# Aspect model 기반의 차원 축소를 이용한 유전자 발현데이터 분석

장정호*O 엄재홍* 김유섭** 장병탁*
*서울대학교 컴퓨터공학부 **한림대학교 정보통신공학부
jhchang@bi.snu.ac.krO jheom@bi.snu.ac.kr yskim01@hallym.ac.kr btzhang@bi.snu.ac.kr

# Gene Expression Pattern Analysis Using Aspect Model-based Dimensionality Reduction

Jeong-Ho Chang*O Jae-Hong Eom* Yu-Seop Kim** Byoung-Tak Zhang*
*School of Computer Science and Engineering, Seoul National University
**Division of Information Engineering and Telecommunication, Hallym University

**요 약**

본 논문에서는 aspect model을 이용한 차원 축소 기반의 유전자 발현 데이터 분석을 제시한다. Aspect model은 은닉변수모델의 하나로서, 이를 이용하여 유전자 발현 데이터에 대한 확률적 학습 과정을 통해 특징적 발현 패턴을 추출할 수 있다. 또한 모델로부터 커널함수를 유도함으로써 발현패턴에 기반한 유전자 간의 유사도를 자연스럽게 측정할 수 있다. 모델에 의해 정의되는 은닉공간 차원 수는 데이터 permutation 기반의 검증을 통해 결정한다. 효모(yeast)의 세포 주기(cell cycle) 관련 발현데이터에 대한 실험에서, 주기별 특징 발현 패턴을 추출할 수 있었다. 또한 aspect model로부터 유도된 커널 기반의 유사도 척도를 이용함으로써, 동일 기능 또는 동일 complex 범주에 속하는 유전자 쌍 예측에서 기본적인 상관계수에 의한 방법에 비해 보다 향상된 성능을 얻을 수 있었다.

## 1. 서 론

Oligonucleotide 칩이나 cDNA 칩으로부터의 마이크로어레이 데이터는 특정 실험환경이나 세포 대사 과정하에서의 수백 수천 개 유전자들의 전체적인 발현 양상을 한꺼번에 제공한다. 이러한 대규모의 유전자 발현 데이터에 대한 효과적인 분석을 위해서는 한 번의 실험에서 소규모 유전자 집합에 대한 탐구를 시도하는 것과는 다른 데이터 분석 기법을 필요로 한다.

차원축소(dimensionality reduction) 기법은 군집화(clustering)와 더불어 마이크로어레이 분석을 위해 가장 널리 사용되는 무감독학습(unsupervised learning)에 기반 분석법 중의 하나로서[2], PCA[10], SVD[1], NMF[8] 등의 예가 있다. 대규모 마이크로어레이 데이터에 대한 차원축소에 의한 분석은 내재된 특징적 의미 있는 발현 패턴을 추출하거나 2차원 또는 3차원 공간상에 유전자들을 발현패턴에 따라 가시적으로 제시할 수 있다는 장점이 있다. 또한 차원축소에 의한 고차원 발현 데이터의 저차원으로의 변환은 이후 군집화나 분류 문제를 해결하기 위한 전처리 단계로 활용되기도 한다.

본 논문에서는 은닉변수모델의 일종인 aspect model에 기반한 차원축소에 의한 유전자 발현 데이터의 분석을 제시한다. 유전자 발현데이터는 유전자와 실험조건의 결합쌍(paired) 데이터의 행렬로 표현하고, 각 발현 값은 해당 결합쌍이 구체화된 수치로 간주한다. 이러한 데이터에 대해, aspect model은 은닉변수들을 통해 특징적 발현패턴 추출하고, 또한 학습결과로부터 샘플간의 유사도 측정을 위한 유효한 커널 함수를 제공함으로써 유전자 발현패턴간의 유사도를 효과적으로 측정할 있도록 한다.

2절에서는 aspect model과 그 모델 학습을 설명하고 3절에서는 유전자 발현데이터에 대한 적용과 유전자간의 발현 패턴 유사도 척도를 서술한다. 4절에서는 효모(yeast)의 세포 주기(cell cycle)에 대한 패턴 추출 및 aspect model기반 유사도 척도에 의한 유전자간 상호관계 분석 실험 결과를 제시한다. 5절에서는 결론 및 향후 연구방향을 제시한다.

