NMF기반 하이퍼그래프 구조 분석을 통한 전립선 암 특이적 유전자 모듈 탐색

김수진 O1 하정우 2 장병탁 1,2
서울대학교 생물정보학 협동과정1
서울대학교 컴퓨터공학부2
{sjkim, jwha, btzhang}@bi.snu.ac.kr

Identifying prostate cancer-specific gene modules via hypergraph analysis based on NMF

Soo-Jin Kim O1 Jung-Woo Ha 2 Byoung-Tak Zhang 1,2
Interdisciplinary Program in Bioinformatics, Seoul National University 1
School of Computer Science & Engineering, Seoul National University 2

요 약

암 조절 메커니즘을 포함하여 생체 내에서 발생하는 대부분의 생물학적 조절 메커니즘은 여러 다양한 요소들이 서로 영향을 주고 받으며 발생하는 복잡한 문제이기 때문에 최근 유전체 수준에서 유전자간 모듈 단위로 분석하는 연구의 중요성이 부각되고 있다. 본 논문에서는 기존 그래프 모델과 달리 다수 인자들간의 고차원적 상호 연관 관계 분석이 가능한 하이퍼그래프(hypergraph) 모델을 이용하여 유전자들간의 연관관계를 학습한 후 학습된 하이퍼그래프의 구조를 NMF (non-negative matrix factorization)를 기반으로 분석하여 협력적으로 특정 암 관련 조절 메커니즘에 참여하는 중요 유전자 모듈을 탐색한다. NMF 기법은 음이 아닌 값으로 구성된 데이터를 두 종류의 양의 행렬의 곱 형식으로 분할하는 데이터 분석 방법으로 파트기반 패턴을 가지고 있는 인자들로 구성된 모듈을 추출하는데 유용하다. 이에 특정 조건에 영향을 주는 인자간 상호작용 관계를 학습하여 구축된 하이퍼그래프 구조를 분석하는데 적합하다. 본 논문에서는 다른 두 유형의 전립선 암 조직 miRNA (microRNA)와 mRNA 발현 데이터를 이용하여 기능적으로 상관관계가 존재하는 공격성 전립선 암의 특이적 miRNA-mRNA 조절 모듈을 탐색하고 생물학적으로 유의함을 실험결과로서 제시한다.

1. 서론

암을 비롯한 다양한 질병 메커니즘을 포함한 복잡한 다원 발생의 생물학적 조절 기전을 이해하는데 표현적 변화에 관계된 유전자들의 고차원적 상호 연관관계를 발굴하기 위해서는 높은 성능의 실험적 데이터와 분석 방법의 조합이 필수적이다. 이에 최근 대용량으로 산출되는 유전체 수준의 데이터를 기반으로 시스템적으로 유전자를 포함한 다양한 생물학적 요소간 상호작용 분석을 위해 모듈 단위로 탐색하는 연구의 중요성이 더욱 부각되고 있다. 따라서 많은 연구자들이 생물학적 시스템에서 세포적 기능을 가지고 있는 유전자 셋(set)을 분석하기 위한 다양한 전산학적 방법을 제안하였다[1]. 특히 최근 miRNA는 여러 종양에 수반하여 주요 유전자의 발현을 제어하고 세포를 기능적으로 조절함으로써 암의 발생과 진행을 유도하는데 중요한 역할을 하는 조절자로 주목 받고 있다. 이에 miRNA의 기능과 활동 기전을 이해하기 위해서는 모듈 단위의 miRNA와 mRNA간 연관관계 셋을 탐색하는 것은 필수적이라 할 수 있다. 그러나 여러 다양한 생물학적 요소와 더불어 miRNA와 mRNA 상호작용의 다양성과 복잡성 때문에 유전체 수준에서 기능적인 miRNA-mRNA 조절 모듈을 추론하는 것은 여전히 어려운 문제로 남아 있다.

본 논문에서는 두 유형의 전립선 암 조직의 발현 데이터로부터 인자간 고차적 상호작용 탐색이 가능한 하이퍼그래프 모델을 구축하고 NMF을 기반으로 학습된 하이퍼그래프 구조를 분석하여 암 특이적 miRNA-mRNA 조절 모듈을 추출하는 방법을 제안한다(그림 1). 하이퍼그래프 분류 모델은 기존 일반 그래프 모델과 달리 다수 인자들 간의 고차 관계가 표현이 가능하기 때문에 복잡한 상호작용 현상이 다수 일어나는 생물학

