단백질 상호작용 데이터의 신뢰도 검증 기법

홍진선° 한경숙
인하대학교 컴퓨터정보공학과
jsblue01°@freechal.com, khan@inha.ac.kr

A scoring method for evaluating the reliability of protein-protein interaction data

Jinsun Hong° Kyungsook Han
Department of Computer Science and Information Engineering, Inha University

요 약

단백질 상호작용 검출 방법의 발달로 많은 양의 데이터가 산출되고 있고, 이러한 상호작용 데이터의 방대한 양으로 인해 통계적 방법을 이용하여 데이터를 처리함으로서 유용한 지식을 얻을 수 있다. 예측한 상호작용 데이터는 첫째, 대량의 데이터를 생산해내므로, 많은 false-positive를 내포하고 있고, 둘째, 예측한 상호작용을 검증시 실험을 하는 방법 외에는 신뢰도를 측정하기가 어렵다는 문제점이 있다. 본 연구에서는 점수 할당 시스템을 사용함으로서 예측한 인간 단백질 상호작용 데이터의 false-positive를 줄이고,각각 상호작용에 점수를 부여함으로서 상호작용 데이터의 신뢰도를 검증하는 방법을 제안하고 있다.

1. 서 론

최근 yeast two-hybrid나 mass spectrometry techniques 같은 high-throughput을 통한 상호작용 검출방법의 발달은 단백질 상호작용 데이터의 빠른 증가를 가져왔다. 생물정보학은 이러한 생물 데이터의 방대한 양을 통계적인 방법이나 데이터 마이닝 기법을 사용하여 유용한 지식을 얻는 생명 과학 연구이다.

단백질 상호작용 데이터의 증가는 새로운 단백질 사이에서 일어나는 상호작용을 발견하는데 기초가 되고 있다. 단백질 상호작용 예측에는 도메인을 사용하거나 [1] 염기서열의 유사도를 이용하는 방법이 사용되고 있다. 도메인을 사용하는 이유는 원래 존재하는 단백질의 기능 부위인 도메인만을 재조합하여 새로운 단백질을 만들 수 있으므로, 더 복잡한 단백질 생산을 위해 새로운 물질을 찾는 것보다 다른 생물에서 이미 검증되어 있는 단백질을 재배열 하는 것이 효과적이기 때문이다. 도메인은 단백질 패밀리들에 대하여 Hidden Markov 모델을 적용한 Pfam이나 Rosetta Stone 방법을 사용하는 InterDom, 알려진 단백질들로부터 단백질 패밀리와 도메인을 추출해 낸 InterPro 데이터베이스로부터 얻을 수 있다. 염기 서열의 유사도를 측정하기 위해서 BLAST [2]나 FASTA같은 툴을 사용하며, 이 외에도 단백질의 구조 상호작용 맵 [3]도 사용된다. 그러나, 예측한 단백질 상호작용 데이터는 첫째, 대량의 데이터를 생산해내므로, 많은 false-positive를 내포하고 있고, 둘째, 예측한 상호작용을 검증는 문제점이 있다. 따라서, 이 논문에서는 단백질의 세포내 소기관 정보를 사용하여 점수 할당 시스템을 개발함으로서, 예측한 상호작용의 false-positive를 줄이고, 데이터의 신뢰도를 검증하는 방법을 제안하고 있다.

본 논문은 다음과 같이 구성된다. 2장에서는 점수 할당 시스템을 구축하는 방법에 관해 소개하고, 3장에서는 점수 할당 시스템이 구축되어 예측한 데이터에 적용시킨 결과를 설명한다. 그리고 4장에서는 결론 및 향후 과제에 대하여 기술한다.

