Inter-Species Validation for Domain Combination Based Protein-Protein Interaction Prediction Method

Woo-Hyuk Jang[1] Dong-Soo Han[1] Hong-Soog Kim[1] Sung-Doke Lee[1]
[1]*School of Engineering, Information and Communications University,*
119, Munjiro, Yuseong-gu, Daejeon 305-732, Korea
Email : torajim@icu.ac.kr, dshan@icu.ac.kr, kimkk@icu.ac.kr, sdlee@icu.ac.kr

ABSTRACT: 도메인 조합에 기반한 단백질 상호작용 예측 기법은 효모와 같은 특정 종에 대하여 우수한 예측 정확도를 보이는 것으로 알려졌으나, 인간과 같은 고등 생명체의 단백질에 대한 상호작용 예측을 수행하기 위하여는 여러종에 대한 기법의 적절성 검증과 최적의 학습집단 구성 방안에 대한 연구가 선행되어야 한다. 본 논문†에서는, 초파리 단백질을 이용한 예측 정확도 검증으로 도메인 조합 기법의 일반화 가능성을 타진 하고 이종간의 상호작용 예측 실험 및 정확도 검증을 통하여 비교적 연구가 덜 되어진 종의 단백질 상호작용 예측을 위한 학습집단 구성 방법에 대하여 기술한다. 초파리 실험에서는 10351개의 상호작용이 있는 단백질 쌍 가운데, 80%와 20%를 각각 학습집단 및 실험집단으로 사용하였으며, 상호작용이 없는 단백질 쌍의 학습집단은 1배에서 5배까지 변화시키면서 예측 정확도를 관찰하였다. 이 결과 77.58%의 민감도와 92.61%의 특이도를 확인하였다. 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 INTRODUCTION

단백질 상호작용 예측 분야에 있어서 해당 단백질의 도메인 정보를 이용하는 방법은 다양한 형태로 연구되어 왔다. 대부분의 도메인 기반 단백질 상호작용 예측 틀들은 단백질 상호작용 정보와 해당 단백질에 속한다고 알려진 도메인 정보를 이용하여 도메인-도메인 상호작용을 유추하고, 다시 이 정보를 바탕으로 미지의 단백질 상호작용을 예측하는 방식을 취하고 있다 [1, 10, 13]. 현재 인터넷을 통하여 획득할 수 있는 단백질 관련 데이터의 양은 폭발적으로 증가하고 있는 추세이며, 따라서 대부분의 도메인 기반 상호작용 예측 기법들은 예측에 필요한 정보를 이와 같이 인터넷에 공개된 데이터를 수집 및 가공하여 사용하고 있다. 학습집단으로 사용가능한 정보가 인터넷을 통하여 지속적으로 공급된다는 점은 도메인 기반의 상호작용 예측과 같은 계산적인 접근 방법의 점진적 발전을 지원하고 있다.

도메인 기반의 상호작용 연구 초기에 소개된 많은 예측 기법들은 주로 단일 도메인 쌍을 모델로 하고 있다. 이러한 초창기 모델은 단백질 쌍의 상호작용을 단백질 내의 도메인간 상호작용으로 해석하는 확률적 기반을 마련하였으나, 실제 활용되기에는 낮은 예측 정확도를 보여주었다. 이에 예측 정확도 향상을 위하여 도메인 조합 기반의 상호작용 예측 틀이 Han 그룹에 의하여 개발되었다 [3, 4, 5, 6]. 이 예측 방법에서는 단일 도메인 쌍 뿐만 아니라 도메인 조합 쌍을 단백질 상호작용의 기본 단위로 가정함으로써 특정한 도메인간의 작용에 있어서 주변 도메인들의 영향을 고려하였다. 그 결과 높은 예측 정확도가 달성되었고 기법의 안정성 역시 DIP, DIP CORE, HMS-PCI, TAP 데이터를 통하여 검증되었다 [2, 7, 14]. 도메인 조합 기반 예측 기법은 우수한 정확도에도 불구하고 실제 활용이 가능하도록 고등 생물체로의 확장이 필요하다. 그러나 현재 인터넷을 통해 공개되는 단백질-단백질 상호작용 정보 및 도메인 정보의 양은 효모[1,2] 및 초파리[14]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 초파리를 이용한 예측 실험과 이종간의 상호작용 예측 실험을 수행하였다. *Database of Interacting Protein (DIP)*[14] 에서 확보한 초파리의 상호작용 단백질 쌍 20988개 중, *Integrated documentation resource of Protein families, domains and functional sites(InterPro)*[12] 와 *Protein Information Resource (PIR)*[15] 에서 도메인 정보를 찾을 수 있는 10351개의 단백질 쌍이 실험에 사용되었다.