## 2. Aspect Model

Aspect model은 공기 데이터 분석을 위한 은닉변수 모델의 일종으로서[5], *샘플과 자질들은 각기 해당 이산 확률변수가 갖는 하나의 값으로* 간주된다. 샘플 수가 $n$개이고, 자질 수가 $m$개인 데이터 집합을 $n \times m$ 행렬 $\boldsymbol{D}$로 표현할 때, $\boldsymbol{D}$를 구성하는 각 요소(element)들은 결합쌍 $(x_i, s_j)$에 의해 인덱싱된다. 이 때, aspect model은 이 각각의 결합쌍에 대해 관찰되지 않는 은닉변수 $Z = \{z_1, z_2, \ldots, z_K\}$를 도입하여 이를 확률적으로 모델링한다.

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

즉, 특정 샘플데이터에 대한 자질의 조건부 확률값은 $K$개의 은닉요인들 $(P(s_j|z_k))$의 가중치$(P(z_k|x_i))$가 주어진 합으로 표현된다. Aspect model에 의한 차원 감소는 바로 이 은닉변수의 도입에 의해 이루어진다. 그림 1은 이러한 모델링을 그래프 구조를 이용하여 도시한 것이다.

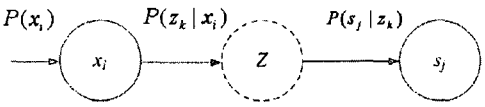


그림 1 aspect model 기반의 데이터 모델링의 그래프에 의한 표현

모델의 학습은 전체 데이터 집합 $\boldsymbol{D}$에 대한 로그 우도 함수 $L$을 최대화하는 모델 매개변수 $P(s_j|z_k)$와 $P(z_k|x_i)$을 추정하는 과정이며, EM 알고리즘이 활용된다.

$$L = \sum_{i=1}^{n}\sum_{j=1}^{m} v(x_i, s_j)\log P(x_i, s_j) \quad (2)$$

$v(x_i, s_j)$은 결합쌍 $(x_i, s_j)$에 대한 실제 데이터 관찰값이다. EM 알고리즘의 E-step에서는 각 결합쌍에 대한 사후 확률 $P(z_k|x_i, s_j)$를 추정하며, M-step에는 이를 이용하여 모델의 매개변수들을 갱신한다[5].

## 3. Aspect model을 이용한 유전자 발현 패턴 분석

### 3.1 데이터 변환

Aspect model을 적용하기 위해 먼저 주어진 유전자 발현 데이터를 일반적인 공기 데이터(cooccurrrence data)와 같은 형태의 빈도수 데이터로 표현할 필요가 있다. 본 논문에서는 이를 위해 우선 각 측정치에 충분히 큰 상수 (예를 들어, 100)을 곱한 후 정수부분의 값만 취하였다.

그리고, 모든 측정치들이 음이 아닌 값을 갖도록 하기 위해 [8]에서 제안된 데이터 중복 표현 (data folding) 방식을 활용하였다. 이 방식에서는 어떤 유전자에 대한 전체 발현 패턴이 벡터 $\boldsymbol{x}$로 주어질 때, 이를 길이가 같은 두 개의 벡터 $\boldsymbol{x}^P$와 $\boldsymbol{x}^N$으로 표현한다.

$$v(x^P, s_j) = \begin{cases} v(x, s_j) & \text{if } v(x, s_j) > 0 \\ 0 & otherwise \end{cases},$$

$$v(x^N, s_j) = \begin{cases} -v(x, s_j) & \text{if } v(x, s_j) < 0 \\ 0 & otherwise \end{cases}$$

요약하자면 원래의 $n \times m$ 유전자 발현 데이터는 $2n \times m$ 행렬로 변환되는 것이다.

### 3.2 은닉공간 차원 설정

은닉변수모델을 실제 문제에 적용할 때 한 가지 고려해야 할 사항은 은닉 변수의 개수 또는 정의되는 은닉공간의 차원수를 어떻게 결정할 것인가에 관한 문제이다. 본 논문에서 데이터 permuation에 기반한 검증 방법을 이용한다. 먼저 데이터 $\boldsymbol{D}$에 대해 임의로 $T$개의 permutation 데이터들($\boldsymbol{D}_P$)을 생성한 후, $\boldsymbol{D}$와 $\boldsymbol{D}_P$들에 대해 각각 aspect 모델을 학습 시킨다. 그런 후, 로그-우도 값 차이 ($L_{diff} = L(D) - L(D_P)/T$)가 가장 큰 $K$값을 모델에서의 주어진 데이터 $\boldsymbol{D}$에 대한 최적 은닉공간 차원으로 결정한다.