2. 점수 할당 시스템 구축

HPRD (http://www.hprd.org) 의 인간 단백질 상호작용 데이터의 세포내 위치 정보에 근거하여 점수 할당 시스템을 개발하였다. HPRD는 인간의 건강과 질병에 관하여 인간 단백질의 적절한 기능과 세포내 소기관 정보를 통합한 관계형 데이터베이스이다 [4]. 세포 소기관 정보를 사용하여 잉여의 단백질 상호작용을 제거하는 실험을 하였고, 이 실험으로부터 동일한 세포 소기관내에 있는 단백질 사이의 상호작용은 다른 소기관 내에 있는 단백질 사이의 상호작용보다 많이 일어나고, 특히 세포질과 핵에서는 다른 소기관들에 비해 많은 상호작용이 일어난다 [5] 는 규칙을 발견하였다. 또한 단백질 상호작용의 초기 예측 데이터는 동일한 세포소기관 내에 있는 단백질의 쌍을 선택함으로서 false-positive인 상호작용이 제거된다는 규칙을 찾아내었고, 이러한 규칙들을 기본으로 하여 점수 할당 시스템을 개발하게 되었다.

점수 할당 시스템은 그림 1. 에서 상호작용 데이터를 기존의 데이터베이스에서 추출하고, 이를 이용해서 점수 행

시 실험을 하는 방법 외에는 신뢰도를 측정하기가 어렵
렬을 만들고 [6], 인간 단백질의 상호작용을 예측한 후에, 예측한 상호작용에 대하여 점수를 부여하는 4 단계로 구성되어 있다.

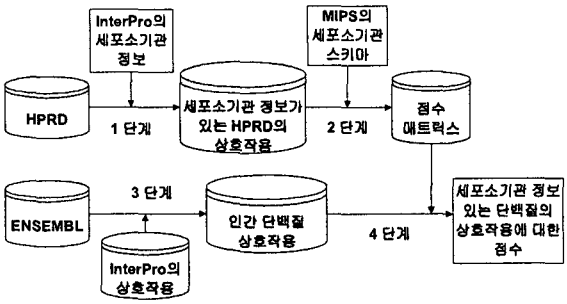

그림 1. 점수 할당 시스템 구성도.

처음 단계인 1 단계에서 세포내 소기관 정보는 HPRD와 InterPro (http://www.ebi.ac.uk/interpro/)로부터 가져와서 중복된 데이터를 제거하여 로컬 데이터베이스에 저장한다. HPRD에는 5818개의 소기관 정보가 있는 단백질이 5879개의 상호작용을 하고, InterPro는 1086개의 소기관 정보가 있는 도메인 사이에서 1712개의 상호작용을 한다.

다음으로 2 단계에는 소기관의 분류 데이터를 얻기 위하여 MIPS ftp://ftp.pasteur.fr/pub/databases/yeast/catalogues/subcell/subcellcat.scheme) 의 스키마를 가져와서 HPRD와 InterPro의 형용사형으로 표시된 소기관을 20개의 명사형으로 구분하고, 각각에 대해서 아이디를 부여하였다. 세포내 소기관에 대한 점수 행렬은 20*20 으로 구성되었다. 본 시스템의 1단계에서 축적된 HPRD의 상호작용 데이터에 대하여 식 1과 같이 계산한 상호작용의 확률 값으로 점수 행렬의 행과 열이 채워진다. 점수(S)는 S_I 과 S_A로 구성되며, S가 높은 값을 갖을 수록 예측한 단백질의 상호작용은 신뢰할 수 있다는 것을 보여준다.

$$S_I = \frac{\sum_{j} I(i,j)}{\sum_{i=1}^{j} I_i} \qquad (1)$$

$I(i,j)$ 는 단백질 i 와 j사이에 상호작용을 나타내고, $\sum_{j}^{i} I_i$ 는 상호작용에 참여하는 단백질의 수이다. S_I 은 세포소기관 정보가 있는 상호작용에 대해서 상호작용에 참여하는 단백질의 확률임을 나타낸다.