실험은 기존의 효모에서와 같이 사용가능한 단백질 쌍 중 80%와 20%를 각각 학습집단 및 실험집단으로 구분하여 정확도를 분석하였다. 실험 결과 예측 정확도는 효모의 경우와 유사한 정도 (민감도: 77.58%, 특이도: 92.61%)를 나타냄으로써, 도메인 조합 기법이 특별한 변형없이 다른 종에서도 적용될

† 본 과제는 과학재단의 2005 년도 특정연구개발 사업 (M10529000011-05N2900-01110)의 지원을 받고 있음.

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 클수록 예측 정확도가 높게 관찰되었다. 이것은 곧 학습집단과 실험집단 도메인 사이의 유사도가 예측 기법상의 정확도에 중요한 척도임을 의미한다. 본 논문에서는 인터넷을 통하여 수집할 수 있는 단백질 및 도메인 데이터들을 사용하여 구축한 이종 단백질 집단 사이의 도메인 유사도를 나타내기 위하여 *Domain Overlapping Rate*(DOR)를 소개하였다. 실험에 사용된 종들에 대한 도메인 유사도 조사 결과, 예상대로 생물학적으로 유사한 종들 사이의 DOR이 높게 나타나는 것을 확인하였다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 도메인 조합 기반의 단백질 상호작용 예측 기법에 대한 간략한 설명을 한다. 3장에서는 이종간의 검증에 사용된 데이터와 검증 절차에 대하여 기술한다. 4장에서는 결과와 그 의미에 대한 분석이 이루어 지며, 마지막으로 5장에서 결론을 맺는다.

2 DOMAIN COMBINATION BASED PROTEIN-PROTEIN INTERACTION PREDICTION METHOD

여기서는 본 논문의 검증에 사용된 도메인 조합 기반의 단백질 상호작용 예측 기법에 대한 간략한 소개를 한다. 예측 기법에 대한 자세한 내용은 [3]과 [6]에서 살펴 볼 수 있다.

2.1 Motivation

도메인 조합 기반 단백질 상호작용 예측 기법은 초기 도메인 기반의 예측 기법[1, 10, 13]에 기초하고 있으면서도 기존의 단일 도메인 기반 접근방법에서 나타나는 몇가지 문제점들을 개선한 기법이다. 대부분의 도메인 기반 상호작용 예측 기법은 단백질 사이의 상호작용은 도메인끼리의 상호작용 결과라고 가정한다. 이들은 단백질 상호작용 정보로부터 도메인간의 상호작용 정보를 유추하고, 이를 기초로 하여 새로운 단백질 쌍의 상호작용을 예측한다. 그러나 초기의 예측 방법들은 단일 도메인 쌍만을 고려하였고, 계산적 편의를 위하여 도메인 간의 상호작용은 주변의 다른 도메인에 독립적으로 일어난다고 가정하였다. 그 결과 예측 정확도는 실제 서비스로 활용되기에는 무리가 있었다.

