Neural Feature Association Rule을 이용한 효모 단백질-단백질 상호작용의 예측

엄재홍[0] 장병탁
서울대학교 컴퓨터공학부
{jheom, btzhang}@bi.snu.ac.kr

Prediction of Yeast Protein-Protein Interactions by Neural Feature Association Rule

Jae-Hong Eom Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

단백질들은 서로 다른 단백질들과 상호작용하거나 복합물을 형성함으로써 생물학적으로 중요한 기능을 한다고 알려져 있다. 때문에 대부분의 세포작용에 있어 중요한 역할을 하는 단백질들 간의 상호작용 분석 및 예측에 대한 연구는 여러 연구그룹으로부터 풍부한 데이터가 산출된 후게놈시대(post-genomic era)에서 또 하나의 중요한 이슈가 되고 있다. 본 논문에서는 효모에 대해 공개되어있는 단백질 상호작용 데이터들에서 속성들 간의 연관규칙 학습을 통해 잠재적 단백질 상호작용들을 예측하기 위한 연관규칙 기반의 상호작용 예측 방법을 제시한다. 단백질들 간의 상호작용 예측을 위해 고려되는 각 단백질의 다수의 속성차원은 정보이론 기반의 속성선택 알고리즘을 이용하여 효율적으로 줄이며 상호작용의 속성집합을 이용하여 신경망을 훈련시키고 이렇게 훈련된 신경망에서 속성들 간의 연관규칙을 디코딩하여 연관규칙 기반의 상호작용 예측에 활용한다. 연관속성 발굴을 통한 상호작용 예측을 위한 마이닝 방법으로는 연관규칙 발견 알고리즘을 사용하였으며 예측 정확도를 높이기 위하여 신경망 예측 모델의 학습 결과를 디코딩한 규칙들이 추가적으로 사용하였다. 논문에서 제안한 방법은 발견된 연관규칙을 통한 단백질 상호작용 예측문제에 있어 평균 약 94.8%의 예측 정확도를 보였다.

1. 서론

1.1 연구의 배경

단백질-단백질의 상호작용(PPI; protein-protein interaction)이 생물의 기관에서 일어나는 매우 기초적인 생화학 반응들 중의 하나이며 여러 가지 생물학적 반응 과정에서 중요한 기능을 한다는 것은 이미 여러 연구자들에 의해 밝혀졌으며 지금까지도 다양한 생물학 도메인(domain)에 대하여 활발하게 연구되고 있다[1].

효모(yeast)는 비교적 단순한 구조, 높은 대사활성 및 빠른 성장속도와 함께 그 응용의 다양성 등으로 인해 많은 연구가 이루어진 생물중의 하나이다. 1997년 발아효모(budding yeast; *Saccharomyces cerevisiae*)의 DNA 염기서열이 밝혀진 후 많은 연구자들이 약 6,300여개가 넘는 효모 단백질들의 기능적 분석과 관련된 연구를 수행하였으며, 이에 따라 효모의 단백질-단백질 상호작용과 관련된 풍부한 실험적 데이터 및 분석적 데이터가 존재하게 되었다[2].

1.2 관련 연구

유전자발현 데이터나 단백질 상호작용 데이터와 같은 여러 데이터들을 이용한 단백질 기능 및 상호작용 예측과 관련된 여러 연구들이 수행되어 왔다. Eisen, Pavlidis등은 유사기능 유전자들의 발현 특성을 이용하여 유전자발현 데이터에 대한 군집화 분석을 시도하였다[3][4]. Wu등은 효모 염기서열 전사클러스터들 간의 중복성 분석을 통하여 효모 단백질의 기능을 분석하였다[5]. 또한, Park등은 진화적으로 연관성 있는 단백질 도메인들의 구조적 그룹들 간의 상호작용 분석법을 이용하여 단백질들 도메인들 간의 상호작용을 분석하였으며[6], Iossifov, Ng등은 밝혀진 단백질들 간의 상호작용 데이터를 이용하여 각각 확률추론과 계산학적 추론방법을 이용하여 새로운 상호작용을 예측하였다[7][8]. Ito등은 Y2H (yeast two-hybrid)[9] 방법을 이용하여 발아효모 단백질들의 기능적 상호작용에 대한 포괄적인 분석을 하였고[10][11], Uetz 등은 어레이스크리닝과 라이브러리스크리닝을 함께 이용하여 효모 단백질의 광범위한 단백질-단백질 상호작용을 조사하였다[12]. Bu등은 단백질 상호작용 네트워크 구조분석을 통해서 새로운 상호작용을 예측하였다[13].

