Latent variable model에 의한 바이러스 유형 분석

김수진[O1,2] 정제균[1,2] 태강수[3], 장병탁[1,2,4]
서울대학교 생물정보학 협동과정[1]
서울대학교 바이오정보기술 연구센터[2]
전주대학교 정보기술공학부[3]
서울대학교 컴퓨터공학부[4]
{sjkim[O], jgjoung, kstae, btzhang}@bi.snu.ac.kr

Analysis of Virus Types by a Latent Variable Model

Soo-Jin Kim[01,2] Je-Gun Joung[1,2] Kang Soo Tae[3] Byoung-Tak Zhang[1,2,4]
Graduate Program in Bioinformatics, Seoul National University[1]
Center for Bioinformation Technology, Seoul National University[2]
School of Information Technology and Engineering, Jeonju University [3]
School of Computer Science and Engineering, Seoul National University[4]

요 약

인유두종 바이러스(Human Papillomavirus: HPV)는 사마귀로부터 생식기 및 배설기의 침윤성 암에 이르기까지 여러 질병과 연관되어 있음이 알려져 있다. 현재 200종 이상이 알려져 있고, 이 중 85개는 전체 유전자가 밝혀져 있다. HPV 감염 시 만들어지는 단백질 중 E6, E7 단백질은 암 억제 유전자(p53, pRb)에 결합하여 세포의 암 억제 기능을 저하시키고 이로 인해 암을 발생시킨다. 본 논문은 암 발생과 밀접한 관련이 있는 HPV의 E6 단백질 서열과 HPV 유형(HPV Type)을 가지고, PLSA (Probabilistic Latent Semantic Analysis) 방법을 이용하여 HPV를 클러스터링(clustering) 해 보았다. 실험 결과, 특정 클러스터는 질병과 밀접하게 연관되어 있으며, 이와 관련된 주요 서열 분석이 가능함을 보여주고 있다.

1. 서 론

현대 의학의 발전에도 불구하고 암은 그 발생 원인이 아직도 확실히 밝혀지지 않고 있다. 그러나 여성에게 발생하는 가장 흔한 악성 종양 중 하나인 자궁경부암에서는 인유두종 바이러스(Human Papillomavirus: HPV)가 암 발생 과정에서 중요한 역할을 한다는 것이 증명되었다[1].

HPV는 약 8kb의 환상의 이중나선 DNA 바이러스로 papovavirus 과에 속해있다. 흔히 손과 발에 사마귀를 만드는 바이러스로서, 접촉에 의하여 전염되며 상피층의 미세한 손상부위를 통하여 침범한다. 현재까지 약 200여종 이상이 알려졌고, 이 중에서 많은 종이 자궁경부암과 밀접한 관련이 있는 것으로 확인되었다. 또 그 중 85 종에 달하는 유전자형(genotype)의 염기 서열이 완전히 밝혀져 있어[2], 암을 진단하는데 분자 생물학적인 방법을 이용하거나, 유전자 칩 등 새로운 의학 도구의 개발 등에 있어 중요한 정보가 되고 있다.

HPV와 연관되어 있는 임상적인 병변은 종류에 따라 자궁경부암을 유발시키는 것과 사마귀 등의 질환을 일으키는 것으로 나누어진다. 예를 들어 HPV-6, -11형들은 주로 바이러스성 사마귀인 콘딜로마(condyloma)를 유발하고 HPV-16, -18형들은 자궁경부암과 관련이 있다. 이는 HPV가 상피세포에 특이적으로 감염되고, 바이러스의 증식주기가 상피세포의 분화정도에 따라 발현되는 특정 요인에 의해 좌우되기 때문이다. 따라서 어떤 요인에 의해 HPV 유형이 나눠지는지 파악하는 것은 아주 중요한 일이다.

이전부터 HPV의 위험군 분류를 위해 여러 연구들이 시행 되어져 왔다. 결정 트리(decision tree)를 이용한 텍스트 마이닝(text mining) 기법이 HPV의 자동 분류에 적용된 연구[3], HPV 서열을 이용하여 커널(kernel)을 기반으로 고위험군의 HPV 유형을 분류하는 연구[4] 등 다양한 방법으로 HPV 유형 분류에 대한 연구를 해왔다.