### 3.3 유전자 발현 데이터간의 유사도

유전자 발현 데이터를 분석 시 aspect model은 특징적 패턴 추출 기능 외에도 발현 패턴간의 유사도 측정을 위한 Fisher kernel[6] 기반의 척도를 제공한다. 주어진 모델의 매개변수 집합을 $\theta = \{\theta_1, \theta_2\}$, $\theta_1 = \{\sqrt{P(z_k)}\}$, $\theta_2 = \{\sqrt{(P(s_j|z_k)}\}$라고 정의할 때, 매개변수 $\theta_1$과 $\theta_2$에 대해 커널 $k_1$, $k_2$는 다음과 같이 주어진다[4].

$$k_1(x_i, x_j) = \sum_k P(z_k|x_i)P(z_k|x_j)/P(z_k) \quad (3)$$

$$k_2(x_i, x_j) = \sum_l \tilde{P}(s_l|x_i)\tilde{P}(s_l|x_j)\sum_k \frac{P(z_k|x_i, s_l)P(z_k|x_j, s_l)}{P(s_l|z_k)} \quad (4)$$

식 (4)에서 $\tilde{P}(s_l|x_i) = v(x_i, s_l)/\sum_m v(x_i, s_m)$으로 주어진다. 간략히 말해, 커널 $k_1$은 은닉공간상에서의 두 유전자간의 유사도이고, $k_2$는 입력공간상에서의 유사도라고 할 수 있다. 본 논문에서는 하나의 유전자 발현 벡터를 두 개의 벡터로 나누어서 표현하기 때문에, 실제 두 유전자에 대한 커널은 다음과 같이 정의된다.

$$k_a(x_i, x_j) = k_a(x_i^P, x_j^P) + k_a(x_i^N, x_j^N), \qquad a = 1, 2.$$

## 4. 실 험

### 4.1 데이터

Spellman et al.[11]의 Saccharomyces cerevisiae의 세포 주기 관련 발현 데이터 중 $\alpha$-factor 데이터에 대해 실험하였다. $\alpha$-factor 발현 데이터는 $\alpha$-mating factor 페로몬을 이용하여 세포들을 G1기에 정지(arrest)시켰다가 해제(release)시킨 후 매 7분마다 유전자들의 발현양상을 측정한 것으로 6,000개 이상의 효모 유전자들에 대해 총 18개의 시점(timepoint)에서의 측정치를 담고 있다. 전체 유전자들 중에서 800개의 유전자들이 세포주기에 관여하는 것으로 추정되었는데, 본 논문에서는 이 중 결측치(missing value)가 하나 이하인 753개의 유전자를 선택하였다. 결측치들은 $k$-최근접 이웃 방법[12]을 이용하여 채워넣었는데, 거리 척도는 유클리디안 거리를 이용하고 $k$값은 10으로 하였다. 그리고 나서 각 유전자에 대해 발현값은 평균이 0이고 표준편차값은 1의 값을 갖도록 표준 정규화한 후, 3.1절에서 설명된 전처리 방법에 의해 모든 발현값이 양의 값을 갖도록 변환하였다. 최종적으로 데이터는 $1506 \times 18$ 행렬로 주어진다.

### 4.2 특징적 발현 패턴 추출

Aspect model을 적용하여 먼저 내재된 주요 발현 패턴의 추출에 관한 실험을 진행하였다. 모델의 은닉공간 차원 수는 3.2절에서 설명한 방법을 이용하여 결정하였다. 원래 데이터에 대해 총 100개의 임의 데이터를 생성하였다. 모델 학습시 매개변수의 초기화에 따라 학습 결과는 영향을 받으므로, 각 차원에 대해 임의 초기치를 가지고 EM 학습을 20번 수행한 후, 로그-우도(log-likelihood)값이 가장 높은 모델을 선택하였다. 이는 원래 데이터와 임의 데이터 모두에 해당된다. 그림 2는 모델의 은닉공간 차원에 따른 원래 데이터와 임의 데이터에 대한 로그-우도 값의 차이 $L_{diff}$를 보인다. 차원이 5일 때 그 값이 가장 높음을 알 수 있으며, 따라서 aspect model에서의 은닉공간 차원은 5로 설정하였다.