본 논문에서는 단백질들의 상호작용을 속성집합들 간의 상호작용으로 고려하고 이들 속성들 간의 연관성을 연관규칙발견(association rule discovery) 방법을 이용하여 추출한 후 이를 이용하여 새로운 상호작용을 찾는 방법을 제시한다. 또한 각각의 단백질들에 대하여 고려되는 다수의 속성들로 인한 속성차원 문제를 정보이론 기반의 속성필터로 줄여 사용하며 단순연관규칙 추출 결과를 향상시키기 위하여 상호작용하는 단백질들의 속성벡터들을 이용하여 신경망을 학습하고 학습된 신경망에서 연관규칙을 추출하는 방법을 사용하였다.

2. 연관규칙 기반의 단백질-단백질 상호작용 예측

2.1 상호작용들에 대한 속성벡터 인코딩

상호작용하는 각 단백질 쌍의 속성은 그림 1과 같이 해당 속성의 존재 여부를 이진 코드로 인코딩 하였다.

2.2 속성차원의 축소

논문에서 상호작용하는 단백질들에 대해 고려하는 속성들은 약 6,000여개가 넘으며 이같이 많은 속성들에는 상호작용을 예측하는데 불필요한 다수의 속성들이 존재할 수 있다. 논문에서는 Eom등이 사용한 속성차원축소필터 FDRF[14]를 이용하여 그림 1의 우측과 같이 속성차원(속성들)을 축소(필터링)하였다. FDRF에서는 속성들의 엔트로피를 이용한 정보이득을 계산하여 상대적으로 많은 값을 가지는 속성으로 정보이득 값이 편향되는 것을 방지하기위하여 새로운 속성 평가척도 *SU*[14]를 사용하며 *SU*는 다음과 같이 정의된다.

$$SU(X,Y)=2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right]$$

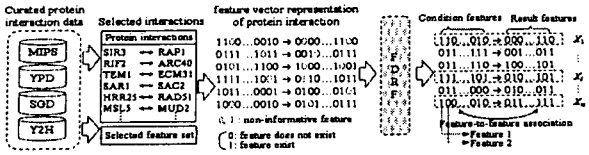

그림 1. 상호작용하는 단백질의 이진 속성벡터 표현. 각 단백질들의 속성 벡터는 그림의 왼쪽과 같은 데이터베이스와 Oyama 등의 실험방법[15]을 이용하여 구성하였다.

2.3 속성 연관규칙의 추출

Oyama[15]등은 가장먼저 단백질 상호작용을 단백질 속성들 간의 상호작용으로 살펴보았다. 본 논문에는 Oyama등이 사용한 속성들의 정의와 속성 코딩 등의 기본적인 연관규칙 추출 방법을 사용하였다. 일반적으로 연관규칙(association rule)은 지지도 $SP(A\Rightarrow B)=P(A\cup B)$와 신뢰도 $CF(A\Rightarrow B)=P(B|A)$ 값으로 해당 규칙의 적용 범위와 규칙의 질을 나타낸다. 본 논문에서는 Oyama 방법에 기초한 Eom [14]등의 방법을 사용하였다.

2.4 신경망을 이용한 속성들 간의 연관성 학습

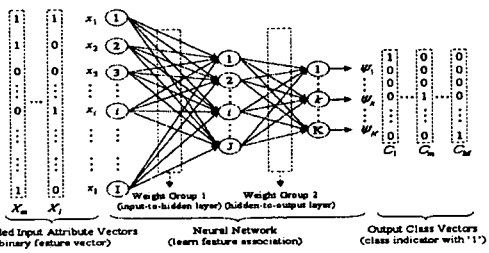

그림 2. 단백질 상호작용 예측을 위한 상호작용 단백질들의 속성 학습 신경망 모델. 2.1에서 인코딩된 상호작용별 속성 벡터는 신경망의 입력으로 제공되고 신경망은 벡터속성들 간의 연관성을 학습한다.

그림 2에서와 같이 신경망은 상호작용하는 단백질들의 속성벡터를 이용하여 학습되며 학습이 학습된 결과는 2.5와 같은 과정을 통하여 속성들 간의 연관 규칙으로 디코딩된다.

