단백질 상호작용 네트워크에서 단백질 기능 예측을 위한 Modified Chi-square 기법

강태호*, 유재수*
*충북대학교 전기전자컴퓨터공학부

e-mail:thkang@chungbuk.ac.kr, yjs@chungbuk.ac.kr

Modified Chi-square Method for Prediction of Unannotated Proteins from Protein Interaction Network

Tae-Ho Kang*, Jae-Soo Yoo*
*School of Electrical & Computer Engineering, Chungbuk National University

요 약

생명체의 생명현상을 주관하는 각종 화학반응들은 단백질이 관여하고 있다. 단백질은 일정한 질서에 따라 서로 조립되기도 하고, 기능적으로 연관돼 네트워크를 이루고 있다. 이 네트워크를 구성하는 단백질-단백질 상호작용은 단백질의 기능과 밀접하게 관련되어 있다. 즉, 상호작용하는 단백질은 같은 기능을 수행할 가능성이 크다. 이러한 사실은 단백질-단백질 상호작용을 통해 기능이 알려지지 않은 미지 단백질의 기능을 예측할 수 있게 한다. 대표적인 연구로는 이웃 노드에 존재하는 기능분포를 이용하는 이웃노드 카운트(Neighborhood Counting)방식과 특정 기능의 나타날 빈도를 계산하여 기능을 예측하는 카이-제곱(Chi-Square)방식 등이 있다. 본 논문에서는 단백질 기능 예측의 정확성을 높이기 위해 이들 두 방식의 장점을 취합한 보완된 카이-제곱 방식을 제안한다. 그리고 다양한 단백질 상호작용 네트워크 데이터를 비교 분석하여 보완된 카이-제곱 방식이 기능 예측의 정확성이 높음을 증명한다.

1. 서론

단백질의 기능을 밝혀내기 위해 현재까지는 생물학적 실험에 의한 방법을 주로 의존하고 있다. 이러한 실험을 위해서는 많은 비용과 시간이 요구되므로 최근에는 불필요한 실험을 막고 막대한 시간과 비용을 절약하고자 정보기술을 활용하는 노력들이 많이 시도되고 있다.

생명체 내에서 일어나는 대부분의 생명현상은 여러 단백질들이 복합적으로 상호작용함으로서 발생된다. 단백질들은 서로 매우 복잡한 상호작용 관계를 형성하는데 이들 전체 단백질들의 상호작용 관계를 연결하면 하나의 거대한 네트워크를 형성한다. 단백질의 상호작용은 기능과 밀접한 관계가 있다. 따라서 이들 상호작용들을 분석하여 단백질들의 기능적 관계를 파악하거나 밀접한 기능적 관계를 이용하여 기능이 알려지지 않은 미지 단백질의 기능을 예측할 수도 있다.

전체 단백질-단백질 상호작용 네트워크에서 기능의 분포를 살펴보면 특정 기능을 가지는 단백질들이 서로 상호작용하는 기능모듈들을 확인할 수 있다[1]. 이러한 단백질 상호작용 네트워크는 매우 방대하기 때문에 이러한 기능 모듈을 찾아내는 것은 쉽지 않다. 또한 이러한 모듈들은 특정 기능만이 모여 있는 형태만 존재하는 것이 아니고 다양한 기능들이 매우 복잡하게 연결되어 있다.

상호작용과 기능 사이의 관계를 분석한 대표적인 연구는 이웃노드 카운트(Neighborhood counting) 방식과 카이-제곱(Chi-Square)방식이 있다. 하지만 이들 방식들은 서로 다른 특징을 집중적으로 부각시킴으로서 기능 예측의 정확성을 떨어뜨리기도 한다. 따라서 본 논문에서는 이러한 문제점을 제시하고 이를 해결하기 위해 각 방식의 장점을 취합하여 보다 정확성 높은 상호작용 과 기능의 관계 분석 방법을 제시하고자 한다.

2. 단백질 기능 발견을 위한 기존 방법론

단백질 상호작용으로 부터의 기능예측 방법은 크게 네트워크에서 직접적인 연결을 기반으로 단백질의 기능을 예측하는 직접(Direct) 방식과 관계가 있는 단백질들의 모듈을 식별하고 모듈 내 단백질들의 알

★ 이 논문은 2008년 정부(교육과학기술부)의 지원을 받아 수행된 연구임" (지역거점연구단육성사업/충북BIT연구중심대학육성사업단)

려진 기능을 기반으로 각 모듈의 기능을 추측하는 모듈(Module-assisted) 방식으로 나눌 수 있다[2]. 이 중 본 논문에서 제안하는 방법과 관련이 있는 대표적인 기존 연구들에 관해 설명한다.