예측 정확도가 낮은 데에는 여러 이유가 있을 수 있으나, 단지 계산적 편의를 위하여 단백질 상호작용의 기본단위를 단일. 도메인으로 가정한 것이 주요한 문제점이라고 할 수 있다. 이를 해결하기 위하여 도메인 조합 기법에서는 도메인 조합과 도메인 조합쌍을 도입하였다. 이 방법에서는 특정 단백질이 포함하는 도메인들의 멱집합을 상호작용의 기본단위로 사용하여 주변 도메인의 영향 및 도메인 그룹들간의 작용을 해석하고자 하였다. 예측 기법 관점에서 단일 도메인 기반의 예측과 도메인 조합 기법 사이의 명확한 차이점은 [3]에서 찾아 볼 수 있다.

2.2 Prediction Method

도메인 조합 예측 기법에서는 상호작용이 일어나는 단백질과 그렇지 않은 단백질의 도메인 조합쌍에 대하여 출현빈도를 나타내기 위해 *Appearance Probability* (AP) 행렬을 생성한다. 예측 기법에서는 이 행렬을 바탕으로 단백질 쌍을 0 과 1 사이의 실수로 변환하는 확률식을 개발하였다. 본 논문에서는 변환되는 실수값을 *Primary Interaction Probability* (PIP) 로 정의한다. 확률식을 모든 상호작용 단백질쌍과 비상호작용 단백질쌍에 적용시키면 두개의 PIP 분산이 얻어진다. 미지의 단백질 쌍이 주어지면 이 분산들을 기준으로 PIP 값이 계산되고 PIP 값의 카테고리에 따라서 상호작용 확률이 구해지게 된다. 도메인 조합기법의 세부 내용은 [3]에서 확인할 수 있다.

3 MATERIALS AND METHODS

3.1 Databases

대부분의 도메인 기반 상호작용 예측 기법에서는 단백질 상호작용 정보와 해당 도메인 정보를 필요로 한다. 도메인 조합에 기반한 본 논문의 예측 기법 역시 이러한 정보들을 사용하고 있다. 단백질 상호작용 정보는 *Database of Interacting Proteins* (DIP) [8, 14]에서 추출하였고, 도메인 정보는 *Integrated documentation resource of Protein families, domains and functional sites* (InterPro) 와 *Protein Information Resource* (PIR) [9, 12, 15] 에서 수집하였다. DIP 은 대표적인 종(*D. melanogaster, S. cerevisiae, C. elegans, H. pylori, H. sapiens, E. coli, and M. musculus*)에 대한 단백질 상호작용 데이터를 제공하며 각각의 데이터들은 InterPro 에서 사용되는 SWISSPROT ID 를 포함하고 있어 우리의 예측 기법에서 주요하게 사용되었다. 또한 상호작용 데이터가 지속적(The full sets: a month, CORE subsets: 3 months)으로 업데이트 되기 때문에 예측 시스템의 학습집단의 질을 꾸준히 향상시킬 수 있다는 장점을 가지고 있다. 본 논문에서는 Oct. 3, 2004 에 릴리즈된 DIP 의 단백질 상호작용 데이터를 사용하여 실험하였다. InterPro 는 초기에 분산되어 있는 여러 생물 데이터베이스를 하나로 통합하고자 시작되었고, 현재 풍부한 양의 도메인 데이터들을 보유하고 있다. InterPro 에서는 DIP 에서 제공하는 종들의 taxonomy ID 를 통하여 'IPRXXXX' 형태의 도메인 정보를 찾아낼 수 있다. DIP 과 InterPro 를 이용하여 단백질-도메인 사전(dictionary)을 구축하는 한가지 방법은 SWISSPROT 을 이용하여 사상(mapping) 시키는 것이다. InterPro 는 SWISSPROT ID 를 가지는

단백질에 대하여 90% 이상의 coverage 를 가진다. 이는 단백질-도메인 사전 구축에서 SWISSPROT ID 를 알 수 있는 단백질의 대부분은 도메인 정보를 찾아낼 수 있다는 것을 의미한다. 본 논문에서 사용한 또다른 유용한 단백질 데이터베이스로 *Protein Information Resource*(PIR)이 사용되었다. PIR 은 DIP 에서 제공하는 단백질을 InterPro 의 도메인으로 사상하는 과정에서 생겨나는 정보의 손실을 막기 위하여 사용되었다. 도메인 조합 예측 기법이 높은 예측 정확도를 보여주기 위해서는 충분한 학습집단이 준비되어야 한다. 따라서 학습집단 구축의 핵심인 단백질-도메인 사전 구축은 매우 중요한 작업이며, 본 논문의 실험을 위하여 최대한 많은 도메인 정보를 찾아내기 위하여 노력하였다.