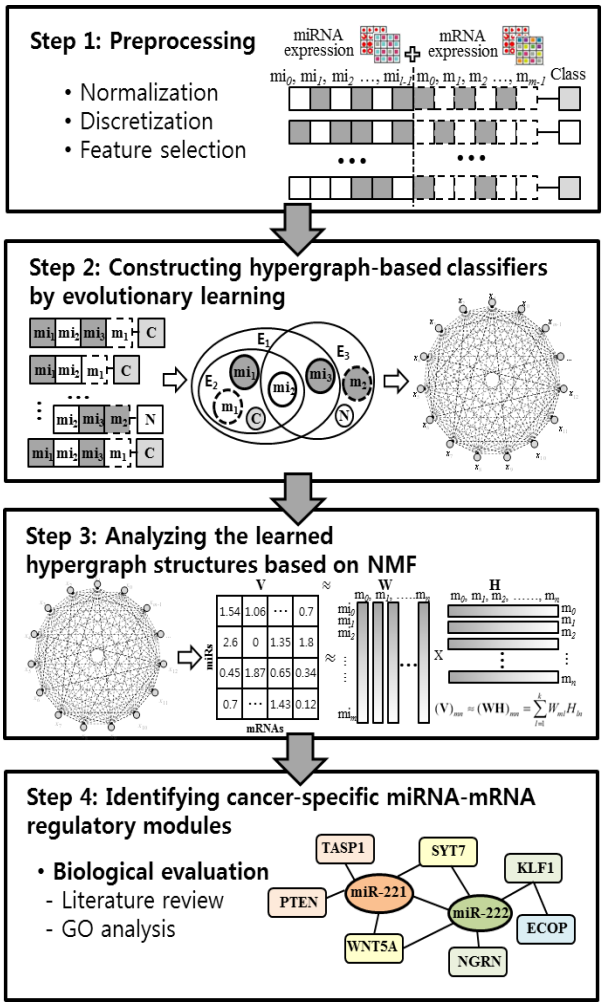

그림 1 miRNA 와 mRNA 발현 데이터로부터 암 특이적 miRNA-mRNA 조절 모듈 탐색을 위한 전체 개요도

문제를 모델링 하는데 적합하다. 이에 부분적 패턴 분석에 유용한 NMF (non-negative matrix factorization)를 기반으로 높은 정확도로 암을 분류하도록 학습된 하이퍼그래프 모델의 구조를 분석하여 협력적으로 특정 암 관련 조절 메커니즘에 참여하는 중요 miRNA-mRNA 모듈을 추출한다. 전립선 암 발현 데이터에 적용한 실험 결과에서 제안한 방법은 다른 기계학습 방법에 비해 높은 분류 성능을 보여준다. 더욱이 학습된 모델 구조 분석을 통해 암 조직에서 특이적인 패턴을 가지고 있는 miRNA-mRNA 조절 모듈을 추출하여 문헌 등을 통해 생물학적으로 유의함을 검증하여 제시한다.

2. 진화연산을 이용한 하이퍼그래프 모델 학습

하이퍼그래프 모델[2]은 복잡한 생물학계에서 다수 인자 간의 상호작용을 분석하는데 적합한 모델이다. 이는 2개 인자간 연결 관계 표현만 가능했던 기존 일반 그래프와 달리 2개 이상 인자간 관계 표현이 가능하여 수 많은 인자들간 복잡한 고차적 상호작용을 분석이 가능하기 때문이다. 하이퍼그래프 모델은 2개 이상의 정점을 동시에 연결 가능한 하이퍼에지(hyperedge)로 구성되어 있으며, 정점은 데이터를 구성하는 인자(attribute)를 의미하고 하이퍼에지는 인자들간의 고차적 조합을 의미한다. 즉, 하이퍼그래프 H는 $H = (V, E)$로 V 와 E는 정점 v의 셋, 하이퍼에지 e의 셋을 각각 의미한다. 또한 하이퍼에지는 V의 부분집합으로 표현되며 각 하이퍼에지는 가중치(weight)를 가지고 있다. 이에 제안한 하이퍼그래프 모델은 인자들간 고차적 관계를 반영한 weak learner 역할을 하는 수많은 하이퍼에지의 앙상블 머신으로서 인식된다. 최적의 하이퍼그래프 모델 H^*는 데이터 셋 D, $D = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$ 가 주어졌을 때 다음과 같은 식으로 정의 될 수 있다.

$$H^* = \arg\max_{H} P(Y \mid \mathbf{x}, H) \quad (1)$$

$$P(Y \mid \mathbf{x}, H) \cong \frac{1}{N}\sum_{n=1}^{N} \delta(y^{(n)}, \hat{y}^{(n)}; H) \quad (2)$$