대표적인 직접 방식으로 이웃노드 카운트방식과 이를 변형한 방식들이 있다. 이웃노드 카운트 방식은 기능이 밝혀지지 않은 단백질과 직접 상호작용하는 단백질들의 이미 알려진 기능을 기반으로 단백질 기능을 예측한다[3]. 이 방식은 단백질 기능 예측을 위한 가장 간단하면서 직접적인 방법이다. 하지만 기능예측을 위해 다른 의미 있는 값들에 대한 연관성이 전혀 부여되지 않으며 전체 네트워크 토폴로지를 고려하지 않는다는 문제점이 있다. 그리고 특정 기능 클래스가 전체 기능 클래스에 차지하는 크기를 무시함으로써 기능 예측에 대한 편차를 가져온다.

이웃노드 카운트 방식과 마찬가지로 네트워크의 토폴로지는 고려하지 않으나. 한 단백질의 k-이웃에 대해 카이제곱을 적용해 기능을 예측하는 카이제곱 방식이 있다[4]. 카이제곱 방식은 네트워크의 토폴로지를 고려하지 않고 단지 특정 단백질의 k-이웃만을 고려하기 때문에 네트워크 크기가 매우 작거나 큰 경우 기능 예측의 정확성이 떨어진다. 카이-제곱 방식은 다음의 수식으로 표현된다.

$$Si(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)}$$

N(i) - 노드 i의 이웃
ni(j) - N(i)에서 기능 j를 갖는 단백질의 수
f(j) - 전체 단백질에서 기능 j의 빈도
ei(j) = |N(i)|f(j) - N(i)에서 기능 j를 갖을 확률

위의 수식에서 볼 때 카이-제곱 *Si(j)*는 *ei(j)*가 작을수록 큰 값을 갖게 된다. 즉 전체 단백질에서 기능 *j*의 빈도인 *f(j)*가 낮을 경우 이웃노드에 존재하는 기능(*j*)의 출현 빈도가 낮더라도 *Si(j)*의 값이 커지게 되어 기능예측의 정확성을 떨어뜨릴 수 있다.

예를 들면 표 1의 경우를 확인할 수 있다.

먼저 YER161C 단백질은 실제로 29, 50번 기능을 갖는다. 하지만 표 1의 예에서 기능번호 196의 경우 이웃노드에 기능 196이 존재하는 경우 *ni(j)*는 2지만 *f(j)*의 값이 매우 작아 결과적으로 ni(j)가 19인 기능 29보다 높은 *Si(J)*값을 갖게 된다. 이러한 이유로 인해 기능 예측의 정확성이 떨어질 수 있다.

<표 1> 카이-제곱 점수

단백질	N(i)	n(i)j	f(j)	기능번호	카이제곱
YER161C	27	19	0.3037	29	14.23
YER161C	27	8	0.11191	128	8.2
YER161C	27	5	0.02957	189	22.11
YER161C	27	2	0.06653	83	0.02
YER161C	27	2	0.05175	109	0.26
YER161C	27	2	0.00082	196	176.39
YER161C	27	1	0.01273	33	1.25
...	...	...	...	...	...

3. 제안하는 보완된 카이-제곱 방법

본 논문에서는 카이-제곱 방법에서 f(j)에 의해 Si(j)가 지나치게 높게 설정되는 것을 방지하기 위해 다음과 같은 보완된 카이-제곱 방식을 제안한다.

$$MSi(j) = \alpha \frac{(n_i(j) - ei(j))^2}{ei(j)} * \beta \frac{n_i(j)}{N(i)}$$

먼저 기존의 카이-제곱에 *ni(j)/N(i)* 값이 추가적으로 반영된다. 이는 이웃노드에 존재하는 기능*(j)*의 빈도를 반영함으로서 *MSi(j)* 결과 값을 보완하는 역할을 수행한다. 즉, 표1에서의 기능 29와 같이 기능빈도가 높은 경우에 *Si(j)*값이 지나치게 낮게 설정되지 않도록 방지한다.

여기에서 α와 β값은 각각의 수식에 대한 가중치를 의미한다. 여기에서의 가중치는 실험을 통해 최적의 가중치를 부여하도록 하였다.

