인간 및 초파리 단백질을 대상으로 한 도메인 조합 기반 단백질-단백질 상호작용 예측 기법 검증

장우혁[1O] 한동수[1] 김홍숙[2] 이성독[1]
[1]한국정보통신대학교 [2]한국전자통신연구원
[1]{torajim[O], dshan, sdlee}@icu.ac.kr, [2]kimkk@etri.re.kr

Validation of Domain Combination Based Protein-Protein Interaction Prediction Method Using Human and Fly Proteins

Woo-Hyuk Jang[1O] Dong-Soo Han[1] Hong-Soog Kim[2] Sung-Doke Lee[1]
[1]Information and Communications University
[2]Electronics and Telecommunications Research Institute

요 약

도메인 조합 기반의 단백질-단백질 상호작용 예측 기법(DCPPIP)은 효모 단백질에 대하여 뛰어난 정확도를 보여준다. 그러나 다른 종에서의 예측 정확도 및 기법의 유효성은 아직까지 검증되지 않고 있다. 본 논문에서는, 초파리 및 인간 단백질을 이용한 예측 정확도 검증 및 이종간의 상호작용 예측 실험의 결과를 기술한다. 초파리와 인간 단백질의 실험에서는 각각 10,351개와 2,345개의 상호작용 단백질 쌍이 사용되었다. 초파리와 인간의 상호작용 단백질 쌍 중 80%와 20%를 각각 학습집단 및 실험집단으로 사용하였으며, 상호작용이 없는 단백질 쌍의 학습집단은 1배에서 5배까지 변화시키면서 예측 정확도를 관찰하였다. 정확도는 실험집단 중 학습집단과 도메인이 완전히 혹은 부분적으로 겹치는 쌍들에 대하여 계산하였다.이 결과 초파리에서는 약 77%의 민감도와 92%의 특이도가 확인되었고, 인간 단백질에 대하여는 약 96%의 민감도와 95%의 특이도를 보여주었다. 이종간의 상호작용 예측 실험은 효모, 초파리, 효모+초파리에 해당하는 학습집단 각각을 바탕으로 Human, Mouse, *H. pylori, E. coli, C. elegans* 등의 단백질 상호작용 예측을 수행하였다. 실험 결과 학습집단의 도메인이 실험집단의 도메인과 많이 겹칠 수록 높은 정확도를 보여주었으며, 도메인 집단간의 유사도를 나타내기 위해 고안한 *Domain Overlapping Rate(DOR)* 는 상호작용 예측 정확도의 중요한 요소임을 찾아내었다.

1. 서 론

도메인 조합 기반 예측 기법[3,4,5]은 효모 단백질에 대한 우수한 정확도에도 불구하고 고등 생물체에 대한 예측 정확도 검증이 이루어지지 않았다. 그러나 예측 기법의 실제 활용을 위해서는 고등 생물체에 대한 유효성 검증을 통한 기법의 확장이 필요하다. 그러나 현재 인터넷을 통해 공개되는 단백질-단백질 상호작용 정보 및 도메인 정보의 양은 효모[1,2] 및 초파리[6]와 같이 비교적 활발히 연구가 진행되어온 몇몇 종을 제외하고는 충분하지 않으며 인간 및 쥐와 같은 고등 생물체에 대한 정보는 더욱 부족한 실정이다. 문제 해결을 위해서는 먼저 예측 기법이 다른 종에서도 유효한지를 검증해야 하며, 부족한 정보에 상관없이 종에 따라 적절한 학습집단을 구성하는 방안을 마련해야 한다. 본 논문에서는 초파리 및 인간 단백질을 이용한 예측 실험과 이종간의 상호작용 예측 실험을 수행하였다. *Database of Interacting Protein (DIP)*[6] 에서 확보한 초파리의 상호작용 단백질 쌍 20988개 중, *Integrated documentation resource of Protein families, domains and functional sites(InterPro)* 와 *Protein Information Resource (PIR)* 에서 도메인 정보를 찾을 수 있는 10351개의 단백질 쌍이 실험에 사용되었다. 인간 단백질은 *Human Protein Reference Database (HPRD)* [7]에서 19,514개의 상호작용 단백질 쌍 중 InterPro에서 도메인 정보를 찾을 수 있는 2,345개의 쌍이 실험에 사용되었다. 실험은 기존의 효모에서와 같이 사용가능한 단백질 쌍 중 80%와 20%를 각각 학습집단 및 실험집단으로 구분하여 정확도를 분석하였다. 정확도 는 실험집단 중 학습집단과 공통되는 도메인을 가진 단백질 쌍에 대하여만 측정하였다. 실험 결과 예측 정확도는 초파리의 경우 효모와 유사한 정도 (민감도: 77%, 특이도: 92%)를 나타내었고, 인간의 경우에는 매우 높은 정확도(민감도: 96%, 특이도: 95%)를 나타내어 도메인 조합 기법이 특별한 변형없이 다른 종에서도 적용될 수 있음을 보였다. 이종간의 실험은 효모로 구성된 학습집단, 초파리로 구성된 학습집단, 효모 및 초파리의 조합으로 구성된 학습집단 각각을 기반으로 Human, Mouse, *H. pylori, E. coli, C. elegans* 등의 단백질 상호작용을 예측하였다. 실험에서는 학습집단과 실험집단의 도메인이 겹치는 정도가 클 수록 예측 정확도가 높게 관찰되었으며, 집단간의 도메인 유사도를 나타내는 *Domain Overlapping Rate(DOR)* 를 높히는 것이 학습집단 구성의 중요한 요소임을 알 수 있었다.