위의 식에서 $\delta(y^{(n)}, \hat{y}^{(n)}; H)$ 는 데이터 샘플의 클래스 $y^{(n)}$ 과 모델에 의해 분류된 클래스 $\hat{y}^{(n)}$ 가 동일하면 1, 그렇지 않으면 0을 나타내는 표시함수이다. 위의 식에 따라 분류 에러가 최소가 되도록 하는 하이퍼에지 생성 확률을 최대로 하여 최적의 하이퍼그래프 모델을 구축한다. 그러나 하이퍼그래프 모델은 데이터를 구성하는 인자들간 조합 공간을 확률적으로 표현하는 모델이므로 인자의 수가 증가함에 따라 문제 공간이 기하급수적으로 증가한다. 이와 같은 거대한 문제 공간을 점진적 방법을 통해 탐색하는 것은 현실적으로 불가능하기 때문에 학습을 현실적으로 가능하게 하기 위해 샘플링 기반의 진화 연산 방법을 적용한다. 즉, 하이퍼에지들의 선택, 증폭, 제거 등 진화연산 기법을 통해 하이퍼그래프의 구조와 파라미터를 학습하여 최적의 하이퍼그래프를 구축한다. 또한 학습의 효율성을 강화하기 위해 하이퍼에지 생성에 있어서 각 인자와 데이터의 클래스간 상호정보(mutual information) 값을 반영한 샘플링 기법을 이용한다[3].

3. NMF 기반 학습된 하이퍼그래프 구조 분석

NMF기법[4]은 음이 아닌 값으로 구성된 데이터를 두 개의 저차원 양의 행렬의 곱으로 분할하는 방법으로 데이터에 내재된 부분적 특징 패턴을 추출하는데 유용하다. NMF는 m개의 다변량, n차원의 데이터 벡터의 집합으로 구성된 $m \times n$행렬의 데이터 셋 V가 주어졌을 때 이를 기저행렬(basis matrix)과 상관행렬(coefficient

matrix)로 분할하여 V로 근사화 한다.

$$V_{m\times n} \approx W_{m\times k} H_{k\times n} \quad W, H \geq 0 \tag{3}$$

식(3)에서 $m\times k$ 행렬 W와 $k\times n$ 행렬 H는 각각 기저행렬과 상관행렬이고 k는 n이나 m보다 작게 선택하여 W나 H의 크기가 V의 크기보다 작게 한다. 이 두 요소를 기반으로 업데이트 룰을 적용하여 목적함수 값이 수렴 허용 오차보다 작아지거나 지정한 반복 횟수를 초과할 때까지 과정을 반복하여 최적화한다. 식(4)는 Frobenius norm을 사용한 목적함수이다.

$$\Theta_E = \min \left\| V - WH \right\|_F^2 \quad W, H \geq 0 \tag{4}$$

본 연구에서는 miRNA와 mRNA 관계를 학습한 하이퍼그래프로부터 miRNA와 mRNA간 가중치를 $m\times n$ 행렬 V로 표현하여 squared error가 최소화되는 W와 H로 분해하여 biclustering에 의해 부분적 특징패턴을 가진 miRNA-mRNA 모듈을 추출한다. 행렬 V의 각 원소 값은 다음과 같이 정의된다.

$$v_{ij} = \sum_{e\subset E} \sum_{x_i^m \in X^m} \sum_{x_j^{mi} \in X^{mi}} w(e)\delta(x_i^m, x_j^{mi}, e) \tag{5}$$

또한 NMF가 biclustering에 적용될 때 행렬 V를 구성하는 데이터 포인트(v_j)는 기저행렬의 열의 비음수 선형 조합으로 표현되며, W를 구성하는 각 기저벡터는 클러스터의 중심점(centroid)이 되고, 행렬 H내의 각 열에서 가장 큰 값을 갖는 원소의 행 값이 클러스터 인덱스가 된다.

