연관속성개념공간으로의 사상을 이용한 단백질 상호작용 예측

엄재홍[0] 장병탁

바이오지능연구실
서 울 대 학 교 컴 퓨 터 공 학 부
{jheom, btzhang}@bi.snu.ac.kr

Prediction of Protein Interactions using the Associative Feature Concept Space Mapping

Jae-Hong Eom Byoung-Tak Zhang

Biointelligence Laboratory
School of Computer Science and Engineering, Seoul National University

요 약

생물체 내에서 중요 생물학적 기능을 수행하는 기본 단위인 단백질 및 이들의 상호작용 대한 많은 연구가 이루어져 다양한 생물체에 대한 단백질 상호작용 데이터베이스가 구축되었다. 본 논문에서는 효모에 대해 공개되어있는 단백질 상호작용 데이터를 이용하여 새로운 단백질 상호작용을 예측하는 방법을 제안한다. 논문에서는 문헌에서 연관 정보를 효율적으로 찾아내기 위하여 제안된 연관개념공간 탐색 방법을 확장하여 단백질 상호작용 예측에 사용한다. 단백질들은 각각이 가지는 다양한 속성들의 벡터로 간주되며, 상호작용은 해당 단백질들의 연관성을 통해 이루어지는 것으로 표현된다. 상호작용하는 두 단백질들의 속성은 단어의 공동 출현과 같이 고려되어 단백질 상호작용은 두 단백질 벡터의 요소로 표현되고 벡터의 요소 속성들 간의 연관성을 표현하기위해 연관속성개념공간으로 사상되어 공간상의 거리 기반으로 연관속성을 추출한다. 추출된 연관속성을 최대로 포함하는 단백질들 간의 상호작용을 예측하는 방식으로 단백질 상호작용을 예측한다. 논문에서 제안한 방법은 효모의 단백질 상호작용 예측에 대해 평균 약 91.8%의 예측 정확도를 보여, 연관속성개념공간을 이용한 방법이 단백질 상호작용을 예측하는 또 다른 대안으로 사용 될 수 있음을 확인하였다.

1. 서론

1.1 연구의 배경

생물체 내부에서의 일어나는 중요 생·화학적 반응들은 대부분 해당 생물체의 DNA상의 유전자들에 의해 조절된다. 유전자는 생물체에 따라 수백~수천, 또는 수만 개가 존재한다. 이들은 상호작용을 위해 mRNA를 통해 해석되어 단백질(protein)로 구성되고 이 단백질들에 의해 생물체 내에서의 대부분의 실질적 상호작용이 일어난다. 때문에 단백질-단백질의 상호작용(PPI; protein-protein interaction)을 생물의 기관에서 일어나는 매우 기초적인 생화학 반응들 중의 하나로 간주되고 있다.

단백질 상호작용이 생물학적 반응과정에서 중요한 기능을 한다는 것은 이미 여러 연구자들에 의해 밝혀졌으며 지금까지도 다양한 생물학 도메인(domain)에 대하여 활발하게 연구되고 있다[1]. 때문에 생물학 실험에 기초한 방법에 더해 상호작용 예측에 도움을 줄 수 있는 계산학적 접근법의 고안 및 이를 통한 새로운 상호작용 예측에 대한 필요성이 그동안 대두되어왔다.

1.2 관련 연구

개념공간(Concept Space)은 서로 다른 도메인에 대한 정보검색 수행에의 어휘문제 해결을 위해서 90년대 후반 Chen 등에 의해 처음 제안되었다[2]. 이 후 생물의학 문서에서 추출된 정보를 표현하기 위하여 연관개념공간(ACS; Associative Concept Space)이 van der Eijk등에 의해 처음으로 제안되었다[3]. ACS는 시소러스 개념들이 서로간의 연관성에 따라 거리를 두고 분포해 있는 Euclidean 공간으로 두 개념이 함께 나타나는 것을 '연관성'의 기본 정보로 사용한다.

단백질 상호작용 예측에 대한 연구는 주요 생물체들의 유전자 지도의 완성 이후부터 폭넓게 연구되어왔다. 효모(yeast)의 경우 Y2H(yeast two-hybrid)라는 고효율 방법을 이용하여 Ito 등이 다수의 상호작용 정보를 밝혀내기도 하였다[4]. 또한 Bu 등은 기존의 효모 단백질들 간의 상호작용 네트워크 구조 분석을 통해서 상호작용을 예측하였다[5]. Eom 등은 상호작용하는 단백질이 가지는 속성들 간의 연관규칙 발굴을 통해 상호작용을 예측하는 방법을 제시하였으며[6], 속성들 간의 연관성을 신경망을 이용하여 학습한 후 이를 기초로 예측하는 방법도 제시되었다[7].

