보완된 카이-제곱 기법을 이용한 단백질 기능 예측 기법

Fucntional Prediction Method for Proteins by using Modified Chi-square Measure

강태호*, 유재수*, 김학용**
충북대학교 전기전자컴퓨터공학부*, 충북대학교 생명과학부**

Tae-Ho Kang(thkang@chungbuk.ac.kr)*, Jae-Soo Yoo(yjs@chungbuk.ac.kr)*,
Hak-Yong Kim(hykim@chungbuk.ac.kr)**

요약

유전체 분석에서 중요한 부분 중 하나는 기능이 알려지지 않은 미지 단백질에 대한 기능 예측이다. 단백질-단백질 상호작용 네트워크를 분석하는 것은 미지 단백질에 대한 기능을 보다 쉽게 예측할 수 있게 한다. 단백질-단백질 상호작용 네트워크로부터 미지 단백질의 기능을 예측하기 위한 다양한 연구들이 시도되어 왔다. 카이-제곱(Chi-square) 방식은 단백질-단백질 상호작용 네트워크를 통해 기능을 예측하고자 하는 연구 중 대표적인 방식이다. 하지만 카이-제곱 방식은 네트워크의 토폴로지를 반영하지 않아 네트워크 크기에 따라 예측의 정확성이 떨어지는 문제점이 있다. 따라서 본 논문에서는 카이-제곱 방식을 보완하여 정확성을 높인 새로운 기능 예측 방법을 제안한다. 이를 위해 MIPS, DIP 그리고 SGD와 같은 공개된 단백질 상호작용 데이터베이스들로부터 데이터를 수집하여 분석하였다. 그리고 제안된 방식의 우수성을 입증하기 위해 각 데이터베이스들에 대해 카이-제곱방식과 제안하는 보완된 카이-제곱(Modified Chi-square)방식으로 예측해보고 이들의 정확성을 평가하였다.

■ 중심어 : | 단백질 기능예측 | 단백질 상호작용 네트워크 |

Abstract

Functional prediction of unannotated proteins is one of the most important tasks in yeast genomics. Analysis of a protein-protein interaction network leads to a better understanding of the functions of unannotated proteins. A number of researches have been performed for the functional prediction of unannotated proteins from a protein-protein interaction network. A chi-square method is one of the existing methods for the functional prediction of unannotated proteins from a protein-protein interaction network. But, the method does not consider the topology of network. In this paper, we propose a novel method that is able to predict specific molecular functions for unannotated proteins from a protein-protein interaction network. To do this, we investigated all protein interaction DBs of yeast in the public sites such as MIPS, DIP, and SGD. For the prediction of unannotated proteins, we employed a modified chi-square measure based on neighborhood counting and we assess the prediction accuracy of protein function from a protein-protein interaction network.

■ keyword : | Protein Function Prediction | Protein-Protein Interaction Network |

* 본 논문은 2009년 교육과학기술부와 한국산업기술재단의 "지역혁신인력양성사업"과 정부(교육과학기술부)의 "지역거점연구단육성사업/충북BIT연구중심대학육성사업단"의 지원을 받아 수행된 연구임.

접수번호 : #081209-004
접수일자 : 2008년 12월 09일
심사완료일 : 2009년 01월 21일
교신저자 : 유재수, e-mail : yjs@chungbuk.ac.kr

I. 서 론

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]. 기능적 모듈은 대부분의 생체 시스템에서 확인할 수 있다. [2] 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.

II. 관련연구

단백질 상호작용으로 부터의 기능예측 방법은 크게 네트워크에서 직접적인 연결을 기반으로 단백질의 기능을 예측하는 직접(Direct) 방식과 관계가 있는 단백질들의 모듈을 식별하고 모듈 내 단백질들의 알려진 기능을 기반으로 각 모듈의 기능을 추측하는 모듈(Module-assisted) 방식으로 나눌 수 있다[3]. 이중 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[4]. 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 카이-제곱을 적용해 기능을 예측하는 카이-제곱방식이 있다[5]. 카이-제곱 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.

$$Si(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)} \quad (1)$$

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 j의 빈도인 f(j)가 낮을 경우 이웃노드에 존재하는 기능(j)의 출현 빈도가 낮더라도 Si(j)의 값이 커지게 되어 기

능예측의 정확성을 떨어뜨릴 수 있다.