3.2 Validation Procedure

이종간의 단백질 상호작용 예측 검증에 앞서 도메인 조합 기반 예측 기법의 일반화 가능성에 대한 실험이 수행되었다. 도메인 조합 기법은 효모에서 높은 정확도를 보여주고 있지만 다른 종에서도 유사한 정확도를 보여줄 수 있는지는 명확하지 않았다. 이 실험에서는 그동안 비교적 많은 연구를 통해 충분한 학습집단을 얻을 수 있는 초파리를 사용하여 예측 정확도를 측정하고 효모에서의 실험결과와 비교하였다.

이종간의 검증을 위해서 세 종류의 출현 확률 행렬(AP matrices)이 준비되었다. 첫번째는 효모 단백질 쌍으로만 이루어진 행렬과, 초파리 단백질로 이루어진 것, 마지막으로 효모와 초파리 단백질 쌍을 합한 것으로 만들어진 행렬이다. 효모와 초파리는 인터넷을 통하여 얻을 수 있는 단백질 상호작용 및 도메인 데이터가 풍부하여 출현 확률 행렬을 만드는 과정에서 손쉽게 학습집단을 구성할 수 있기 때문에 주요하게 사용되었다. 세가지 행렬이 구성된 후, 각각의 행렬을 이용하여 Human, Mouse, *H. pylori, E. coli, C.elegans* 의 단백질 쌍에 대한 상호작용 예측이 수행되었다. 이 과정에서 학습집단과 실험집단의 도메인 겹침 정도와 예측 정확도 사이의 상관 관계를 살펴 보았다. 한편 도메인 겹침 정도를 더욱 정량적으로 나타내기 위하여 본 논문에서는 *Domain Overlapping Rate* (DOR)를 고안하였다. 이종간의 상호작용 실험에 앞서 이종간의 DOR 을 조사하여 상호작용 실험의 결과를 분석하는데 사용하였다.

3.3 Domain Overlapping Rate

우리는 두 도메인 집합간의 유사도 정도를 측정하기 위하여 Domain Overlapping Rate(DOR)을 고안하였다. DOR은 두개의 도메인 집단 D1과 D2에 대하여 $DOR_{D1\rightarrow D2}$, $DOR_{D1\leftarrow D2}$, $DOR_{D1\leftrightarrow D2}$를 정의하며, $DOR_{D1\rightarrow D2}$는 D2에 대한 D1의 겹침정도, $DOR_{D1\leftrightarrow D2}$는 D1과 D2 집단의 전체적인 유사도를 나타낸다. 세가지 DOR을 수식으로 나타내면 다음과 같다.