4. 실험결과

4-1. 데이터 전처리

본 연구에서는 두 유형의 전립선 암 조직의 miRNA와 mRNA의 발현 프로파일 데이터를 사용하여 공격성 전립선 암 특이적 miRNA-mRNA 조절 모듈을 탐색하였다[5,6]. 실험 데이터는 공격성 전립선 암 조직과 비공격성 전립선 암 조직 각각 45개 총 90개의 샘플로부터 470개 miRNA와 24,519개 mRNA의 발현 정도를 측정한 것이다. 실험을 위해 각 샘플 단위의 발현 평균값을 기준으로 발현 수준을 정규화시키고, 정규화 된 발현 값을 각 유전자 별로 샘플 평균값을 기준으로 miRNA와 mRNA의 발현 수준을 0 또는 1로 이산화하였다. 제안하는 하이퍼그래프 모델은 어떤 형태의 인자 값이라도 프로세스가 가능하지만, 본 실험에 있어서는 *in silico* 상에서 알고리즘의 효율적인 구현과 탐색한 miRNA-mRNA 관계 분석을 용이하게 하기 위해 이산화 데이터를 사용하였다. 또한 생물학적으로 더 유의성 있는 결과를 얻기 위해 전체 mRNA 목록에서 [7]에 따라 mRNA를 추출하여 실험을 수행하였다.

표 1 다른 기계학습 방법들과 성능 비교 결과

분류기법	정확도(표준편차)
하이퍼그래프 모델	0.921 (±0.014)
SMO (2nd poly)	0.911 (±0.011)
J48 결정트리	0.819 (±0.028)
나이브 베이즈	0.815 (±0.014)

4-2. 분류성능

표 1는 제안하는 알고리즘과 다른 기계학습 모델들과의 분류 성능을 비교한 결과이다. 다른 기계학습 방법으로는 WEKA에서 제공되는 SVMs (support vector machines), J48 결정트리, 나이브 베이즈(naïve Bayes)를 사용하였으며 10-fold-cross validation을 이용하여 10번 반복 수행하여 분류 정확도의 평균을 구하였다. 그 결과 하이퍼그래프 모델은 정확도 92.1%로서 다른 모델들에 비해 우수한 분류 성능을 보였다.

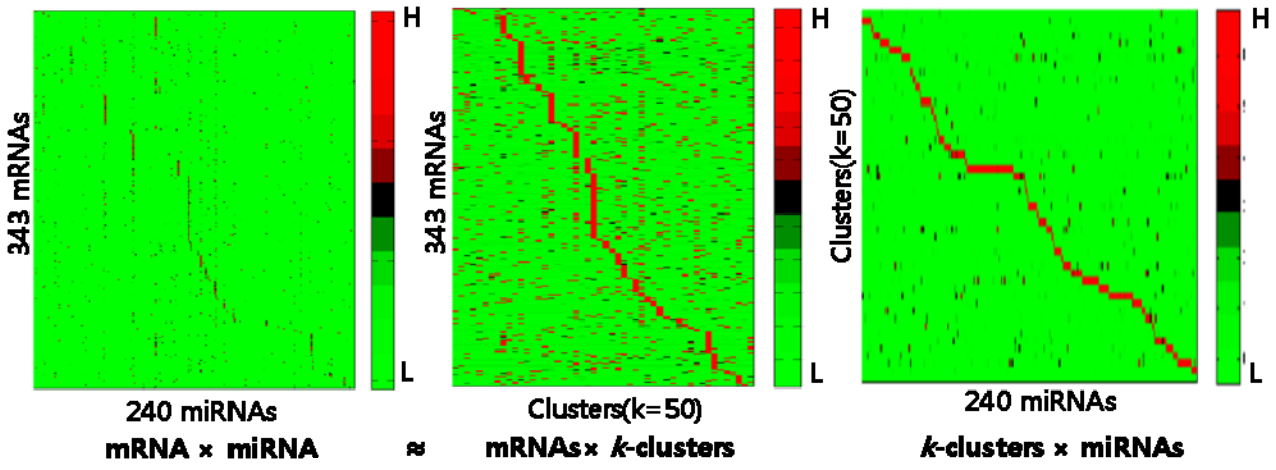

그림 2 학습된 하이퍼그래프 구조를 NMF 기반으로 biclustering을 이용하여 분석한 결과

표 2 추출된 공격성 전립선 암 특이적 MIRNA-MRNA 조절 모듈의 기능적 분석 결과

No.	GO biological process terms	*p*-value	miRNAs	mRNAs
10	lymphocyte & leukocyte proliferation	3.54E-4	hsa-miR-101	CASK, SGK3, PRKCD
	adaptive immune response	4.84E-3		
21	response to stress	3.21E-5	hsa-miR-331	C1R, CASP10, CASP9, ESPL1, CDKN1A, MMP7, PSMC3, CSNK1E, ELF3, EXO1, RAD23A
	proteolysis	8.84E-4		
	cellular response to steroid hormone stimulus	2.25E-2		
50	negative regulation of organ growth	7.34E-6	hsa-miR-222	ARF3, BCL6, BUB3, CAMK2G, CAMK4, CDK5R1, EIF1, ELF5, ELK1, LTK, MAP2K7, WNK1
	estrogen metabolic process	4.30E-3		
	male genitalia development	7.69E-3		