예를 들면 [표 1]의 경우를 확인할 수 있다. 먼저 YER161C 단백질은 실제로 29, 50번 기능을 갖는다. 하지만 [표 1]의 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	카이-제곱
YER161C	27	19	0.3037	29	14.23
YER161C	27	8	0.1119	128	8.20
YER161C	27	5	0.0297	189	22.11
YER161C	27	2	0.0663	83	0.02
YER161C	27	2	0.0515	109	0.26
YER161C	27	2	0.0002	196	176.39
YER161C	27	1	0.0127	33	1.25
...	...	...	...	...	...

III. 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.

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

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(2006년도)[6], DIP[7], SGD[8]등의 3개의 단백질 상호작용 네트워크 데이터베이스를 사용하였다. 그리고 수집된 상호작용 정보는 부정확한 데이터를 제거하기 위하여 정제하였다. 단백질-단백질 상호작용 정보는 기능이 알려지지 않은 미지 단백질의 기능을 예측하는데 효과적으로 이용될 수 있다. 하지만 이들 단백질 상호작용 데이터는 추출한 실험 방식에 따른 오류를 포함하고 있어 상호작용이 생물학적으로 미치는 영향을 평가하는데 오히려 오류를 범할 수 있게 된다[9]. 따라서 상호작용 데이터의 정확성을 높이기 위해 단백질의 세포내 위치 정보(localization)를 기반으로 상호작용을 정제하여 사용한다[10]. 사용된 데이터는 [표 2]와 같다.

표 2. 단백질 상호작용 네트워크

데이터베이스	단백질 수	상호작용 수
MIPS	4,469	11,912
DIP	4,870	16,604
SGD	5,154	69,109

그리고 단백질의 기능은 GO(Gene ontology)의 분자기능(Molecular Function)에 따라 기능 분류(266가지 기능)를 수행하였다.

먼저 각각의 데이터베이스를 네트워크로 구축하고 이들의 상호작용 관계를 조사하여 카이-제곱방식과 제안하는 방식의 스코어를 계산하여 기능을 예측하였다. 각 단백질 네트워크를 조사해본 결과 1차 이웃의 기능을 따를 가능성이 약 50%정도이고, 2차 이웃의 기능을 따를 가능성은 약 20%정도로 낮았다. 1차 이웃과 2차 이웃에서 공유하는 단백질 기능의 분포를 조사한 결과를 [표 3]에서 보이고 있다.

표 3. 1, 2차 이웃의 기능과 동일한 경우

데이터베이스	1차 이웃과 기능공유 (단백질 수)	2차 이웃과 기능공유 (단백질 수)
MIPS	2,121	951
DIP	2,613	975
SGD	1,734	642

표 3을 근거로 본 논문에서는 1차-이웃 단백질들 중에서 스코어가 가장 높은 기능을 예측하고자 하는 단백질의 기능으로 예측하여 성능평가를 수행하였다.

예측 결과를 측정하기 위해 먼저 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 4 - 표 6]과 같다.

표 4. MIPS 데이터베이스

예측가능한 단백질 수 : 2121/2121			
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. DIP 데이터베이스

예측가능한 단백질 수 : 2613/4870			
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 6. SGD 데이터베이스

예측가능한 단백질 수 : 1734/5154			
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

각 표 에서 예측 가능한 단백질 수의 의미는 전체 단백질 중 단백질의 기능이 1차 이웃에 상호작용 하는 단백질의 기능과 같은 경우를 의미한다. 예를 들어 [표 2]에서 2121/4870은 전체 4870개의 단백질 중 2121개의 단백질이 1차 이웃 단백질과 기능이 일치함을 의미하며 상호작용을 통해 예측 가능한 단백질 수를 말한다.

Chi-square는 기존의 카이-제곱 방식으로 예측한 값을 나타내며, M α10 β90 은 제안된 방식에서 α를 10% 비율로 β를 90% 비율로 계산한 경우를 말한다. 그리고 표 안의 수치는 정확하게 예측된 단백질의 수를 의미한다. 실험 결과 α를 40% 이상으로 했을 경우 정확성이 최고를 기록하다, α의 비율이 더 높아질수록 정확성이 다시 떨어지는 결과를 확인하였다. α와 β의 비율에 따른 정확성의 차이를 [그림 1]에서 보이고 있다.