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

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

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

그림 1. $DOR_{D1\leftrightarrow D2}$ Heptagon

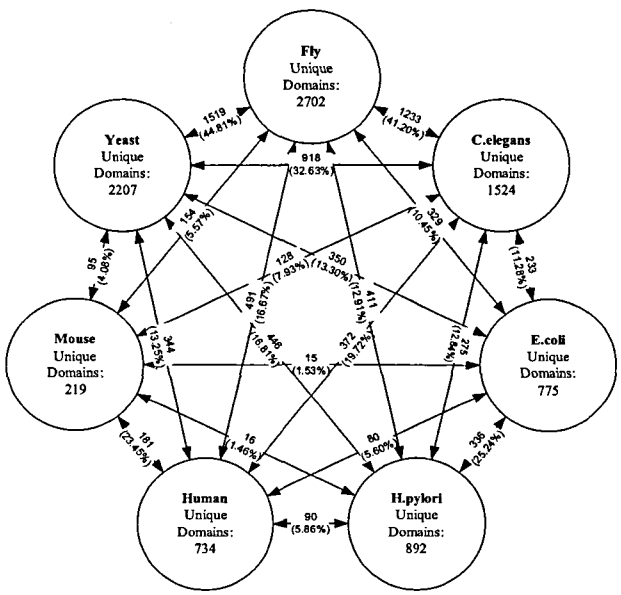


그림 1은 이종간의 DOR에 대하여 조사한 결과이다. 예상대로 비슷한 종들간의 DOR은 그렇지 않은 종들간의 DOR보다 대체로 높은 값을 보여준다. 도메인 조합 기반의 예측 방법에서 도메인 정보는 예측 정확도에 큰 영향을 미치는 요인이다. 인터넷을 통하여 구할 수 있는 단백질 상호작용 및 도메인 정보가 충분하다면 동종의 학습집단을 구성하는 것이 가장 좋은 방법이겠지만, 그렇지 못할 경우 DOR이 학습집단의 질을 평가하는 중요한 수치가 될 수 있으며, 좀더 높은 DOR을 가지는 학습집단을 구성할 수록 예측 정확도는 향상된다고 말할 수 있다.

4 VALIDATION AND RESULTS

4.1 D. melanogaster Validation

초파리는 그동안 활발히 연구가 진행되어온 종으로, 예측 정확도 실험에 필요한 충분한 크기의 학습집단을 쉽게 구성할 수 있는 장점이 있다. 실험을 시작하기 전에 우리는 DIP 에서 제공하는 종들의 단백질과 해당 도메인들에 대하여 통계를 구하였다 (표 1). 효모와 초파리는 하나의 단백질이 보유하는 평균 도메인 갯수는 거의 유사하나, 초파리의 최대 도메인 갯수는 현재 시스템에서 처리하기에 무리가 있었다. 도메인 조합을 이용하는 예측 기법에서 하나의 단백질이 n 개의 도메인을 가지고 있다고 가정하면 도메인 조합의 갯수는

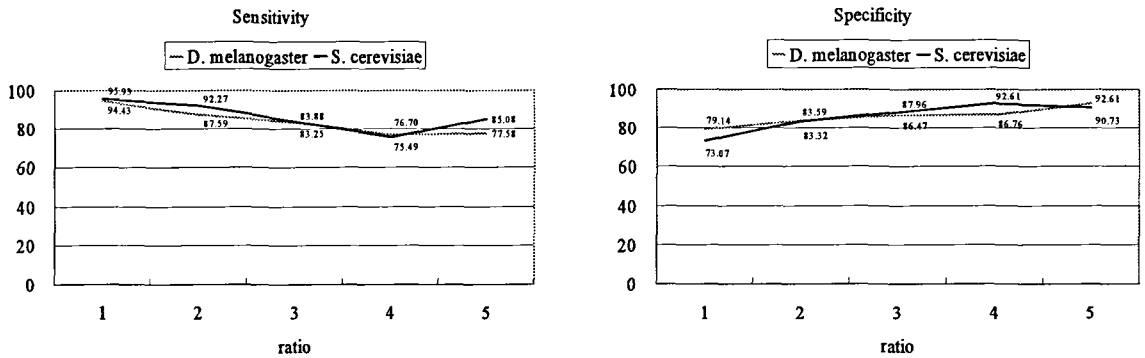
2^n-1 개가 된다. 초파리의 경우 2^{21}-1 개의 도메인 조합을 생성하는 단백질을 포함하고 있었으며 이는 시스템에서 다루기에 많은 비용이 소모되어 학습집단에서 제외하였다. 제외된 단백질은 FlyBase ID FBgn0003189 를 가지는 *rudimentary* 이다. 효모와 초파리 사이의 평균 도메인 갯수는 비슷하게 나타났