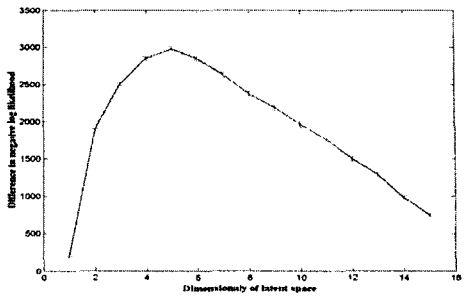


그림 2. Aspect model에서의 은닉공간 차원 수에 따른 로그-우도(log-likelihood) 값 차이의 변화

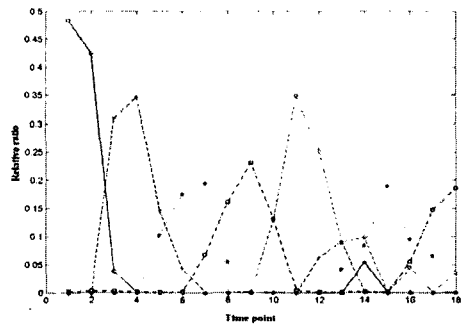


그림 3. 효모 세포 주기 데이터에서 추출된 특징적 발현패턴

그림 3은 aspect model의 은닉변수에 의해 추출된 5개의 특징적 패턴을 보인다. 데이터 전처리 시 평균보다 억제된 발현 측정치는 모두 양의 값으로 변환되었기에 약간의 차이는 있지만, 분석 결과는 [11]에서 분석된 세포 주기별 특징적 패턴과 상당히 유사한 결과를 보인다. 예를 들어 시간점 1과 2에서 높은 발현양상을 보이는 것은 대략 M/G1 주기에 관련된 유전자들의 전형적 발현 패턴이고 시간점 3과 4의 경우 G1 주기에 해당되며 시간점 9와 18에서의 높은 발현 양상은 M 주기에 해당된다. 이를 통해, aspect model을 적용함으로써 발현 패턴상의 내재된 특징적 형태를 효과적으로 추출할 수 있음을 알 수 있다.

### 4.3 유전자쌍에 대한 관계 분석

유전자들간에 발현패턴이 비슷하다는 것은 해당 유전자들이 비슷한 세포내 기능을 갖거나 해당 단백질 산물들이 물리적으로 상호작용할 가능성이 있음을 의미할 수 있다[3, 7]. 이러한 가정하에, 분석된 753개의 유전자들의 각 쌍들에 대해 유사도를 측정하고, 그 결과를 MIPS Comprehensive Yeast Genome Database(CYGD)[1][9]로부터 얻은 두 종류의 데이터에 대해 검증하였다. 하나는 functional catalogue 데이터(본 논문에서는, 2번째 level까지만 선정)이고, 다른 하나는 protein complex catalogue 데이터이다.

이 실험을 위해, 먼저 모든 유전자 쌍에 대해 유사도를 측정하였으

1) http://mips.gsf.de/genre/proj/yeast/

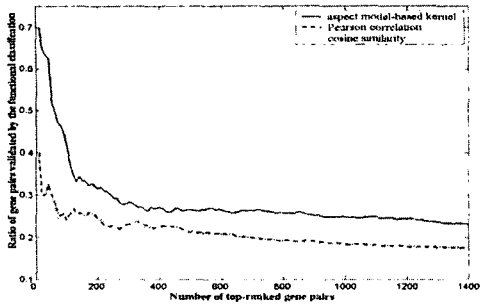


그림 4. MIPS functional catalogue에 대한 결과

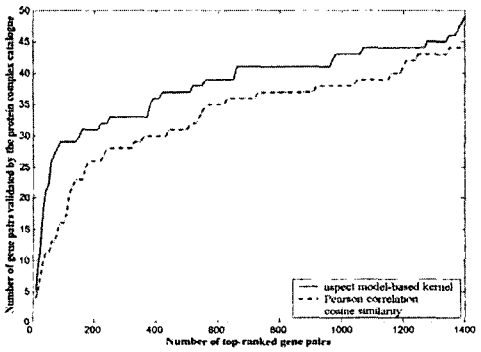


그림 5 MIPS protein complex catalogue에 대한 결과

며, 유사도 계산은 3.3절에서 설명된 두 개의 커널 함수들을 조합하여 계산하였다. 즉, 두 유전자의 발현 패턴 $x_i$와 $x_j$간의 유사도는 $sim(i,j)=k_1(x_i,x_j)\times k_2^{norm}(x_i,x_j)$로 정의하였다. 여기서, $k_2^{norm}(x_i,x_j)$는 $k_2(x_i,x_j)$의 정규화된 값이다 ($0 \le k_2^{norm}(x_i,x_j) \le 1$). 모든 유전자쌍 (총 238,128쌍)을 그 유사도 값에 따라 내림차순으로 정렬한 다음, MIPS의 두 데이터에 대해 검증하였다. 각 유전자 쌍은 동일한 기능을 갖거나(functional catalogue) 동일 complex를 구성할 때(protein complex catalogue), 올바르게 예측된 것으로 간주된다.