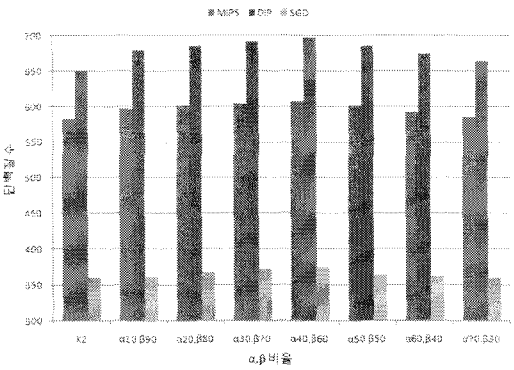

그림 1. α, β 비율에 따른 정확성 비교

실험을 통해 α와 β의 비율이 40%, 60% 일 때 보다 정확한 예측이 가능함을 알 수 있었다.

결과적으로 성능평가를 통해 본 논문에서 제안하는 보완된 카이-제곱 방식이 기존의 카이-제곱에서의 문제점을 보완하였고 정확도가 높아졌음을 확인하였다.

IV. 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. MIPS, DIP 그리고 SGD등의 다양한 크기의 공개된 단백질 상호작용 데이터베이스를 활용해 성능평가를 수행하여 제안된 방식의 정확성이 보다 높음을 입증하였다. 향후에는 보다 다양한 실험을 통해 보완하여 단백질 상호작용 네트워크의 상

호작용 관계 분석 및 기능분석 등에 활용할 예정이다.

참 고 문 헌

[1] J. W. Ryu, "Prediction of Unannotated Proteins from Protein Interaction Network Filtered by Using Localization and Domains in Yeast," JKPS, Vol.51, No.5, 2007.

[2] P. Holme, M. Huss, and H. Jeong, "Subnetwork Hierarchies of Biochemical Pathways," Bioinformatics, Vol.19, No.4, 2003.

[3] X. He and J. Zhang, "Why Do Hubs Tend to Be Essential in Protein Networks?", PLoS Genetics, Vol.2, No.6, 2006.

[4] R. Aragues, A. Sali, J. Bonet, M. A. Marti-Renom, and B. Oliva, "Characterization of Protein Hubs by Inferring Interacting Motifs from Protein Interactions," PLoS Comput. Biol., Vol.3 No.9, 2007.

[5] P. Uetz, "A Comprehensive Analysis of Protein-protein Interactions in Saccharomyces Cerevisiae". Nature Vol. 403, pp. 623-627, 2001.

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

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

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

[9] T. R. Hazbun and S. fields. "Networking Proteins in Yeast", Proc. Natl. Acad. Sci. USA, 2001.

[10] H. Y. Kim, H. Y. Kang, J. W. Ryu, C. N. Yoon, S. K. Han, "Phase Specific Activated Modules from a Protein Interaction Network of Yeast Cell Cycle," JKPS, Vol.50, No.91, 2007.

저 자 소 개

강 태 호(Tae-Ho Kang) 정회원

- 1999년 2월 : 호원대학교 정보통신공학과(공학사)
- 2002년 8월 : 충북대학교 정보산업공학과(공학석사)
- 2007년 8월 : 충북대학교 정보통신공학과(공학박사)
- 2007년 9월 ~ 현재 : 충북대학교 전기전자컴퓨터공학부 Post-doc.

<관심분야> : 데이터베이스 시스템, 데이터 마이닝, 생물정보학, 시스템 바이오

유 재 수(Jae-Soo Yoo) 종신회원

- 1989년 2월 : 전북대학교 컴퓨터공학과(공학사)
- 1991년 2월 : 한국과학기술연구원 전산학과(공학석사)
- 2007년 8월 : 한국과학기술연구원 전산학과(공학박사)
- 1996년 ~ 현재 : 충북대학교 전기전자컴퓨터공학부 교수.

<관심분야> : 데이터베이스 시스템, 정보검색, 멀티미디어 데이터베이스, 분산객체 컴퓨팅, 생물정보학

김 학 용(Hak-Yong Kim) 종신회원

- 1985년 2월 : 충북대학교 농화학과(공학사)
- 2002년 8월 : 충북대학교 화학과(공학석사)
- 2007년 8월 : Connecticut University, Molecular Cell Biology. (공학박사)
- 1998년 ~ 현재 : 충북대학교 생명과학부 교수

<관심분야> : 시스템 바이오, 신호 전이, 단백질 네트워크, 생체동역학