2.5 신경망으로부터의 학습된 속성 연관규칙 추출

N개의 속성 $A_n(n=1,\cdots,N)$은 고정길이의 벡터 $x_1,\cdots,x_i,\cdots,x_{m(n)}$으로 표현할 수 있다 ($m(n)$은 속성 A_n이 가질 수 있는 가능한 값의 수). 이때 신경망의 입력은 아래의 I와 같으며 신경망의 입력 속성벡터는 $X_m=(x_1,\cdots,x_i,\cdots,x_I)$, $X_m=m=(1,\cdots,2,\cdots,M)$과 같이 표현 된다 ($M$은 입력 후련 패턴의 총 개수). 이때, 고정 벡터길이 K의 출력 클래스 벡터 $C_k(k=1,2,\cdots,K)$는 아래와 같다. 이러한 신경망은 그림 2와 같이 구성되며 은닉층으로의 입력 IHN_j 은닉층의 출력 OHN_j 출력노드로의 입력 ION_k은 각각 아래와 같이 정의할 수 있다 (WG는 가중치들의 그룹).

$$I=\sum_{n=1}^{N}m(n),\qquad C_k(\Psi_1,\cdots,\Psi_k,\cdots,\Psi_K)$$

$$IHN_j=\sum_{i=1}^{I}x_i(WG1)_{i,j},\qquad OHN_j=\frac{1}{1+e^{-\left[\sum_i^I x_i(WG1)_{i,j}\right]}}$$

$$ION_k=\sum_{j=1}^{J}(WG2)_{j,k}\frac{1}{1+e^{-\left[\sum_i^I x_i(WG1)_{i,j}\right]}}$$

$$\Psi_k=\left\{\frac{1}{1+e^{-\left[\sum_j^J WG2_{j,k}\left(\frac{1}{1+e^{-\left[\sum_i^I x_i(WG1)_{i,j}\right]}}\right)\right]}}\right\}$$

이 경우 출력함수 $\Psi_k=f(x_i,(WG1)_{i,j},(WG2)_{j,k})$는 $WG1$·2가 각각 상수이기 때문에 x_i에 대한 지수함수가 된다. 따라서, 신경망에서 규칙을 구성하는 작업은 입력 속성들 중에서 Ψ_k를 최대화 시키는 입력 속성벡터 X_m을 찾아내어 규칙으로 구성함으로써 처리할 수 있다. 이 최적화 문제는 이진속성벡터 x를 고려한 $\Psi_k(x_i)$를 목적함수로 갖는 비선형 정수최적화 문제로 생각 할 수 있다. 논문에서는 GA를 이용하여 아래의 목적함수를 최적화 시키는 이진속성벡터를 찾아 규칙화하여 새로운 상호작용 예측에 사용한다. 신경망에서 원하는 출력 값에 부합하는 이진속성벡터는 그림 3의 과정을 거쳐 규칙으로 변환된다.

With the given parameters,
- A : set of attributes.
- α : set of conditional attributes, β: set of result attributes.
- n : the number of total attribute, γ: the length of feature n.
- G : set of the best b chromosome, g: a chromosome in G.
- b : the number of total chromosome ($|G|=b$).
- μ : the number of total rule found by association rule mining.

Repeat Step 1 to Step 5, for all g in G.

1. **Create** temporary empty rule t $(\alpha)\rightarrow(\beta)$, **Set** $\alpha=\beta=\phi$.
2. **Divide** best chromosome into $2n$ segments.
   (Each segment in 1 to n is corresponds to each attribute of A_n for condition of rule and each segment in $n+1$ to $2n$ is corresponds to the each rest attribute of A_n for result of rule).
3. For all i, i = 1 to n.
   3.1 For all j, j = 1 to γ.
   3.1.1 If the corresponding bit of conditional chromosome is equal to '1 then **Update** α, $\alpha\leftarrow\alpha\cup A_j$.
   3.2 **Connect** all feature in α with operator 'AND'.
4. For all i, i = $n+1$ to $2n$.
   4.1 For all j, j = 1 to γ.
   4.1.1 If the corresponding bit of result chromosome is equal to '1 then **Update** β, $\beta\leftarrow\beta\cup A_j$.
   4.2 **Connect** all feature in β with operator 'AND'.
5. For all k, k = 1 to μ.
   5.1 If any $R(k)\equiv t$ then $R\leftarrow R-R(k)$ else $R\leftarrow R\cup t$.

Return final rule set R.
(R = rules mined by association mining $\cup$ rules decoded from top b chromosomes)

그림 3. 최적 Chromosome에서의 속성연관규칙 추출 과정.

