The Application of Machine Learning Algorithm In The Analysis of Tissue Microarray; for the Prediction of Clinical Status

Sung Bum Cho*, Woo Ho Kim†, Ju Han Kim*

** Seoul National University Biomedical Informatics*, †Department of Pathology, College of Medicine, Seoul National University, Seoul 110-799, Korea, correspondence to juhan@snu.ac.kr*

ABSTRCT: Tissue microarry is one of the high throughput technologies in the post-genomic era. Using tissue microarray, the researchers are able to investigate large amount of gene expressions at the level of DNA, RNA, and protein The important aspect of tissue microarry is its ability to assess a lot of biomarkers which have been used in clinical practice. To manipulate the categorical data of tissue microarray, we applied Bayesian network classifier algorithm. We identified that Bayesian network classifier algorithm could analyze tissue microarray data and integrating prior knowledge about gastric cancer could achieve better performance result. The results showed that relevant integration of prior knowledge promote the prediction accuracy of survival status of the immunohistochemical tissue microarray data of 18 tumor suppressor genes. In conclusion, the application of Bayesian network classifier seemed appropriate for the analysis of the tissue microarray data with clinical information.

Keyword: Tissue microarray, Gastric cancer, Bayesian network classifier, Prior knowledge

1. 배경

조직 마이크로어레이는 마이크로어레이 기술의 영향을 받아서 발전된 기법으로, 슬라이드위에 미세하게 제작된 조직 절편을 하나의 슬라이드 위에 배열하는 실험 방법을 말한다. 조직 마이크로어레이에서는 면역 형광 염색, in situ hybridization, in situ RT PCR등의 조직을 이용하는 대부분의 기법이 가능하여, 각 조직에서 DNA, RNA 그리고 protein 수준에서 유전자 발현을 관찰할 수 있다.[1] 이 방법의 장점은 한 번에 여러 가지 유전자의 발현을 검사할 수 있다는 것이다. 특히, 면역 형광 염색의 경우 이제까지 임상 진료에 있어서 분자 생물학적인 정보를 제공하는데 중요한 역할을 했었고, 진료 방침의 결정에도 기여했으므로 많은 유전자에 대한 면역형광 염색법을 이용한 연구가 있어왔다. 이러한 사실을 감안한다면, 조직 마이크로어레이가 임상 진료에 있어서 중요한 역할을 담당할 가능성은 크다고 할 수 있다. 상대적으로 저렴한 비용으로 검사가 가능하고 검사하려는 조직이 안정적으로 장기간 보관할 수 있다는 점도 이러한 가정을 뒷받침해 주고 있다.

이러한 장점과 가능성에도 불구하고 마이크로어레이의 경우와는 달리, 면역 형광 염색을 이용한 조직 마이크로어레이의 분석은, 저자들이 조사한 바로는 이제까지 발표된 논문은 조직 마이크로 어레이상에서 나타난 각 유전자의 발현과 임상 양상과의 개별적인 연관성을 조사하는 것이 대부분이었다. [2] 따라서 저자들은 조직 마이크로어레이상에서 나타난 유전자 발현양상의 분석을 위해서 패턴 인식 알고리즘의 하나인 베이지안 망 분류기를 사용하였다.

베이지안망 분류기(Bayesian network classifier)는 나이브베이즈 망(Naive Bayes net)의 분류 효율을 높이기 위해 고안된 방법이다. 나이브베이즈 망이 노드간에 독립성을 가정하는 반면에 베이지안 망 분류기는 노드간에 의존성이나 계층을 두거나, 일반 노드처럼 취급하여 학습을 수행한다.[3] 그동안 많

은 분야에서 베이지안망 분류기의 우수한 수행능력이 입증되었고, 이산형 자료의 처리와 학습 시에 사전 지식의 도입이 용이하다는 장점을 가지고 있다.

본 연구에서는 사전 지식의 도입이라는 베이지안망 분류기의 장점을 이용하여 조직 마이크로어레이 상에서 나타난 다양한 암 억제 유전자의 발현 양상이 환자의 예후를 어느 정도의 정확성을 가지고 예측할 수 있는지 조사하였다.

2. 방법

저자들은 베이지안망 분류기를 이용하여 위암 환자 197명의 면역 형광 염색된 조직 마이크로어레이 자료를 분석하였다. 분석에 사용된 자료는 기존에 발암 억제 유전자로 알려진 총 18개의 유전자에 대한 면역 형광 염색 결과를 포함하였다(표 1). 그리고, 위장관계에서 발생하는 종양의 mucin과 cytokeratin의 발현 각 유전자의 발현 여부는 기존의 문헌에서 정한 기준에 따라 병리학자들의 평가에 의해 판단되었다. 그리고 각 환자의 수술 후 생존 여부와 TNM 병기도 입력 변수로 사용되었다.