제안하는 방식의 정확성을 검증하기 위해 먼저 MIPS(2003년도), MIPS(2006년도)[5], DIP[6], SGD[7]등의 4개의 단백질 상호작용 네트워크 데이터베이스를 수집하였다. 그리고 단백질의 기능 정보는 GO(Gene ontology)의 기능 분류(266가지 기능)를 사용하였다.

먼저 각각의 데이터베이스를 네트워크로 구축하고 이들의 상호작용 관계를 조사하여 카이-제곱방식과 제안하는 방식의 스코어를 계산하여 기능 1개를 예측하였을 경우와 2개를 예측하였을 경우 그리고 3개를 예측하였을 경우에 대해 예측된 결과가 기존에 알려진 기능과 일치하는 경우의 수를 구하였다. 각각의 데이터베이스에 대해 예측된 결과는 다음의 표 2,3,4,5와 같다.

표에서의 예측 가능한 단백질 수의 의미는 전체 단백질 중 단백질의 기능이 이웃에 상호작용 하는 단백질의 기능과 같은 경우를 의미한다. 예를 들어 표 2에서 1704/3373은 전체 3373개의 단백질 중 1704개의 단백질이 이웃하는

단백질과 기능이 일치함을 의미하며 상호작용을 통해 예측 가능한 단백질 수를 말한다. Chi-square는 카이-제곱 방식으로 예측한 값을, M α10 β90 은 제안된 방식에서 α를 10% 비율로 β를 90% 비율로 계산한 경우를 말한다.

<표 2> MIPS 2003년 데이터베이스

예측가능한 단백질 수 : 1704/3373			
예측 개수	1	2	3
Chi-square	515	962	1252
M α10 β90	529	965	1252
M α20 β80	535	966	1252
M α30 β70	537	967	1252
M α40 β60	538	969	1252

<표 3> MIPS 2006년 데이터베이스

예측가능한 단백질 수 : 1704/4469			
예측 개수	1	2	3
Chi-square	650	1158	1592
M α10 β90	680	1176	1606
M α20 β80	686	1191	1620
M α30 β70	692	1195	1627
M α40 β60	697	1203	1632

<표 4> DIP 데이터베이스

예측가능한 단백질 수 : 1704/4870			
예측 개수	1	2	3
Chi-square	583	1066	1427
M α10 β90	597	1068	1427
M α20 β80	602	1071	1431
M α30 β70	604	1078	1437
M α40 β60	608	1080	1440

<표 5> SGD 데이터베이스

예측가능한 단백질 수 : 1704/5154			
예측 개수	1	2	3
Chi-square	360	638	867
M α10 β90	361	639	867
M α20 β80	367	644	868
M α30 β70	372	648	867
M α40 β60	375	648	867

표 안의 수치는 정확하게 예측된 단백질의 수를 의미한다. 실험 결과 α를 40% 이상으로 했을 경우 정확성이 최대이다 α의 비율이 더 높아질수록 정확성이 다시 떨어지는 결과를 확인하였다. 따라서 실험결과는 α를 40%까지만 제시하였다. 결과적으로 표2에서 표5 까지의 실험 결과 제안하는 방식이 카이제곱 보다 정확하게 예측함을 확인할 수 있다. 그리고 보다 정확한 예측이 가능하도록 하는 α와 β의 비율은 40%, 60% 임을 실험을 통해 도출하였다.

4. 결론

본 논문에서는 단백질 상호작용과 기능사이의 연관성을 분석하여 기능이 알려지지 않은 미지 단백질의 기능을 보다 정확하게 예측할 수 있도록 하는 보완된 카이-제곱 방식을 제안하였다. 그리고 여러 단백질 상호작용 데이터베이스에 대한 실험을 통해 정확성이 높아졌음을 확인하였다. 향후에는 보다 다양한 실험을 통해 보완하여 단백질 상호작용 네트워크의 상호작용 관계 분석 및 기능분석 등에 활용할 예정이다.

참고문헌

[1] Jea Woon Ryu, et al, prediction of Unannotated Proteins from Protein Interaction Network Filtered by Using Localization and Domains in Yeast, JKPS, Vol51, No5, (2007)

[2] He X, Zhang J. Why Do Hubs Tend to Be Essential in Protein Networks? PLos Genet. 2:6, 0826-0834 (2006)

[3] Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B. Characterization of protein hubs by inferring interacting motifs from protein interactions. PLoS Comput Biol. 3(9), 1761-71(2007)

[4] Uetz P, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623-627(2001)

[5] http://mips.gsf.de/

[6] http://dip.doe-mbi.ucla.edu/

[7] http://www.yeastgenome.org/