표 1. 실험에 사용한 상호작용 데이터 집합 및 상호작용 개수

데이터베이스	상호작용 개수	속성 개수	선택된 속성 개수
MIPS	10,641		
YPD	2,952		
SGD	1,482	6,232 (total)	1,293 (total)
Y2H(by Ito)	957		
Y2H(by Uetz)	5,086		

2.6 연관규칙을 이용한 상호작용의 예측

앞의 2.3~2.5의 단계를 거쳐 추출되고 디코딩된 규칙들은 본래의 속성벡터들에 대하여 구해진 연관규칙들과 함께 최종적으로 테스트 데이터의 상호작용 예측에 사용되었다. 각 방법들의 효과는 표 3과 같이 각 방법들을 조합해 봄으로써 살펴보았다.

3. 실험, 결과 및 분석

3.1 데이터 집합

실험데이터는 대표적인 단백질 상호작용 데이터베이스인 MIPS, SGD, YPD등을 활용하였으며 Ito와 Uetz의 Y2H 실험결과도 함께 사용하였다. 아래의 표 1은 논문에서 실험에 사용한 상호작용 데이터의 통계정보를 나타낸다.

3.2 실험 결과

표 3은 정보이론 기반의 속성 필터링과 속성 연관규칙 기반의 예측 및 신경망 학습결과의 디코딩을 통해 얻은 연관규칙의 조합에 따른 상호작용 예측 정확률 변화를 나타낸다. 상호작용 단백질들의 속성에 대한 속성 필터링과 이러한 속성들에 대하여 추출된 연관규칙 및 필터링 되지 않은 속성벡터를 이용하여 학습한 신경망 학습결과에서 추출된 연관규칙을 이용한 예측 방법(☆)인 ⓓ 조합이 가장 좋은 예측성능을 보였다.

표 3. MIPS 데이터에 대한 제안한 방법의 예측 결과 (표에서 'Asc.'는 연관규칙 기반의 예측을, 'FDRF+Asc.'는 정보이론 기반의 속성필터링 후의 속성벡터를 이용해 추출된 연관규칙 기반의 예측을, 그리고 'Asc.+N-Asc.'는 속성벡터에서 추출된 연관규칙과 속성벡터를 이용하여 훈련된 신경망에서 추출된 연관규칙들 기반의 예측을 의미하여 'FDRF+Asc.+N-Asc.'는 속성벡터에 대한 필터링과 필터링 된 속성벡터를 기반으로 추출된 연관규칙 및 필터링 되지 않은 속성벡터에 대해 학습된 신경망에서 추출된 연관규칙 기반의 예측 방법을 의미).

예측방법	상호작용 개수			예측률 (|P|/|T|)
	훈련 집합	테스트집합 (T)	예측된 상호작용(P)	
ⓐ: Asc.	4,628	463	423	91.4%
ⓑ: FDRF+ⓐ	4,628	463	439	94.8%
ⓒ: ⓐ+N-Asc.	4,628	463	432	93.3%
ⓓ: ⓑ+N-Asc.(☆)	4,628	463	445	96.1%

3.3 결과 분석

표 3에서 볼 수 있듯이 단순히 속성벡터들에서 추출된 연관규칙들을 이용한 예측 방법(ⓐ)이 가장 낮은 성능을 보였다. 이러한 결과는 지나치게 많은 속성들(약 6,232개)을 연관규칙 학습에 사용함으로써 의미 없는 속성들에 대한 잘못된 연관규칙을 구성하게 되었고 결국 예측 성능을 저하시킨 것으로 생각된다. 이는 속성벡터들에 대하여 정보이론 기반의 필터링을 적용한 ⓑ가 ⓐ보다 약 3.4% 좋은 성능을 내는 것으로 확인 할 수 있었다. ⓑ의 결과는 속성벡터를 적절히 필터링하여 보다 의미 있는 속성들을 골라내는 것이 유용하다는 것을 나타낸다.

단순연관규칙의 추론과 신경망을 이용해 구성한 연관규칙을 함께 사용한 ⓒ는 단순 연관규칙만을 이용한 예측 결과인 ⓐ에 비해 약 1.9%의 성능 향상을 보였다. 이는 신경망을 이용한 연관규칙의 확장이 어느 정도 의미가 있음을 나타낸다고 할 수 있다. 그렇지만 성능향상이 ⓑ의 결과보다 낮은 1.9%에 머무른 것은 ⓒ의 방법에서도 역시 정보성이 적은 속성들을 필터링 할 필요가 있음을 나타낸다. 속성벡터에 대한 필터링 적용 후 추출한 연관규칙과 신경망을 이용하여 추가로 획득한 연관규칙을 이용한 예측 방법인 ⓓ는 의미 있는 속성들의 선택과 신경망을 이용한 속성들 간의 비선형적 연관성을 학습하고 규칙으로 구성할 수 있어 ⓐ에 비하여 약 4.7%의 예측 성능 향상을 보였다고 생각된다.