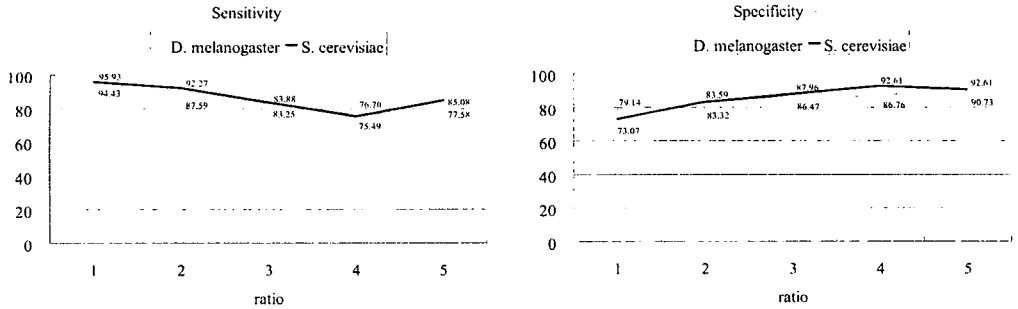

그림 1. 효모와 초파리 단백질의 상호작용 예측 정확도 비교

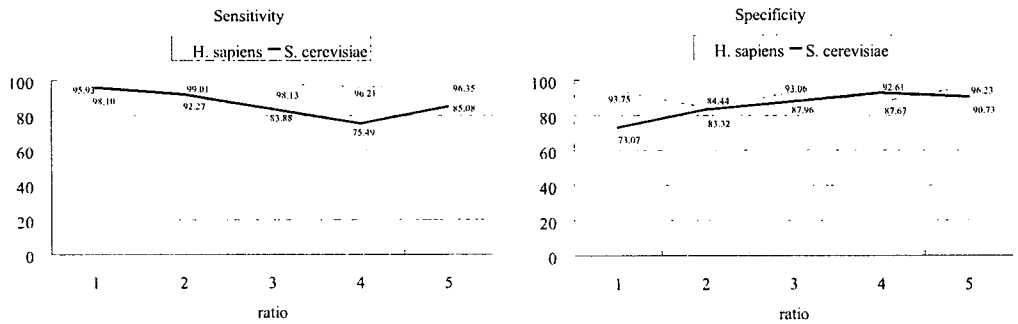

그림 2. 효모와 인간 단백질의 상호작용 예측 정확도 비교

2. 검증 및 결과

2.1 초파리(*D. melanogaster*) 검증

초파리는 그동안 활발히 연구가 진행되어온 종으로, 예측 정확도 실험에 필요한 충분한 크기의 학습집단을 쉽게 구성할 수 있는 장점이 있다. 검증은 효모와 마찬가지로, 8,280개의 상호작용 단백질 쌍(전체 상호작용 쌍의 80%)을 학습집단으로 사용하였고, 2,071개의 쌍(전체 상호작용 쌍의 20%)을 실험집단으로 사용하였다. 또한 초파리에서 발견되는 모든 단백질 중 도메인 정보를 알 수 있는 단백질로 임의의 쌍을 만들어 상호작용이 없는 학습집단 및 실험집단으로 사용하였다. 이 때, 생성된 임의의 쌍 가운데 상호작용 단백질 쌍과의 중복은 제거하였다. 상호작용이 없는 단백질의 실험집단의 크기는 상호작용 단백질과 동일하게 유지하였고, 학습집단의 크기는 상호작용 단백질에 대하여 1배에서 5배로 늘려가면서 예측 정확도를 관찰하였다. 그림 1은 초파리 단백질에서 도메인 조합 기반 예측 기법의 정확도를 효모 단백질의 경우와 비교한 그래프이다. 민감도는 학습집단의 비율에 따라 94.43%~77.58%, 특이도는 79.14%~92.61%를 나타내어, 효모의 경우와 유사한 정도의 정확도를 보여주었다. 이는 곧 초파리에서도 효모의 경우와 마찬가지로 상호작용이 있는 단백질 집단에서의 도메인 패턴은 상호작용이 없는 단백질 집단에서의 도메인 패턴과 비교적 잘 구분이 됨을 뜻하며, 도메인 조합 기반 예측 기법을 별다른 수정 없이 다른 종에 적용할 수 있음을 의미하고 있다.