본 논문은 관찰되어 지지 않은 latent variable을 매개로 하여 두 도메인간의 관계를 학습하는 PLSA 방법을 이용하여 HPV를 클러스터링 해 보았다. 암 발생에 밀접한 관련이 있는 E6 단백질의 서열과 HPV 유형을 사용하여, 하나의 도메인으로 각각 클러스터링 한 것이 아니라, 두 도메인 특성 모두를 동시에 고려하여 클러스터링을 한 것이다. 실험 결과, 특정 클러스터는 질병과 밀접하게 연관되어 있으며, 이와 관련된 주요 서열 분석이 가능함을 보여주고 있다.

2. Latent variable model을 이용한 클러스터링

본 논문에서는 HPV E6 단백질의 서열 데이터와 HPV 유형을 사용해, 두 형태(2-mode)의 데이터를 동시에 분

석 할 수 있는 확률적 모델 PLSA[5]를 이용하여 클러스터링 해 보았다.

데이터의 분포에 대한 제약을 두지 않고 likelihood를 정의하여 클러스터링 모델을 만든다. 즉, K개의 latent variable (Z_k)에 대해 확률 분포를 학습하고, latent variable을 클러스터로 한다. 두 도메인으로 HPV E6 단백질의 서열 데이터와 HPV 유형을 $X = \{x_1, x_2, \ldots, x_N\}$, $Y = \{y_1, y_2, \ldots, y_M\}$로 두고, PLSA를 이용하여 $p(z_k)$, $p(x_i|z_k)$, $p(y_j|z_k)$를 학습하면, 그림 1과 같이 latent variable, HPV 서열, HPV 유형 간의 관계가 학습된다.

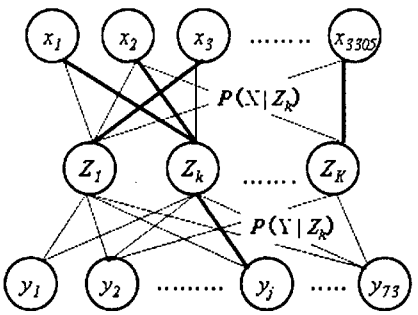

그림 1 PLSA를 이용한 클러스터링

다시 말하면, 이러한 데이터에서 latent variable을 사용한 클러스터링은 관찰되어지지 않은(unobserved) class variable $z \in Z = \{z_1,\ldots,z_k\}$를 매개로 하여 관찰된(observed) 두 도메인 X, Y 간의 관계를 학습하는 것이다[5]. 즉 x_i와 y_j의 결합 확률(joint probability)을 학습하는 것인데, x_i와 y_j의 결합 확률 분포는 다음 식과 같이 정의할 수 있다.

$$p(x_i, y_j) = \sum_{k=1}^{K} p(z_k)p(x_i|z_k)p(y_j|z_k) \quad (1)$$

두 도메인 x_i와 y_j의 결합 확률은 식(1)에서와 같이 클러스터 z_k의 확률 $p(z_k)$, 클러스터 z_k가 데이터 x_i를 생성할 확률 $p(x_i|z_k)$, 클러스터 z_k가 데이터 y_j를 생성할 확률 $p(y_j|z_k)$를 모두 결합한 값이다. 결과적으로는 $p(z_k)$, $p(x_i|z_k)$, $p(y_j|z_k)$를 최대화하는 likelihood를 정의해야 한다. 최대화해야 할 likelihood 함수는 다음 식과 같다.

$$L_{LOG} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(x_i, y_j)\log\, p(x_i, y_j) \quad (2)$$

두개의 도메인 X ,Y에 대해 i번째 서열 데이터의 j번째 HPV 유형의 값 $n(x_i, y_j)$이고, 로그를 취하여 log-likelihood를 최대화 하였다. 이 함수를 최대화하기 위해서 EM (Expectation-Maximization) 알고리즘을 사용하였다[5]. EM 알고리즘에서 E-step은 각 데이터를 적절한 클러스터로 할당하는 단계이고, M-step은 목적 함수를 최대화하기 위해 모델을 갱신하는 단계이다. 이 두 단계를 함수가 최대화 되는 값으로 수렴할 때까지 반복하여 최적화시킨다.