그림 4는 MIPS functional catalogue에 대한 검증결과로서, 전체 쌍 중 상위 1,400쌍에 대한 결과를 보인다. 두 경우 모두 aspect model 기반의 커널에 의한 방법이 기본적인 Pearson 상관 계수에 의한 방법보다 향상된 성능을 보였다. 특히 유사도 정렬 리스트 상의 상위 유전자쌍들에 대해 그러한 경향이 두드러짐을 알 수 있다. 검증된 유전자 쌍들로는 *HHF1-HHT1*(mRNA transcription), *HTA2-HHF1*(mRNA transcription), *MSH2-CTF2*(DNA processing), *CTF4-DPB2*(DNA processing) 등이 있다. 그림 5는 MIPS protein complex catalogue에 대한 결과이며, functional catalogue에 대한 결과와 마찬가지로 aspect model 기반 커널 방법이 보다 더 우수한 성능을 보임을 알 수 있다.

이러한 성능 차이가 단지 3.1 절에 명시된 데이터 변환에 의한 것인지를 알아보기 위해, 모델 학습 없이 오직 변환된 데이터에 대한 코사인 (cosine) 척도를 유전자 발현패턴 간 유사도로 설정하고 앞서와 같은 실험을 하였다. 이 경우 변환 이전의 데이터에 대한 Pearson 상관 계수에 의한 결과와 비교해 볼 때 성능차가 거의 없었다(그림 4, 5). 따라서 데이터 변환 그 자체 보다는 aspect model 학습으로부터 유도된 커널에 기초한 유사도 측정이 그림 4와 5에 제시된 성능 향상의 주 요인이라고 할 수 있다.

## 5. 결 론

본 논문에서는 은닉변수모델의 하나인 aspect model기반의 차원 축소에 의한 유전자 발현 데이터 분석 기법을 제시하였다. 효모의 세포 주기 관련 데이터에 대해 적용하여, 내재된 특징적 패턴들을 잘 추출할 수 있었으며 이는 기존에 알려진 세포 주기와 어느 정도 일치함을 보였다. 모델의 은닉변수 차원은 데이터 permutation 기반의 테스트를 통해 결정하였다. 특히 모델로부터 유도된 kernel 함수를 두 유전자간의 발현패턴 상의 유사도를 계산하는 데 적용하였을 때, 기존의 전통적인 상관계수에 의한 방법에 비해 보다 향상된 성능을 보였다. 이와 같은 모델로부터 유도된 커널 기반의 유사도 척도는 향후 데이터 샘플 또는 유전자들의 군집화 작업에 충분히 활용 가능할 것이다.

## 감사의 글

본 연구는 과학기술부 국가지정연구실(NRL)사업에 의해 지원되었음.

## 참고문헌

[1] Alter, O., Brown, P. O., and Botstein, D., Singular value decomposition for genome-wide expression data processing, *Proceedings of the National Academic Sciences*, vol. 97, no. 18, 10101-10106, 2000

[2] Baldi, P. and Hatfield, G. W., *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, 2002

[3] Ge, H., Liu, Z., Church, G., and Vidal, M., Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae, *Nature Genetics*, vol. 29, pp. 482-486, 2001

[4] Hofmann, T., Learning the similarity of documents: an information-geometric approach to document retrieval and categorization, In *Advances in Neural Information Systems* 12, pp. 914-920, 2000

[5] Hofmann, T., Unsupervised learning by probabilistic latent semantic analysis, *Machine Learning*, vol. 42, pp. 177-196, 2001

[6] Jaakkola, T. and Haussler, D., Exploiting generative models in discriminative classifiers, In *Advances in Neural Information Processing Systems* 11, pp. 487-493, 1999

[7] Jansen, R., Greenbaum, D., and Gerstein, M., Relating whole genome expression data with protein-protein interactions, *Genome Research*, vol. 12, pp. 37-46, 2002

[8] Kim, P. M. and Tidor, B., Subsystem identification through dimensionality reduction of large-scale gene expression data, *Genome Research*, vol. 13, pp. 1706-1718, 2003

[9] Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, S., Rudd, S., and Weil, B., MIPS: a database for genomes and protein sequences, *Nucleic Acids Research*, vol. 30, pp. 31-34, 2002

[10] Raychaudhuri, S., Stuart, J. M., and Altman, R. B., Principal component analysis to summarize microarray experiments: application to sporulation time series, In *Proceedings of Pacific Symposium on Biocomputing*, vol. 5, pp. 236-244, 2000

[11] 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*, vol. 9, pp. 3273-3297, 1998

[12] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R., Missing value estimation methods for DNA microarrays, *Bioinformatics*, vol. 17, pp. 520-525, 2001