2.2 인간(*H. sapiens*) 검증

도메인 조합 기반의 단백질 상호작용 예측 기법은 효모와 초파리에 대하여 높은 정확도를 나타냈다. 이에 우리는 인간과 같은 고등 생물체에 대해서도 예측 기법의 유효성을 검증하고자 하였다. 효모 및 초파리에 대한 실험이 예측 기법의 정확도 검증을 위함이라 하면, 인간 단백질에 대한 실험을 통해서는 고등 생물체에 대한 예측 기법의 확장 가능성을 살펴 볼 수 있다.

DIP에서 구할 수 있는 인간 단백질의 갯수는 실험에 필요한 학습집단을 구성하기에 충분하지 않기 때문에 *Human Protein Reference Database (HPRD)* [7]의 단백질을 사용하였다. 19,514개의 상호작용 단백질 쌍 가운데 도메인 정보를 구할 수 있는 2,345개의 쌍이 실험에 사용되었다. 도메인 정보는 효모 및 초파리 실험의 경우와 마찬가지로 InterPro에서 추출하였다. 앞선 실험과 마찬가지로 80%의 상호작용 쌍이 학습집단으로 사용되어 AP 배열 구성에 참여하였고, 나머지 20%는 실험집단으로 사용되었다. 비상호작용 집단은 무작위로 추출된 단백질 쌍 가운데 상호작용 단백질 쌍과의 중복을 제외한 나머지가 사용되었다. 그림 2에서와 같이 매우 높은 정확도(민감도: 96.41%, 특이도: 95.12%)가 나타났다. 여기서 정확도는 실험집단 중 도메인 정보가 학습집단에 포함된 경우만을 계산하였다. 490개의 실험쌍 가운데 293개의 쌍이 학습집단에 포함된 도메인을 가지고 있었고 그중 280개 쌍이 정확히 예측되었다. Ratio가 증가함에 따라 실험집단 중에서 학습집단의 PIP 분산과 겹치면서 공통된 도메인을 포함할 확률이 증가하였다. 그림 3은 학습집단의 Ratio 변화에 따른 예측 가능 단백질 쌍의 비율 변화를 보여준다.

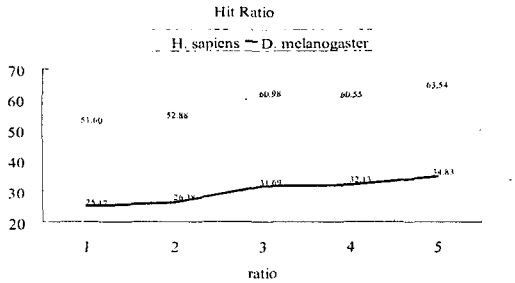

그림 3. Ratio 증가에 따른 Hit Ratio(학습집단과 겹치는 도메인을 가지며 PIP 분산과 겹치는 PIP 값을 가지는 단백질 쌍의 갯수 / 전체 실험집단 단백질 쌍의 갯수)의 변화

2.3 이종간 검증

학습 집단과 실험집단이 포함하고 있는 도메인의 유사도를 측정하기 위해서 본 논문에서는 *domain overlapping rate (DOR)*를 고안하였다. DOR은 두개의 도메인 집단 D1과 D2에 대하여 $DOR_{D1\to D2}$, $DOR_{D1\leftarrow D2}$, $DOR_{D1\leftrightarrow D2}$를 정의하며, $DOR_{D1\to D2}$는 D2에 대한 D1의 겹침정도, $DOR_{D1\leftrightarrow D2}$는 D1과 D2 집단의 전체적인 유사도를 나타낸다. 세가지 DOR을 수식으로 나타내면 다음과 같다.