2. 연관속성개념공간을 이용한 상호작용 연관속성 학습

연관속성개념공간(AFCS; Associative Feature Concept Space)은 ACS를 단백질 상호작용 예측을 위하여 단백질 속성들의 공기정보(co-occurrence information) 활용을 위해 확장한 개념이다. ACS에서의 단어와 단어들 사이의 공기정보는 AFCS에서 단백질의 속성과 상호작용에 따른 두 속성의 공동 출현 정보로 표현된다. 즉, 속성들의 연관성 정보를 AFCS에서는 단백질들이

가지는 각각의 속성들을 시소러스 개념으로 표현하고, 단백질 상호작용에서 함께 나타나는 단백질들의 속성들을 시소러스들의 공동사용 정보와 같이 취급하여 단백질들 간의 연관성을 나타내는 정보를 표현한다.

2.1 단백질 상호작용의 속성정보 인코딩

본 논문에서는 단백질-단백질 상호작용이 두 단백질의 속성들 간의 상호작용으로 표현된다. 이를 위해 상호작용하는 두 단백질 P_a와 P_b가 있다고 할 때 이 두 상호작용은 두 단백질들이 가지는 속성집합의 공동출현(co-occurrence)으로 고려하였다. 여기에서, 상호작용하는 각 단백질 쌍의 속성은 이전 연구에서 사용한 방법과 동일하게 해당 속성의 존재 여부를 이진 코드로 인코딩 하였다[6,7].

2.2 연관속성 개념 공간

단백질이 가지는 속성을 c_i ($i=1,...,N$, N:전체 속성의 수)라 하고, 이러한 속성들로 이루어진 두 단백질들 간의 상호작용에 따른 두 단백질의 속성들의 공동출현 집합인 상호작용 속성집합(interaction feature set)을 f라 할 때, k번째 단백질 상호작용을 표현하는 f_k는 k째 단백질 상호작용에 관여하는 전체 m개의 속성 c_i들의 벡터로 다음과 같이 정의된다.

$$f_k = \{c_1,...,c_m\}$$

이때, 상호작용 k에 대한 상호작용 속성집합 f_k 전체 범위에 대해서 속성 c_i의 존재(occurrence) 여부는 $o_k(c_i)$로 표현되며 다음과 같이 정의된다.

$$o_k(c_i) \Leftrightarrow c_i \in f_k$$

또한, 상호작용 속성집합 f_k에서 두 속성 c_i와 c_j의 공기정보 κ_k는 다음과 같이 정의된다.

$$\kappa_k(c_i,c_j) \Leftrightarrow c_i \in f_k \wedge c_j \in f_k$$

마찬가지 방식으로, 집합 L에 대한 두 속성 c_i와 c_j의 공기정보는 다음과 같이 정의된다.

$$\kappa_L(c_i,c_j) \Leftrightarrow \exists k : \kappa_k(c_i,c_j) \wedge f_k \in L$$

ACFS는 각각의 단백질 상호작용이 속성들 간의 공기정보로 표현된 후 사상(mapping)된 공간을 의미하다. 각각의 단백질 상호작용은 $\mathbf{x}$와 같은 속성 벡터로 표현할 수 있다.

$$\mathbf{x}_i = (x_{i,1} \;\; x_{i,2} \;\; \cdots \;\; x_{i,n})$$

위의 $\mathbf{x}_i$는 총 n개 속성을 가지는 i번째 단백질 상호작용을 상호작용에 관여하는 속성들을 이용한 벡터표현을 나타낸다.

2.3 연관속성 학습

상호작용 관찰에 따른 속성들 간의 연관성 학습은 헵의 학습 규칙(Hebbian learning rule)을 사용한다. 속성들의 벡터로 구성된 상호작용을 관찰함에 따라 각각의 속성개념들을 해당 속성들의 평균점으로 아래이 식을 이용하여 이동시킨다.

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \eta(t)\frac{\mathbf{p}_k(t) - \mathbf{x}_i(t)}{\|\mathbf{p}_k(t) - \mathbf{x}_i(t)\|}$$

여기서 t는 학습 사이클을 의미하며 $\eta(t)$ ($\eta(t) < 0$)는 학습률을 나타낸다. 이 학습률은 속성개념과 $\mathbf{p}_k$의 거리를 줄이는 정도를 나타내며 Van den Berg와 Schuemie에 의해 사용자 설정 파라미터 u를 포함하여 다음과 같이 정의된다[8].

$$\eta(t) = \frac{2}{\min(t,u)}$$

위 식에 따라서 학습률 η가 클 경우 속성들의 공기정보에 따른 공간상의 거리 이동이 그만큼 커지게 된다.