3 단계는 인간 단백질의 상호작용을 예측하는 단계로서, InterPro의 1712개의 도메인 상호작용 데이터를 가져와서 이들을 Ensembl (http://www.ensembl.org/) 에 있는 인간의 단백질에 도메인을 할당하고, 도메인 조합에 의해 상호작용을 예측한 후에, 예측한 데이터를 로컬 데이터베이스에 저장한다.

마지막 단계인4 단계에서는 HPRD로부터 점수 행렬의 값이 세포소기관 내에서 예측한 단백질 상호작용의 점수를 계산하기 위해 사용되었다. Ensembl로부터 가져온 10598개의 인간 단백질에 S_I 를 할당한다. 두개 이상의 소기관을 갖는 인간 단백질의 상호작용인 경우에는 식 2와 같이 평균을 낸 점수 (S_A) 를 상호작용에 부여한다.

$$S_A = \frac{\sum_{I=1}^{m} S_I}{\sum_{i=1}^{m} N_i} \qquad (2)$$

$\sum_{I=1}^{m} S_I$ 는 단백질 상호작용시 S_I 점수의 합을 나타내고, $\sum_{i=1}^{m} N_i$ 는 S_I이 누적된 개수이다.

3. 점수 할당 시스템을 적용한 결과

표 1. 은 HPRD에서 가져온 데이터로부터 계산한 초기 점수 행렬의 값의 분포와 점수 행렬값을 예측한 상호작용에 적용했을 때의 값의 분포를 나타낸다.

표 1. HPRD와 InterPro의 세포내 소기관에 할당된 점수에 대한 상호작용의 분포

S	HPRD의 누적백분율	S	예측한 상호작용의 누적백분율
0.5	1%	0	40%
0.6	2%	0.5	41%
0.7	6%	0.6	59%
0.8	7%	0.7	63%
0.9	9%	0.9	64%
1.0	10%	1.1	70%
1.1	23%	1.3	81%
1.3	39%	1.8	99%
1.6	49%	1.9	100%
1.8	76%		
1.9	100%		

본 연구는 MS SQL SERVER를 이용하여, 윈도운 2000에서 C#으로 점수 할당 시스템을 구현하였다. 본 시스템에서는 InterPro로부터 가져온 데이터의 도메인 상호작용쌍을 사용하여 Ensembl에 있는 인간 단백질의 상호작용을 예측했다. 예측한 단백질 상호작용에 점수 할당 시스템을 적용한 결과, 0에서부터 1.9까지의 값을 갖는다. 이때, 0의 값을 갖는 상호작용은 세포 소기관 정보가 없는

것으로 신뢰도가 낮은 상호작용임을 나타낸다. 표 2. 에서 처럼 Ensembl의 인간 단백질 상호작용이 점수 할당 시스템으로 들어오면 각 상호작용에 대하여 고유한 edgeID를 갖게 된다. 상호작용하는 단백질 A와 단백질 B를 타겟과 소스라 하자. 타겟과 소스 단백질에 대해 점수 행렬에서 표시한 세포 소기관 아이디가 할당되면, 본 시스템에서 제안하여 구한 점수 행렬 함수를 참조하여 S_I과 S_A를 고려한 후에 높은 점수 (S) 순서에 따라 인간의 단백질 상호작용이 나열된다.

표 2. 인간 단백질 상호작용의 EdgeID에따른 타겟과 소스 단백질의 소기관 ID에 해당하는 점수(S)

EdgeID	타겟의 소기관 ID	소스의 소기관 ID	점수(S)
2	12	12	1.8
402	9	9	1.3
762	3	12	1.1
593	12	5	0.6
38	1	1	0.5
409	14	9	0.4
1511	3	7	0.4
...	...	...	...