표 1. 여러 종들에 대한 도메인 통계

Species	I	II	III	IV	1~2	3~4	5~9	10~
D. melanogaster	21	2.12	1.41	2702	3525(70.9%)	1130(22.7%)	309(6.2%)	10(0.2%)
S. cerevisiae	13	2.13	1.40	2207	2133(70.2%)	716(23.6%)	182(6.0%)	7(0.2%)
C. elegans	13	2.23	1.47	1524	1317(67.9%)	470(24.2%)	149(7.7%)	5(0.2%)
H. pylori	13	2.30	1.67	892	378(67.5%)	122(21.8%)	58(10.4%)	2(0.3%)
H. sapiens	12	3.27	1.93	734	259(40.4%)	245(38.2%)	128(20.0%)	9(1.4%)
E. coli	10	2.79	1.77	775	214(53.4%)	125(31.1%)	60(15.0%)	2(0.5%)
M. musculus	12	3.58	2.15	219	45(34.1%)	50(37.9%)	33(25.0%)	4(3.0%)

I: 하나의 단백질이 보유한 도메인의 최대 갯수 II: 하나의 단백질이 보유한 도메인의 평균 갯수
III: 하나의 단백질이 보유한 도메인 갯수의 표준편차 IV: 현재 사용가능한 유니크 도메인의 갯수
X~Y: X~Y 개의 도메인을 가진 단백질의 갯수

그림 2. 효모와 초파리 단백질의 상호작용 예측 정확도 비교



으나 초파리의 경우 유니크 도메인 갯수가 효모의 경우보다 많기 때문에 출현 확률 행렬의 크기가 상대적으로 증가하였고 행렬 처리에 있어 더욱 많은 시간이 소요되었다. 게다가 향후 인간 단백질에 대한 연구를 시작할 경우 초파리의 경우보다 더욱 많은 도메인 조합을 필요로 하기 때문에 이 경우 특별한 조작이 필요할 것으로 전망된다.

실험은 효모의 경우와 마찬가지로, 8,280 개의 상호작용 단백질 쌍(전체 상호작용 쌍의 80%)을 학습집단으로 사용하였고, 2,071 개의 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 1 배에서 5 배로 늘려가면서 예측 정확도를 관찰하였다. 표 2 는 초파리 단백질 쌍에 대한 상호작용 예측 결과를 보여 준다. 학습집단에서 만들어진 PIP 분산과 겹치는 PIP 값을 가지는 단백질 쌍들이 그렇지 않은 쌍들보다 월등한 예측 정확도를 보여준다. 표에서 Ratio 는 학습집단에서 상호작용 하는 단백질 쌍의 갯수에 대한 비상호작용 단백질 쌍의 비율을 나타낸다. 상호작용 한다고 알려진 단백질 쌍의 갯수는 일정하기 때문에 Ratio 1.0~5.0 사이의 비상호작용 단백질 쌍의 갯수는 증가하였다. 우리는 Ratio 가 자연계에 존재하는 상호작용이 있는 단백질 쌍의 갯수와 그렇지 않은 단백질 쌍의 갯수 사이의 비율에 접근할 수록 예측 정확도가 높아질 것이라고 예상한다. 그림 2 는 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.

표 2. 초파리 단백질 쌍에 대한 예측 결과

	Ratio	1.0	2.0	3.0	4.0	5.0
I	Sensitivity	99.54	89.79	86.39	79.55	78.99
	Specificity	82.86	87.20	87.96	89.78	94.20
II	Sensitivity	64.26	58.62	43.33	43.33	34.36
	Specificity	56.52	64.52	72.41	64.29	76.47
Total	Sensitivity	94.43	87.59	83.25	76.70	77.58
	Specificity	79.14	83.59	86.47	86.76	92.61

I: 학습집단의 PIP 분산에 겹치는 PIP 값의 단백질 쌍
II: 학습집단의 PIP 분산에 겹치지 않는 PIP 값의 단백질 쌍

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.