분석 시에는 먼저, 환자의 생존 여부를 분류 기준으로 놓고 기존의 발암 억제 유전자에 대한 지식을 이용하지 않고 조직 마이크로어레이 자료만을 가지고 유전자 발현 양상이 생존 여부를 어느 정도 예측할 수 있는지를 조사하였다. 그리고 기존의 연구 결과를 사전 지식으로 분석에 도입하여 처음 분석 결과와 비교하였다. 마지막으로, 위암의 병기를 임의적으로 초기(TNM 병기 1&2, n=135)인 환자만을 대상으로 했을 때 유전자 발현 양상을 통한 생존여부의 분석 결과가 진행된 병기(병기 3&4, n=62)의 환자를 대상으로 했을 때와 어느 정도 차이를 보이는지를 조사하였다. 모든 분석에서 각 군의 학습 자료(training data)는 각 군의 전체 자료수의 70%로 하고 테스트 자료(test data)는 30%로 하였다.

사전 지식을 도입했을 때에는 J. Chen등이 제안한 알고리즘과 베이지안망 분류기 프로그램(PowerPredictor)을 이용하였다. [4]

Gene Symbol	Gene Name
p53	tumor protein 53
BCL2	B-cell leukemia/lymphoma 2
p16	CDK4 inhibitor
Cycl B2	Cyclin B2
MUC1	mucin 1
MUC2	mucin 2
MUC5ac	mucin 5ac
MUC6	mucin 6
Rb	retinoblastoma protein
CEA	carcinoembryonic antigen
SMAD4	Mothers against DPP homolog4
FHIT	Fragile histidine triad gene
CD44	CD44 antigen
E-cad	E-cadherin
VHL	Von-Hippel Lindau tumor supressor
KAI1	Kangai 1(CD 82 antigen)
MGMT	O-6-methylguanine methyl tranferase
PTEN	Phosphatase and tensin homolog

Table 1. Gene list

3. 결과

전체 결과는 표 2에 정리하였다.

	Without prior (train/test %)	With prior (train/test %)
Prediction Accuracy	93.8/78.4	95/86.3
Error Range	±5.30/±11.29	±4.78/±9.44
Confidence interval	95%	95%

Table 2. Results of Bayesian network classifier analysis with or without prior knowledge

1. 사전 지식을 이용하지 않은 분석

사전 지식을 이용하지 않은 분석 결과 환자

의 생존여부의 예측 정확도는 학습 자료에서는 93.8%를 그리고 학습된 모델로 테스트 자료에서의 예측 정확도는 78.4%를 기록하였다.

2. 사전 지식을 이용한 분석

이제까지 연구 발표된 사전 지식을 표 3과 같이 정리하여 분석에 사용하였고,[5][6] 예측 정확도는 학습 자료와 테스트 자료에서 각각 95%와 86.3%를 기록하였다. 그림 2에 구축된 베이즈망이 있다.

Leaf	Forbidden links
all genes except p53 TNM stage	survival status -x- CEA survival status -x- muc5ac survival status -x- muc6 survival status -x- CD44 sruvival status -x- muc2 survival status -x- cyclin B2

Table 3. Prior knowledge

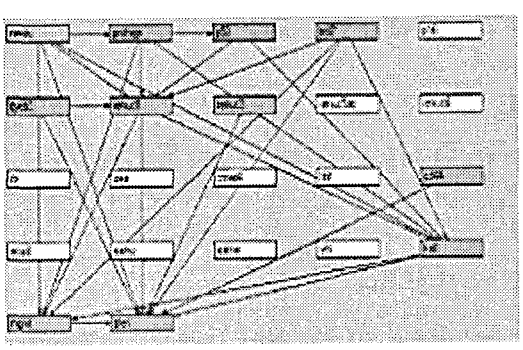

Figure 2. Bayesian network of tissue microarray data with prior knowledge

3. 초기와 진행된 병기에서의 분석

전체 자료 중에서 TNM 병기에서 1기와 2기 환자 132명과 3,4기 환자 65명을 대상으로 분석을 실시하였다. 초기(1,2기)에서 테스트 자료에서 사전지식을 이용하지 않을 때는 78.1%의 정확도를 보였고, 앞서와 같은 사전지식을사용할 때에는 86.2%의 정확도를 보였다(표4, 그림3). 진행된 병기의 경우, 각각 52.2%와 65.2%를 기록하여 초기일 때 보다 낮은 예측 정확도를 기록하였다.

	Without prior (train/test %)		With prior (train/test %)	
Stage	Early	Late	Early	Late
Prediction Accuracy	93/78.1	73.6/52.2	95/86.2	76.3/65.2
Error Range	±5.30/±11.29	±11.98/±20.42	±4.78/±9.44	±13.52/±19.47
Confidence interval	95%	95%	95%	95%

Table 4. Results of Bayesian network classifier analysis with or without prior knowledge, in early-staged Vs. late-staged patients