예측한 전체 1,024,263개의 상호작용 중에서 0 보다 높은 점수를 갖고 있는 데이터는 0 인 값을 갖는 데이터보다 신뢰할 만한 상호작용이라 할수 있고, 이들은 전체 상호작용 중에서 26% (265,242개) 를 차지하고 있다. 0 보다 높은 점수를 갖는 상호작용은 타겟과 소스가 되는 단백질 모두 세포 소기관 정보가 있는 경우에 해당한다. 나머지 74%는 점수 할당 시스템에 적용한 결과 점수가 0 으로서 이들 상호작용 데이터는 상당한 false-positive를 포함하고 있음을 알수 있다. 기존에 도메인 만으로 예측한 상호 작용 데이터의 경우, 상당히 많은 양의 상호작용 데이터를 얻었을 뿐 이들 중에서 어떠한 상호작용이 믿을 만한 것인지 알수 없었다. 그러나 상호작용 데이터의 신뢰도 검증 기법인 점수 할당 시스템을 예측한 상호 작용 데이터에 적용 시키면 0 이상의 점수를 갖는 상호작용 데이터는 신뢰할 수 있음을 제시하므로, 단백질 상호작용에 관한 연구를 하는데 있어서 상당수의 false-positive를 줄일 수 있다. 본 시스템에 의하여 예측한 인간 단백질 상호작용이 높은 점수 순서대로 나열되므로, 사용자가 원한다면 예측한 상호작용 데이터 중에서 신뢰도가 검증된 데이터만을 선정하여 상호작용을 살펴보는 것도 가능하다.

4. 결론

점수 할당 시스템의 목적은 예측한 인간 단백질의 상호작용 중 신뢰할 만한 상호작용을 선택하는 것이다. 상호작용 데이터의 신뢰도를 검증하는 점수 할당 시스템은 세포내 소기관 정보를 갖고 있는 단백질이 상호 작용을 한다면 그렇지 않은 단백질의 상호작용보다 높은 점수를 갖게 되는 원리로 작동한다. 점수 할당 시스템은 통계적인 방법으로부터 구해졌고, 예측한 인간 단백질 상호작용의 데이터에서 0 에서 1.9 사이의 값을 갖는다.

본 연구는 통계학적인 방법에 의해 예측된 상호작용 데이터 중에서 신뢰할 만한 인간 단백질 상호작용을 제시한다. 세포내 소기관 정보를 이용한 점수 할당 시스템을 사용함으로서 InterPro로부터 가져온 도메인을 사용한 인간 상호작용 예측시, 상당한 양의 가짜 상호작용을 제거할 수 있다. 도메인만을 사용한 단백질 상호작용 예측보다 세포내 소기관 정보를 사용한 점수를 할당함으로서 각각 상호작용에 대한 신뢰도를 평가 할수 있다. 그러나, 예측한 상호작용의 26%정도만이 신뢰할 만한 데이터임을 입증하므로, coverage가 떨어진다. 따라서 향후 연구로는 단백질 상호작용의 신뢰도를 높이기 위한 다른 방법들을 시도해보고, 예측한 상호작용의 coverage를 높일수 있는 방안도 모색해보고자 한다.

5. 참고문헌

[1] Deng M, Mehta S, Sun F, Chen T., Inferring domain-domain interactions from protein-protein interactions, Genome Res, 10, 1540-1548, 2002

[2] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucl. Acids. Res.*, 25, 3389-3402,1997

[3] Lappe, M., Park, J., Niggemann, O. and Holm, L., Generating protein interaction maps from incomplete data: application to fold assignment, *Bioinformatics,* 17, S149-S156. 2001

[4] Perì, S. *et al.*, Development of human protein reference database as an initial platform for approaching systems biology in humans, Genome Research, 13, 2363-2371, 2003

[5] Jeong, H., Mason, S.P., Barabasi, A.-L., and Oltvai, Z.N., Lethality and centrality in protein networks, Nature, 411, 41-42, 2001

[6] Won-ki, H., James, V.F., Luke, C.G., Adam, S.C., Russell, W.H., Jonathan, S.W. and Erin, K.O., Global analysis of protein localization in budding yeast, NATURE, 425, 686-691, 2003