을 보여주었으며, 신경망 및 지지벡터머신과 대등한 성능을 보여주었다. 학습된 하이퍼망 분류기를 통해 발견된 miRNA 모듈이 생물학적으로 의미가 있는지를 보기 위해 GO 분석 및 목적 유전자에 대한 통계분석 방법을 이용하였으며, 그 결과 분석 모듈의 생물학적 유의미성을 발견할 수 있었다.

하이퍼망 분류기 및 진화 기반 학습 방법은 좋은 분류 성능과 함께 사람이 이해 및 분석 가능한 해법을 제공한다는 점에서 자질 선택과 같은 전처리 과정에 사용되거나 마이크로어레이 데이타 외에 분석 능력이 중요한 다른 문제에도 활용될 수 있을 것으로 보인다.

참 고 문 헌

[1] Ramaswamy, S. and Golub, T.R., "DNA Microarrays in Clinical Oncology," *Journal of Clinical Oncology*, Vol.20, pp. 1932-1941, 2002.

[2] Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D., "From Signatures to Models: Understanding Cancer Using Microarrays," *Nature Genetics*, Vol.37, s38-s45, 2005.

[3] Segal, E., Friedman, N., Koller, D., and Regev, A., "A Module Map Showing Conditional Activity of Expression Modules in Cancer," *Nature Genetics*, Vol.36, pp. 1090-1098, 2004.

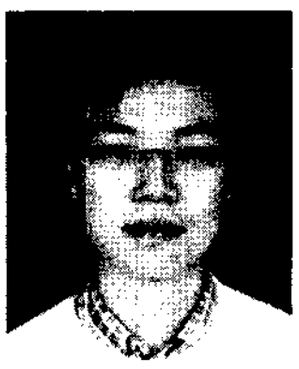
[4] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet C.W., Furey, T.S., Ares, M., Jr., and Haussler, D., "Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," *Proceedings of the National Academy of Sciences*, Vol.97, No.1, pp. 262-267, 2000.

[5] Dettling, M. and Buhlmann, P., "Boosting for Tumor Classification with Gene Expression Data," *Bioinformatics*, Vol.19, pp. 1061-1069, 2003.

[6] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P., "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences*, Vol.102, pp. 15545-15550, 2005.

[7] Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., Kloos, M., Pestell, R.G., West,

M., and Nevins, J.R., "Gene Expression Phenotypic Models That Predict the Activity of Oncogenic Pathways," *Nature Genetics*, Vol.34, pp. 226-230, 2003.
[8] Meltzer, P.S., "Cancer Genomics: Small RNAs with Big Impacts," *Nature*, Vol.435, pp. 745-746, 2005.
[9] Zhang, B.-T., "Random Hypergraph Models of Learning and Memory in Biomolecular Networks: Shorter-Term Adaptability vs. Longer-Term Persistency," *IEEE Symposium on Foundations of Computational Intelligence*, pp. 344-349, 2007.
[10] Kim, S., Kim, S.-J., and Zhang, B.-T., "Evolving Hypernetwork Classifiers for microRNA Expression Profile Analysis," *IEEE Congress on Evolutionary Computation*, pp. 313-319, 2007.
[11] Berge, C., *Graphs and Hypergraphs*, North-Holland Publishing, 1973.
[12] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashitan, N., Chklovskii, D., and Alon, U., "Network Motifs: Simple Building Blocks of Comples Networks," *Science*, Vol.298, pp. 824-827, 2002.
[13] MacKay, D., *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2004.
[14] Kim, S., Heo, M.-O., and Zhang, B.-T., "Text Classifiers Evolved on a Simulated DNA Computer," *IEEE Congress on Evolutionary Computation*, pp. 9196-9202, 2006.
[15] Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R., and Golub, T.R., "MicroRNA Expression Profiles Classify Human Cancers," *Nature*, Vol.435, pp. 834-838, 2005.
[16] Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop, T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M., and Croce, C.M., "Human microRNA Genes are Frequently Located at Fragile Sites and Genomic Regions Involved in Cancers," *Proceedings of the National Academy of Sciences*, Vol.101, No.9, pp. 2999-3004, 2006.
[17] Zhang, L., Huang, J., Yang, N., Greshock, J., Megraw, M.S., Giannakakis, A., Liang, S., Naylor, T.L., Barchetti, A., Ward, M.R., Yao, G., Medina, A., Brien-Jenkins, A.O., Katsaros, D., Hatzigeorgiou, A., Gimotty, P.A., Weber, B.L., and Coukos, G., "MicroRNAs Exhibit High Frequency Genomic Alterations in Human Cancer," *Proceedings of the National Academy of Sciences*, Vol.103, pp. 9136-9141, 2006.
[18] Beissbarth, T., Speed, T.P., "GOstat: Find Statistically Overrepresented Gene Ontologies within a Group of Genes," *Bioinformatics*, Vol.20, No.9, pp. 1464-1465, 2004.
[19] Kashatus, D., Cogswell, P., and Baldwin, A.S., "Expression of the Bcl-3 Proto-Oncogene Suppresses p53 Activation," *Genes and Development*, Vol.20, pp. 225-235, 2006.
[20] Phan, R.T. and Dalla-Favera, R., "The BCL6 Proto-Oncogene Suppresses p53 Expression in Germinal-Centre B Cells," *Nature*, Vol.432, pp. 635-639, 2004.



김 선
1999년 2월 숭실대학교 컴퓨터학부 학사. 2001년 2월 서울대학교 전기·컴퓨터공학부 석사. 2001년~현재 서울대학교 전기·컴퓨터공학부 박사과정. 관심분야는 텍스트마이닝, 생물정보학, 정보검색, 진화연산, 기계학습

김 수 진
2004년 8월 숙명여자대학교 정보과학부 학사. 2005년~현재 서울대학교 협동과정 생물정보학 석박사 통합과정. 관심분야는 생물정보학, 기계학습, 확률 그래프 모델, 진화연산

장 병 탁
1986년 서울대학교 컴퓨터공학 학사. 1988년 서울대학교 컴퓨터공학 석사. 1992년 독일 Bonn대학교 컴퓨터공학 박사. 1992년~1995년 독일국립정보기술연구소(GMD) 연구원. 1995년~1997년 건국대학교 컴퓨터공학과 조교수. 1997년~현재 서울대학교 컴퓨터공학부 교수, 인지과학, 뇌과학, 생물정보학 협동과정 겸임. 관심분야는 Biointelligence, Probabilistic Models of Learning and Evolution, Molecular/DNA Computation