3. 실험 데이터 및 설계

본 논문에서는 Los Alamos National Laboratory의 HPV Database에 있는 73개 유형의 HPV 서열을 이용하였다. 그 중 암 발생에 밀접하게 관련되어 있다고 알려진 E6 유전자의 단백질 서열을 추출하여 실험하였다. 아미노산 20개로 3개씩 조합을 만들어 3-mer 크기의 서열을 만들어, 각 HPV 유형마다 몇 번씩 나타나는지 측정하여 매트릭스를 생성하였다. 생성한 매트릭스를 이용하여 PLSA를 적용시켜 HPV 데이터를 클러스터링 하였다.

4. 실험 결과

본 논문에서는 HPV E6 단백질의 서열 데이터와 HPV 유형을 사용해, PLSA를 이용하여 HPV를 10개의 클러스터로 클러스터링 하였다.

같은 질병을 일으키는 HPV 유형군[6]과 PLSA를 이용하여 HPV를 클러스터링 한 실험 결과는 다음 표와 같다.

Cluster	HPV 유형	임상 병변
Cluster1	Type 16,18,35,39,45	생식기 종양 및 암 (자궁경부암)
Cluster2	Type 16,31,35	후두 및 식도암
Cluster4	Type 31,33,35	Bowen's 질병(피부 종양의 일종)
Cluster3	Type 3,10,28	사마귀, e.verruciformis
Cluster5	Type 6,11	배설기의 condylomas
Cluster6	Type 9,12,17	암으로 진행하는 e.verruciformis
Cluster9	Type 12,14,19,20,21,25,36	epidermodydplasia verruciformis
	Type 5,8	세포 매개 면역 결핍증
Cluster10	Type 15,17,22,23,24	e.verruciformis

표 1 각 클러스터별 $p(y_j|z_k)$값 상위 10순위 중 질병에 관련된 HPV 유형

표 1에서와 같이 클러스터 10개 중 8개의 클러스터에서 질병에 밀접하게 관련되어 있는 HPV 유형들이 클러스터링 되어있음을 보이고 있다. 특히 클러스터 6은 epidermodydplasia verruciformis와 관련되어 있는 HPV 유형 중에서도 암으로 진행되는 유형들만이 클러스터링 되었다. 또 클러스터 9는 10순위 안에 있는 HPV 유형 중 9개의 유형이 질병과 밀접하게 관련되어 있음을 알 수 있다.

또, 클러스터링 결과의 검증을 위해 서열을 기반으로 한 계통발생학적 분석(phylogenetic analysis) 결과[7]와

비교해 보았다. Weighted parsimony 방법을 이용한 HPV E6 계통도는 그림 2와 같다.

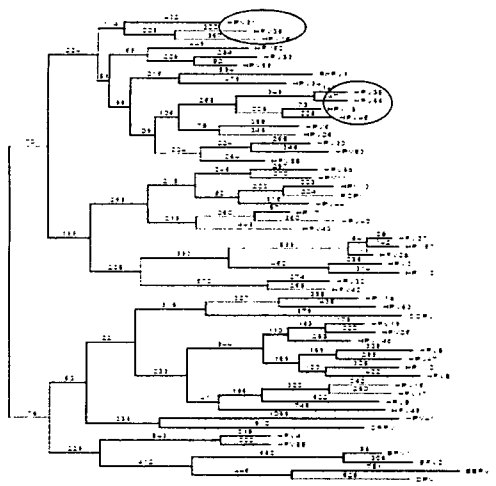

그림 2 Weighted parsimony 방법을 이용한 E6의 계통도

각 클러스터로 클러스터링 된 HPV 유형들이 계통도에서도 가까운 거리에 위치하는 것을 볼 수 있었다. 예를 들어, 클러스터 1의 유형 18, 39, 45를 비롯하여 클러스터 2의 유형 16, 31, 35는 계통도에서 표시된 바와 같이 가까운 거리로 서로 묶여 있음을 알 수 있다.