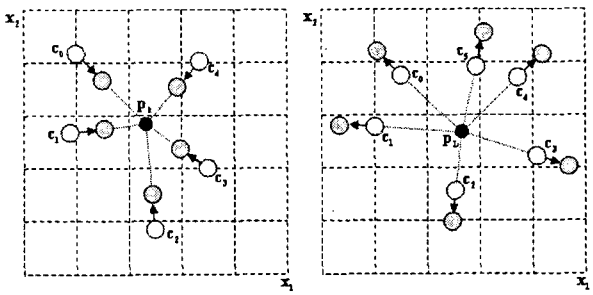

그림 1. 2차원 공간에서의 학습(좌)과 역학습(망각규칙)에 따른 속성 개념의 이동(우)

그림 1에서 단백질 상호작용 k에 대한 상호작용 속성집합 f_k는 $c_0,...,c_4$의 총 5개의 속성을 가진다. 속성의 전체집합 L은 $c_0,...,c_5$의 총 6개의 속성 집합을 가진다. 하나의 속성은 서로 다른 단백질을 표현하는 속성집합들에서 관찰될 수 있다. 학습 규칙은 이러한 상호작용 단백질 집합들을 지속적으로 학습함으로써 AFCS에 표현된 속성들의 위치를 적절히 이동시킨다. 때문에 다수의 상호작용에서 공통적으로 출현하는 속성 정보들은 AFCS상에서 다른 속성들 보다 가까운 거리에 위치하게 된다. AFCS에서는 학습률에 따른 속성들의 위치 이동은 AFCS상에서 속성들이 지나치게 멀어지는 문제를 방지하도록 정규화 하였다.

다음으로, 속성집합으로 표현된 단백질 상호작용들을 이용하여 전체 학습이 완료된 후에는 망각규칙(forgetting rule)을 적용하여 특정 속성이 하나의 점으로 모여지는 문제를 피하고 공기정보가 없는 속성들을 위상적으로 분리할 수 있도록 하였다. 망각규칙의 적용에는 전체 속성집합 L의 모든 속성에 대하여 전체 속성에 대한 중점 $\mathbf{p}_L$에서 정 반대 방향으로 정해진 크기만큼 속성들의 위치를 이동시켰다. 망각규칙의 적용에 따른 속성들의 거리 이동은 다음의 식을 이용하여 조정되었다.

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) - \lambda(\|\mathbf{p}_L(t) - \mathbf{x}_i(t)\|)\frac{\mathbf{p}_L(t) - \mathbf{x}_i(t)}{\|\mathbf{p}_L(t) - \mathbf{x}_i(t)\|}$$

식에서 $\lambda(y)$는 반발함수를 의미하며 본 논문에서는 그 값은 Van den Berg와 Schuemie에 의해 정해진 값을 그대로 사용하였다[8]. 여기서 λ는 다음과 같이 정의된다.

$$\lambda(y) = \begin{cases} 1 & \text{for} \quad y < 1 \\ 1/y & \text{for} \quad y \geq 1 \end{cases}$$

그림 1의 좌측 그림은 각 속성들이 상호작용 속성집합에 따라 학습규칙에 의해 연관성 있는 속성들이 가까워지는 방향으로 이동되는 모습을 나타낸다. 망각규칙은 학습 규칙과는 다르게 그림 1의 우측 그림과 같이 전체 속성집합 L내의 모든 속

성들을 고려하여 하나의 점으로 속성들이 뭉쳐지지 않도록 적용된다.

2.4 연관속성을 이용한 상호작용 예측

논문에서는 단백질 상호작용의 예측은 기본적으로 이전의 연구[6,7]와 같이 일반적인 속성들 보다 상대적으로 밀접한 연관성을 갖는 상호작용 단백질들 사이의 일부 연관속성들을 이용하여 해당 속성들을 갖는 다른 단백질들 간의 상호작용을 예측하는 방법으로 예측을 수행하였다.

3. 실험 결과 및 분석

3.1 데이터

단백질 상호작용 데이터는 MIPS, SGD, YPD의 단백질 상호작용 데이터베이스를 활용하여 총 15,075개의 상호작용을 사용하였다. 상호작용 단백질들은 이전 연구에서 사용한 방법을 이용하여 총 5,812개의 속성으로 인코딩 되었으며 이중에서 정보성이 있는 것으로 추정되는 1,014개의 속성만을 사용하였다. 아래의 표 1은 논문에서 실험에 사용한 상호작용 데이터의 통계정보를 나타낸다.