4.2 Inter-Species Validation

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

H. pylori, E. coli, Mouse 등은 충분한 실험집단이 확보되지 않아, 해석에 무리가 있으나, 표 1, 표 2의 음영에서 학습집단의 출현 확률 행렬에 대한 DOR이 클 수록 높은 예측 정확도를 보여 주었다. 또한 표 3의 음영부분을 통하여 우리는 효모와 초파리 단백질을 합한 학습집단에서도 각각의 종은 동종의 단백질로 구성된 학습집단의 경우에 비해 전혀 예측 정확도가 떨어지지 않음을 알 수 있었다. 이는 곧 상호작용이 일어나는 도메인의 패턴은 종에 상관없이 일정하며, 부족한 단백질 정보를 가진 종의 학습집단 구성시, 실험종과의 DOR을 높힐 수록 예측 정확도 역시 높아진다는 결론을 얻을 수 있었다.

5 CONCLUSION

본 논문에서 우리는 초파리 단백질에 대한 실험을 통하여 도메인 조합 기법이 다른 종에도 적용될 수 있는지에 대해 확인하였다. 그 결과 효모 단백질의 경우와 유사한 정도의 높은 민감도와 특이도를 얻을 수 있었다. 이것은 곧 도메인 조합 기법은 다른 고등 생물에 대하여도 충분한 상호작용 데이터와 도메인 데이터를 확보할 수 있다면 예측 기법의 특별한 수정 없이도 적용이 가능하다는 것을 의미한다.

학습집단의 구성방안을 연구하기 위하여 이종간의 상호작용 예측 실험이 수행되었다. 학습집단의 도메인과 실험집단의 도메인이 유사할 수록 예측 정확도는 향상되었다. 또한 동일한 종으로만 학습집단을 구성한 경우와 다른 종과의 합으로 학습집단을 구성한 경우 각각에 대하여 예측 정확도는 큰 차이를 보이지 않았다. 이것은 곧 상호작용이 일어나는 단백질 패턴은 종에 상관없이 일정하며 실험 집단과 다른 종의 단백질 상호작용 데이터를 학습집단으로 사용할 수 있음을 의미하고 있다. 예를 들면 초파리나 효모와 같은 단백질 상호작용 데이터를 이용하여 학습집단을 구성하고 이를 이용하여 mouse 혹은 human의 단백질 상호작용을 예측하는 것이다.

DOR 은 이종간의 단백질 상호작용 예측에 있어서 중요한 역할을 하지만 조사결과 그리 크지는 않은 것으로 나타났다. 이러한 낮은 DOR 은 기본적으로 종 사이의 유니크 도메인 차이일 수도 있으나, 현재까지 알려진 도메인 자체의 부족 역시 하나의 원인일 수 있다. 그러나 이는 단백질 구조 정보나 서열에서 직접 도메인을 찾아내는 기법을 통하여 어느 정도 해결이 가능할 것으로 예상한다. 또한 인터넷에 공개되는 단백질 상호작용 정보 및 도메인 정보의 질이 개선될 수록 도메인 조합 기법의 예측 정확도도 향상할 것으로 기대한다.

현재 우리는 예측 시스템에 도메인을 찾아 낼 수 있는 여러 기술을 접목할 예정이며 단백질 상호작용 지도를 구성하고 사용자와의 대화를 통하여 상호작용 타겟 단백질 후보군을 찾아내는 방법을 계획하고 있다.

REFERENCES

[1] Minghua Deng and et al. Ingerring 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. Domain combination based proteinprotein interaction possibility ranking method. *Proceedings of 4th IEEE symposium on Bioinformatics and Bioengineering*, pages 434–441, May 2004.

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

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

[7]Yuen Ho and et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(10):180–183, 2002.

[8]Xenarios I. and EisenbergD. Protein interaction databases. *Curr.Opinion in Biotechnology*, 12:239–241, 2001.

[9]Nicola J. and et al. The interpro database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, 31(1):315–318, 2003.

[10]Wojcik J. and Schachter V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17:S296–S305, 2001.

[11]Berman H. M. and et al. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[12]Apweiler R. and et al. The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research*, 29:37–40, 2001.

[13]Ng S., Zhang Z., and Tand S. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19:923–929, 2003.

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

[15]Cathy H. Wu and et al. The protein information resource. *Nucleic Acids Research*, 31(1):345–347, 2003.