cluster	1	2	3	4	5	6	9	10
3-mer seq.	TLE	IRC	RRL	RPR	LCH	EKL	TAT	KLD
	RFH	LCD	KPL	EEK	QLC	RTV	FLD	CCS
	KRR	KPL	GLH	VYR	AST	IRC	WKG	ACC
	DLC	FAF	IRC	RFH	VFC	VRN	ATA	EKL
	TTL	EVL	WRG	PRT	CVF	KLL	EKL	LDL
	SVY	WTG	LLC	IRC	VEK	AEK	CGR	EIE
	VYR	VEE	PYG	EKQ	PLC	DLV	IEK	DLL
	PAE	CCK	ISG	WTG	CLE	LCR	CCR	IRC
	RRF	AFR	LRL	HNI	AAC	SLC	ACC	LTV
	IRC	RCC	CLL	SKI	IRC	LEF	LDI	CLK

표 2 클러스터를 구성하는데 주요 인자로 작용한 서열

표 2는 각 클러스터를 구성하는데 있어 주요하게 작용한 서열을 $p(x_i|z_k)$ 값이 높은 순서대로 10순위까지 보여준다. 질병 발생 메커니즘은 HPV 단백질 구조의 영향이 크다. 이러한 단백질의 구조는 서열들에 의해 결정되기 때문에 HPV 유형을 클러스터링 하는데 있어 단백질 서열은 중요한 역할을 할 것이다. 특히, 클러스터 4의 상위 3개의 서열은 암 억제 유전자 p53에 결합하여 세포의 암 억제 기능을 저하시키는 E6의 zinc-binding 영역에 나타나는 서열이다[8].

5. 결론

본 논문에서는 HPV E6의 단백질 서열과 HPV 유형 데이터를 사용하여 latent variable model을 이용한 PLSA 방법으로 서열과 유형 모두를 동시에 고려하여 HPV를 클러스터링 해 보았다.

실험 결과 특정 클러스터는 질병과 밀접하게 연관되어 있으며, 이와 관련된 주요 서열 분석이 가능함을 보여주었다. 이는 전체 서열에서 주요한 서열 조각(segment)들이 클러스터를 결정하는데 있어서 핵심적인 역할을 하고, 주요한 서열들은 실제로 비슷한 질병을 유발할 가능성이 있음을 알 수 있다. 다시 말해, 클러스터 내의 조건부 확률이 높은 서열 조각들은 특정 질병을 유발하는데 있어서 중요하다고 할 수 있다. 따라서 이 주요 서열들을 분석하면 아직 기능이 알려지지 않은 HPV 유형의 다른 유전자의 기능을 유추해 볼 수 있을 것이다.

향후 현 연구와 관련하여 E6을 비롯한 HPV를 구성하고 있는 다른 7개의 유전자에 대해서도 연구가 수행 되어야 할 것이다. 또 이번 실험에서는 서열을 3-mer 크기로 잘라서 실험을 했는데, 다양한 크기로 서열을 잘라 실험해 보는 것도 의미가 있을 것이라 생각이 된다. 더불어 주요 서열에 대한 생물학적인 분석이 수행되어야 할 것이다.

감사의 글

이 논문은 과학기술부 국가지정연구실 사업(NRL)에 의하여 지원되었음.

참고문헌

[1] M. F. Janicek and H. E. Averette, Cervical cancer: prevention, diagnosis, and therapeutics, *Cancer J. Clin*, vol. 51, pp. 92-144, 2001.

[2] H. Pfister, *et al.*, Classification of the papilomavirues-mapping the genome, *Ciba Found. Symp.*, vol. 120, pp. 3-22, 1986.

[3] 정제균, 오석준, 장병탁, Kernel 기반 학습을 이용한 HPV의 위험군 분류, *한국정보과학회 봄 학술발표 논문집 (B)*, 제30권 1호, pp. 428-430, 2003.

[4] 황소현, 박성배, 장병탁, 결정 트리에 의한 인유두종 바이러스의 위험군 분류, *한국데이터마이닝학회 추계학술대회 논문집*, pp. 148-160, 2002.

[5] T. Hofmann, Probabilistic latent semantic analysis, *In Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI99)*, pp. 289-296, 1999.

[6] http://www.dnachip-link.com/library/default.asp

[7] Human Papillomaviruses 1994 Compendium, *Los Alamos National Laboratory*, 1994.

[8] C. G. Ullman, *et al.*, Predicted α-helix/β-sheet secondary structures for the zinc-binding motifs of human papillomavirus E7 and E6 proteins by consensus prediction averaging and spectroscopic studies of E7, *Biochem. J.*, vol. 319, pp. 229-239, 1996.