표 1. 실험에 사용한 상호작용 데이터 집합 및 상호작용 개수

데이터베이스	상호작용 개수	속성 개수	선택된 속성 개수
MIPS	10,641	5,812 (total)	1,014 (total)
YPD	2,952		
SGD	1,482		
전체 개 수	15,075	5,812	1,014

3.2 실험 결과

실험은 전체 상호작용 데이터를 이용하여 10단 교차검증(10-fold cross validation)을 이용하여 수행하였다. 전체 상호작용에 대한 상호작용 속성집합 f를 이용하여 AFCS내에서의 속성들 간의 위상적 연관성을 학습한 후 전체 속성에 대한 중점 $\mathbf{p_L}$에서 가장 멀리 분포한 속성에 대한 거리($d_{\max}$)와 전체 속성에 대한 중점 $\mathbf{p_L}$에 가장 가까운 속성($d_{\min}$)의 거리 차이를 이용하여 연관성 있는 속성으로 간주할 상호작용 속성들의 거리 기준($d_{association}$)을 계산하였다. $d_{association}$은 아래의 식과 같이 $d_{\max}$와 $d_{\min}$의 차이에 거리분할상수 δ를 고려하여 아래와 같이 계산하였다.

$$d_{association} = \delta(d_{\max} - d_{\min})$$

표 2. 거리분할상수(δ)에 따른 연관속성의 수와 이를 이용한 상호작용 예측 성능

δ	속성 수 ($d \le d_{association}$)	예측 정확도(%)
0.05	82	82.1
0.10	173	91.8
0.20	396	88.6
0.40	740	80.3
0.80	853	78.5

표 2는 전체 거리에 대한 분할상수(δ)에 따른 해당 거리 이내의 평균 속성의 개수와 상호작용 예측 성능변화를 나타낸다. 속성의 개수는 AFCS 상에서 해당 거리분할상수 δ를 고려하여 앞의 식으로 계산한 거리 이내에 위치하는 속성들의 수를 나타낸다. 예측 정확도는 이렇게 선별된 속성들을 공통으로 가지는 단백질들을 상호작용 파트너로 고려하여 예측한 상호작용 예측 성능을 나타낸다.

실험 결과 표 2에서처럼 단백질들의 속성으로 고려한 전체 속성들의 분포 중에서 속성들의 중간점을 기준으로 최근점 속성과 최원점 속성 거리 차이의 10%를 연관속성들의 거리로 고려하여 상호작용을 예측했을 경우의 성능이 가장 좋았다.

4. 결론 및 향후 과제

논문에서는 연관속성들을 이용한 상호작용 예측방법을 제안하고 연관속성을 집합을 선정하기 위하여 연관 속성들의 위상정보를 이용하기 위해 AFCS를 사용하였다. 실험결과 비교적 만족할 만한 예측 성능을 관측할 수 있었다. 그렇지만 AFCS는 전체 속성들의 특징을 제대로 반영하기 위한 사상 공간으로는 아직 부적합 하며 단순한 ACS의 확장 보다는 속성들의 특징을 고려하고 보다 정교한 거리 계량법의 도입이 필요하다.

감사의 글

본 연구는 과학기술부 국가지정연구실(NRL) 사업에 의하여 일부 지원되었음을 밝힙니다.

참고문헌

[1] Deng, M. *et al.*, "Inferring domain-domain interactions from protein-protein interactions," *Genome Res.*, Vol. 12, No. 10, pp. 1540–1548, 2002.

[2] Chen, H. et al. , "A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system," *J. Am. Soc. Inf. Sci. Tec.*,, Vol. 48, No. 1, pp. 17-31, 1997.

[3] Van der Eijk CC et al., "Constructing an associative concept space for literature-based discovery," *J. Am. Soc. Inf. Sci. Tec.*, Vol. 55, pp. 436-44, 2004.

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

[5] Bu, D. et al., "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucl. Acids. Res.*, Vol .31, No. 9, pp. 2443-2450, 2003.

[6] J.-H. Eom et al., "Prediction of Implicit Protein-Protein Interaction by Optimal Associative Feature Mining," In *Proc. of the 5th Int'l. Conf. on Intelli. Data Eng. and Autom. Learn. (IDEAL'04)*, pp. 85-91, 2004.

[7] J.-H. Eom et al., "Prediction of Yeast Protein-Protein Interactions by Neural Feature Association Rule.," In *Proc. of the 15th Int'l Conf. on Artif. Neural Networks (ICANN'05)*, Vol. 2, pp. 491-496, 2005.

[8] van den Berg, J., & Schuemie, M., "Information retrieval systems using an associative conceptual space," In *Proc. of the 7th European Symp on Artif. Neural Networks (ESANN'99)*, pp. 351-356, 1999. ■