4-3. 공격성 전립선 암 특이적 miRNA-mRNA 모듈 추출 및 생물학적 유의성 검증

그림 2는 공격성 전립선 암을 분류하는 데 있어서 miRNA와 mRNA 연관관계가 학습된 최적의 하이퍼그래프 구조를 NMF를 기반으로 biclustering한 결과이다. 이는 470개의 miRNA와 2238개의 mRNA간 관계를 학습하여 얻어진 하이퍼그래프 구조로부터 k 값을 50으로 설정하여 NMF에 의해 W와 H행렬을 구한다. 이 중 신뢰도가 높은 240개의 miRNA와 343개의 mRNA 간의 상관관계를 biclustering을 통해 50개의 클러스터로 부분적으로 특징적 패턴이 나타나는 miRNA와 mRNA 집합을 추출하였다. 표 2는 추출된 대표적 공격성 전립선 암에 특이적인 miRNA-mRNA 조절 모듈의 GO (Gene Ontology) 생물학적 프로세스 분석 결과이다[8]. 추출된 모듈의 hsa-miR-222[9], hsa-miR-331[6], hsa-miR-101[10] 모두 전립선 암 메커니즘에 직접적으로 연관이 있음이 여러 연구를 통해 보고되었다. 또한 같은 모듈 내에 있는 mRNA들 역시 스테로이드 호르몬, 남성 생식기 발달 등 전립선 암에 관련 있는 term들이 유의한 수준으로 나타남으로써 비슷한 생물학적 기능을 공유하고 있음을 보여준다.

5. 결론

본 논문에서는 인자간 고차적 상호작용 탐색이 가능한 하이퍼그래프 모델을 구축하여 NMF을 기반으로 학습된 하이퍼그래프 구조를 분석, 암 특이적 유전자 조합을 추출하는 방법을 제안한다. 이는 높은 분류 성능과 더불어 인자간 고차적 연관관계 이해를 위한 해석 가능한 구조를 제공함으로써 복잡한 생물학 문제를 해결하는데 유용한 방법이다. 실험적으로 제안한 모델을 통해 두 유형의 전립선 암 조직의 발현 데이터로부터 miRNA와 mRNA 고차적 관계를 학습하여 NMF 기반으로 biclustering을 이용하여 공격성 전립선 암에 특이적인 miRNA-mRNA 조절 모듈을 추출하였다. 추출된 모듈을 구성하는 miRNA와 mRNA는 생물학적으로 전립선 암 메커니즘에 관련되어 있음을 확인하였다.

감사의 글

이 논문은 교육과학기술부 국가연구재단의 (2011-0016483, Videome 및 2011-0018299, BrainNet), 지식경제부 산업원천기술개발사업(10035348, mLife), 한국학술진흥재단(314-2008-1-D00377, Xtran)에 의해 지원되었음.

참고문헌

[1] A. Subramanian, *et al*., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *PNAS*, 102: 15545-50, 2005.

[2] B.-T. Zhang, Hypernetworks: a molecular evolutionary architecture for cognitive learning and memory, *IEEE Computational Intelligence Magazine*, 3(3): 49-63, 2008

[3] S.-J. Kim, *et al.*, Evolutionary hypernetworks based on mutual information for cancer gene expression profile analysis, *APBC 2010*, pp.286, 2011.

[4] D. Lee and S. Seung, Learning the parts of objects by non-negative matrix factorization, *Science*, 401: 788-791, 1999.

[5] L. Wang, *et al*., Genome-wide transcriptional profiling reveals microRNA-correlated genes and biological processes in human lymphoblastoid cell lines, *PLoS ONE*, 4(6): e5878, 2009.

[6] L. Wang, *et al*., Gene networks and microRNAs implicated in aggressive prostate cancer, *Cancer Research*, 69(24): 9490-9497, 2009.

[7] P.A. Futreal, *et al*., A census of human cancer genes, *Nature Reviews Cancer*, 4: 177–183, 2004.

[8] Q. Zheng and X. Wang, GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis, *NAR*, 36: W358-W363, 2008.

[9] A. Tong, *et al*., MicroRNA profile analysis of human prostate cancers, *Cancer Gene Ther*., 16(3): 206-16, 2008.

[10] H. Huang, *et al*., Commentary on Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer, *Urol Oncol*., 27(2): 230, 2009.