4. 결론 및 향후 과제

본 논문에서는 단백질들 간의 상호작용을 각 단백질들이 가지는 다수의 속성들 간의 상호작용으로 고려하고, 상호작용하는 단백질의 속성들 간의 연관성을 연관규칙 추론 방법을 이용하여 추론하고 이를 이용하여 새로운 단백질 상호작용을 예측하는 방법을 제시하였다. 각 단백질들의 속성차원은 정보이론 기반의 필터링을 통하여 효율적으로 감소시킬 수 있었다. 또한, 상호작용하는 단백질들의 속성 집합들 간의 연관 관계를 신경망으로 학습한 후, 신경망 학습 결과를 일종의 연관규칙으로 디코딩하는 방법으로 속성들 간의 비선형적 특성을 상호작용 예측에 활용할 수 있음을 살펴볼 수 있었다. 즉, 논문에서 제시한 방법들은 특히 많은 속성들을 가지는 개체들 간의 상호작용 예측과 같은 문제의 전처리 과정으로 유용하게 활용될 수 있을 것이다.

그렇지만 논문에서 사용한 상호작용 데이터들에는 잘못된 상호작용(false positive · negative)들이 어느 정도 포함되어 있으며 이러한 결과들은 최근 추가적인 실험들을 통해서 수정되고 있다. 때문에 보다 정확한 속성들의 연관규칙을 학습하고 예측에 사용하기 위해서는 이와 같은 잘못된 상호작용을 에러로 고려할 수 있는 모델을 구상하는 것이 필요하며 보다 생물학적으로 의미 있는 속성들의 발굴과 인코딩 및 전역적인 상호작용에 대하여 인과관계 네트워크 구축이 가능한 베이지안네트워크와 같은 통계량 기반의 모델을 이용한 추가적인 분석이 필요할 것으로 생각된다.

감사의 글

본 연구는 교육부 BK21-IT 프로그램 및 과학기술부 국가지정연구실(NRL) 사업에 의하여 일부 지원되었음을 밝힙니다.

참고문헌

[1] Deng, M. *et al.*, "Inferring domain-domain interactions from protein-protein interactions," *Genome Res.*, 12(10), pp. 1540–1548, 2002.

[2] Goffeau, A. *et al.*, "Life with 6000 genes," *Science*, 274, pp. 563–567, 1996.

[3] Eisen, M. B. *et al.*, "Cluster analysis and display of genomewide expression patterns," *Proc. Natl. Acad. Sci.*, 95, pp. 14863–14868, 1998.

[4] Pavlidis, P. and Weston, J., "Gene functional classification from heterogeneous data," In *Proc. of the 5th International Conference on Computational Molecular Biology* (RECOMB-2001), pp. 249–55, 2001.

[5] Wu, L. F. *et al.*, "Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters," *Nature Genetics*, 31, pp. 255–265, 2002.

[6] Park, J. *et al.*, "Mapping protein family interactions: intra-molecular and intermolecular protein family interaction repertoires in the PDB and yeast," *J. Mol. Biol.*, 307, pp. 929–39, 2001.

[7] Iossifov, I. *et al.*, "Probabilistic inference of molecular networks from noisy data sources," *Bioinformatics*, 20(8), pp. 1205–12013, 2004.

[8] Ng, S. K. *et al.*, "Integrative approach for computationally inferring protein domain interactions," *Bioinformatics*, 19(8), pp. 923–29, 2003.

[9] Fields, S. and Sternglanz, R., "The two-hybrid system: an assay for protein-protein interactions," *Trends in Genetics*, 10, pp. 286–92, 1994.

[10] Ito, T. *et al.*, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc. Natl Acad. Sci.*, 98, pp. 4569–4574, 2001.

[11] Ito, T. *et al.*, "Novel modular domain PB1 recognizes PC motif to mediate functional protein–protein interactions," *EMBO J.*, 20, pp. 3938–3946, 2001.

[12] Uetz, P. *et al.*, "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*," *Nature*, 403(6770), pp. 623–627, 2000.

[13] Bu, D. *et al.*, "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Res.*, 31(9), pp. 2443–2250, 2003.

[14] Eom, J.-H. and Zhang, B.-T., "Prediction of implicit protein-protein interaction by optimal associative feature mining", *LNCS*, 3177, pp. 85-91, 2004.

[15] Oyama, T. *et al.*, "Extraction of knowledge on protein–protein interaction by association rule discovery," *Bioinformatics*, 18(5), pp. 705–714, 2002. ■