$$DOR_{D1\to D2} = \frac{|D1\cap D2|}{|D1|}$$

$$DOR_{D2\to D1} = \frac{|D2\cap D1|}{|D2|}$$

$$DOR_{D1\leftrightarrow D2} = \frac{|D1\cap D2|}{|D1\cup D2|} \qquad (1)$$

초파리의 경우에서 살펴보았듯이, 다른 종에서도 적정한 크기의 출현 확률 행렬을 만들 수 있을 만큼의 학습 집단을 얻을 수 있다면 도메인 조합 기법을 적용할 수 있을 것이다. 그러나, 표 1에서 살펴보았듯이 그동안 활발하게 진행되어 온 연구는 몇몇 종에 한정되어 있는 것이 사실이다. 이를 해결하기 위하여 우리는 모든 종은 공통의 도메인을 가지며 상호작용 하는 도메인 패턴은 종에 상관없이 일정하다는 가정으로 다른 종의 출현 확률 행렬을 통한 상호작용 예측을 수행하였다. 실험을 위하여 초파리 단백질로 구성된 학습집단, 효모 단백질로 구성된 학습집단, 초파리와 효모 단백질의 조합으로 구성된 합습집단을 준비하였다. 실험집단으로 사용된 종은 DIP에서 공개하고 있는 Yeast, Fly, *C. elegans, Human, H. pylori, E. coli, Mouse*가 사용되었다. 상호작용 실험집단은 공개된 상호작용 단백질 쌍에서 두 단백질의 도메인 정보를 모두 알 수 있는 쌍들로 구성되었고, 비상호작용 실험집단은 도메인 정보를 알 수 있는 단백질로 같은 수의 쌍을 생성하여 구성하였다. 이 때, 학습집단과 실험집단의 DOR을 계산하고 정확도의 추이와 비교하였다.

표 1. 초파리 단백질 기반의 학습집단 사용 결과

	DOR(|D1 ∩ D2|)	I	II	III	IV
Yeast	44.81(1519)	70.40	40.35	86.44	87.27
C. elegans	41.20(1233)	73.80	30.95	81.08	81.52
Human	16.67(491)	68.54	32.83	80.00	83.33
H. pylori	12.91(411)	31.93	70.23	100.00	100.00
E. coli	10.45(329)	46.05	65.51	100.00	100.00
Mouse	5.57(164)	67.69	26.15	100.00	100.00

표 2. 효모 단백질 기반의 학습집단 사용 결과

	DOR(|D1 ∩ D2|)	I	II	III	IV
Fly	44.81(1519)	40.50	64.10	61.96	86.84
C. elegans	32.63(918)	42.10	63.40	70.89	79.59
H.pylori	16.81(446)	22.93	77.55	100.00	-
E. coli	13.30(350)	39.88	70.13	96.55	-
Human	13.25(344)	54.95	62.50	92.00	66.67
Mouse	4.08(95)	53.85	60.00	100.00	-

표 3. 효모, 초파리 조합 단백질 기반의 학습집단 사용 결과

	DOR(|D1 ∩ D2|)	I	II	III	IV
Fly	79.71(2702)	87.75	52.55	94.39	87.13
Yeast	65.1(2207)	92.50	52.40	97.14	79.05
C. elegans	36.65(1318)	53.05	55.85	85.94	84.26
Human	14.30(516)	58.93	49.59	87.65	69.23
H. pylori	13.73(517)	21.61	79.11	100.00	100.00
E. coli	10.51(396)	38.92	73.41	100.00	-
Mouse	4.49(155)	69.23	36.92	88.89	-

I. 전체 실험집단의 민감도
II. 전체 실험집단의 특이도
III. 학습 집단의 AP matrix[4] 와 일치하는 실험집단의 민감도
IV. 학습 집단의 AP matrix 와 일치하는 실험집단의 특이도

3. 결론 및 향후 과제

본 논문을 통하여 우리는 도메인 조합 기법이 다른 고등 생물에 대하여도 충분한 상호작용 데이터와 도메인 데이터를 확보할 수 있다면 예측 기법의 특별한 수정 없이도 적용이 가능하다는 것을 확인하였다. 향후 인터넷을 통한 지속적인 데이터의 확보와 동시에 단백질 구조 정보나 서열에서 직접 도메인을 찾아내는 등의 다양한 기술접목을 통하여 예측 가능한 단백질 쌍의 갯수를 늘려 나갈 것이다.

4. 참고문헌

[1] Minghua Deng and et al. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12:1540-1548, 2002.

[2] Anne-Claude Gavin and et al. Functional organization of the yeast proteome by systematic analysis of protein complexe. *Nature*, 415(10):141-147, 2002.

[3] Dong-Soo Han and et al. Domain combination based probabilistic framework for protein-protein interaction prediction. *Genome Informatics*, 14:250-259, 2003.

[4] Dong-Soo Han and et al. PreSPI: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Research*, 32(21), 2004.

[5] Dong-Soo Han and et al. PreSPI: Design and implementation of protein-protein interaction prediction service system. *Genome Informatics*, 15(2):171-180, 2004.

[6] Lukasz Salwinski and et al. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):d449-d451, 2004.

[7] Suraj Peri and et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13:2363-2371, 2003.