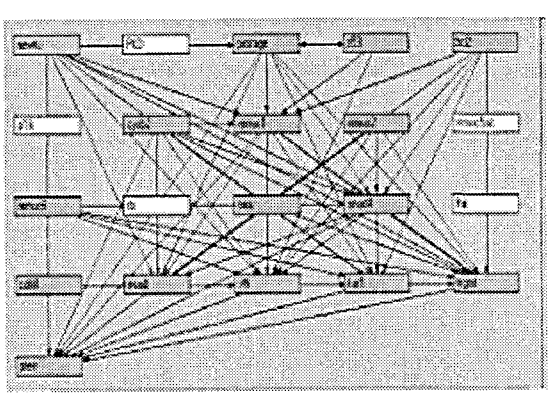

Figure 3. Bayesian network of tissue

microarray data with prior knowledge in early staged patients

4. 결론

본 연구에서는 베이지안망 분류기를 이용하여 조직 마이크로어레이의 암 억제 유전자의 발현 양상으로 위암 환자의 생존 여부를 어느 정도 예측 할 수 있는지를 조사하였다. 전반적으로 사전 지식을 사용할 때는 예측 정확도가 증가하는 것을 관찰 할 수 있었다. 사전 지식의 적용은 문헌에 의존하였고 위암과 연관된 경우만을 적용시켰다. 한 가지 유전자에 대해서 위암과 연관된 경우만을 찾는 경우, 관련된 지식이 있다고 하더라도 임상적인 자료가 아니라면 적용하는 것이 적절하지 않을 수 있다는 가정 하에 참고할 수 없는 지식으로 분류하였다. 사전 지식을 적용한 후에 예측 정확도의 향상을 관찰 할 수 있었으므로, 저자들은 사전 지식이 적용이 위암 종양억제 유전자의 조직 마이크로어레이 자료의 분석에서 유용할 수 있다고 결론지었다.

생존 여부 예측의 정확도가 전체 환자 군을 대상으로 할 때보다 초기 병기 환자 군에서 더 증가하는 것을 관찰 할 수 있었는데, 이 현상이 자료 자체의 특성에 기인하는 것인지 아니면 어떠한 생물학적 의미를 지니는 것인지는 확실하지는 않다. 만약, 진행된 병기에서 암 세포들의 종양 유전자 발현 양상이 질병 진행 과정에서 처음의 양상과는 다른 형태로 나타날 가능성이 초기 위암 세포들보다 크다면 위와 같은 결과가 가능할 것이다.

한 가지 암에서 서로 다른 임상 병기는 서로 다른 유전자 발현 양상을 보일 수 있으므로, 저자들은 병기 정보를 처음부터 분석에 도입하는 것이 좋은 예측율을 보일 수 있다고 가정 하였다. 이 연구에서는 처음부터 병기를 분석에 포함하였다. 이 논문에서 결과를 제시하지는 않았지만, 병기 정보가 없이 분석을 할 때는 예측 정확도가 60%에서 70%사이를 기록하였으므로, 병기 정보의 포함이 예측에 도움을 주는 것이 확인 되었다.

본 연구에서는 유전자 발현 양상과 환자의 생존 유무와의 관계를 분석하였다. 그러나 환자의 생존 기간은 관찰 시점에 따라 다를 수 있기 때문에 결과의 해석에 어려움이 있을 수 있다. 예를 들어 생존한 환자의 관측 기간이 모두 다르다고 한다면, 환자의 생존을 예측할 수 있는 모델의 의미가 모호해질 수 있다. 따라서 저자들은 분석시에 생존한 환자들을 선정할 때 모두 5년 이상 생존 기간을 가진 환자만을 대상으로 하였다.

결론적으로, 저자들은 본 연구를 통해서 임상 자료와 연계된 조직 마이크로어레이 실험의 분석에서 사전 지식을 이용하는 베이지안망 분류기의 적용이 유용할 수 있음을 확인하였다. 만약 이러한 방법이 좀 더 발전한다면, 향후 임상 진료에 있어서 치료 방침의 결정에 도움을 주는 역할을 할 수 있을 것이다.

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Science and Technology, Republic of Korea (2005-00162).

참고문헌

1. 김우호. Clinico-pathologic study를 위한 tissue microarray 방법. 2001. 대한위암학회 추계 학술대회.

2. Donati V, Faviana P, Dell'omodarme M,

Prati MC, Camacci T, De Ieso K, Giannini, R, Lucchi M, Mussi A, Pingitore R, Basolo F, Fontanini G.
Applications of tissue microarray technology in immunohistochemistry: a study on c-kit expression in small cell lung cancer. Hum Pathol. 2004 ;35(11):1347-52.

3. Nir Friedman, Dan Geiger, Moises Goldszmidt. Bayesian network classifiers. Machine learning 1997:29: 131-163.
4. J. Cheng. Belief network Power Predictor.
www.cs.ualberta.ca/~jcheng/bnpp.htm

5. Lee HS, Lee HK, Kim HS, Yang HK, Kim WH.Tumour suppressor gene expression correlates with gastric cancer prognosis. J Pathol. 2003;200(1):
39-46.

6. Lee HS, Lee HK, Kim HS, Yang HK, Kim YI, Kim WH. MUC1, MUC2, MUC5AC, and MUC6 expressions in gastric carcinomas: their roles as prognostic indicators. Cancer. 2001 ;92(6):1427-34.