■ 2009년도 학생논문 경진대회 수상작

상호작용 중요도 행렬을 이용한 단백질-단백질 상호작용 예측

(Protein-Protein Interaction Prediction using Interaction Significance Matrix)

장 우 혁 † 정 석 훈 † 정 휘 성 † 현 보 라 † 한 동 수 ††
(Woo-Hyuk Jang) (Suk-Hoon Jung) (Hwie-Sung Jung) (Bora Hyun) (Dong-Soo Han)

요 약 최근 계산을 통한 단백질 상호작용 예측 기법 중, 단백질 쌍이 포함하고 있는 도메인들 사이의 관계에 중점을 둔 도메인 정보 기반 예측 기법들이 다양하게 제안되고 있다. 하지만, 다수의 도메인 쌍들이 상호작용에 기여하는 정도를 정밀하게 반영하는 계산 기법은 드문 실정이다. 본 논문에서는 단백질 상호작용에 있어 도메인 조합 쌍의 상호작용 영향력을 수치화하여 반영한 상호작용 중요도 행렬을 고안하고 이를 기반으로 한 단백질 상호작용 예측 시스템을 구현한다. 일반적인 도메인 조합 기법과 달리, 상호작용 중요도 행렬에서는 상호작용을 위한 도메인간의 협업 확률이 고려된 Weighted 도메인 조합과, 다수의 Weighted 도메인 조합 중 실제 상호작용 주체가 될 확률을 도메인 조합 쌍의 힘(Domain Combination Pair Power, DCPPW)으로 수치화한다. DIP과 IntAct에서 얻어온 S. cerevisiae의 단백질 상호작용 데이터와 Pfam-A 도메인 정보를 사용한 정확도 검증 결과, 평균 63%의 민감도와 94%의 특이도를 확인하였으며, 학습집단의 증가에 따른 안정적인 예측 정확도 향상을 보였다. 본 논문에서 구현한 예측 시스템과 학습 데이터는 웹(http://code.google.com/p/prespi)을 통하여 내려 받을 수 있다.

키워드 : 도메인 조합 쌍, Weighted 도메인 조합, 도메인 조합 쌍의 힘, 단백질 상호작용 예측, 도메인 조합 협업 확률

Abstract Recently, among the computational methods of protein-protein interaction prediction, vast amounts of domain based methods originated from domain-domain relation consideration have been developed. However, it is true that multi domains collaboration is avowedly ignored because of computational complexity. In this paper, we implemented a protein interaction prediction system based the *Interaction Significance* matrix, which quantified an influence of domain combination pair on a protein interaction. Unlike conventional domain combination methods, IS matrix contains weighted domain combinations and domain combination pair power, which mean possibilities of domain collaboration and being the main body on a protein interaction. About 63% of sensitivity and 94% of specificity were measured when we use interaction data from DIP, IntAct and Pfam-A as a domain database. In addition, prediction accuracy gradually increased by growth of learning set size, The prediction software and learning data are currently available on the web site.

Key words : domain combination pair, weighted domain combination, domain combination pair power, protein interaction prediction, possibility of domain collaboration

· 본 연구는 교육과학기술부와 한국산업기술재단의 지역혁신 인력양성사업과 국가생물자원정보관리센터 운영체제 구축 사업의 지원으로 수행된 연구 결과임

† 학생회원 : KAIST 정보통신공학과
torajim@kaist.ac.kr
sh.jung@kaist.ac.kr
hwiesung@kaist.ac.kr
hyunbr@kaist.ac.kr
†† 종신회원 : KAIST 전산학과 교수
dshan@kaist.ac.kr

논문접수 : 2009년 6월 2일
심사완료 : 2009년 8월 14일

정보과학회논문지: 소프트웨어 및 응용 제36권 제10호(2009.10)

1. 서 론

인터넷을 통하여 공개되는 단백질 관련 정보의 증가와 함께 계산을 통한 단백질 상호작용 예측은 최근 몇 년간 Gene fusion, 단백질 구조 분석, 도메인 정보 활용[1-3] 등의 다양한 방법으로 발전되어 왔다. 그 중에서도 도메인 정보를 이용한 단백질 상호작용 예측 기법은 연상법(association method)[3], 확률적 방법(probabilistic method)[4-6], SVM 기반 예측[7] 등으로 활발히 연구되고 있다. 이러한 도메인 정보 기반 예측 기법들은 단백질은 다수의 도메인으로 구성되어 있고, 단백질 상호작용은 도메인들 사이의 상호작용을 통해 이루어 진다는 것에 바탕을 두고 있다[4,8].

대부분의 도메인 정보 기반 예측은 도메인 상호작용 쌍을 단백질 상호작용의 기본 단위로 여기고 있으나, 기존의 연구에서는 각 도메인 상호작용 쌍들이 단백질 상호 작용에 독립적으로 영향을 끼친다고 간주한 것에 반해 최근에는 두 개 또는 그 이상의 도메인 쌍이 서로 협력한다는 연구가 보고되고 있다[9,14]. Jung[10]은 한 단백질 내의 이웃 도메인들이 서로 영향을 주고받는다는 연구[11,12]를 바탕으로 도메인 간의 상호 의존성을 측정하고자 하였다. 이 연구에서는 진화상에서 특정 기능을 위한 도메인 보존 정도를 계산하기 위하여 연관성 규칙인 all-confidence를 적용하였고, 도메인의 기능 협업도는 all-confidence와 상관관계가 있음을 밝혔다.

Han[13] 그룹은 도메인 간의 상호 영향력을 확률 값에 포함하기 위해 도메인 조합(domain combination)을 제안하고, 이들 조합간의 상호작용 정보를 바탕으로 단백질 쌍의 상호작용을 예측하는 방법을 고안하였다. 이 때, 실험적으로 밝혀진 단백질 상호작용 쌍에서 도메인 조합의 출현 빈도를 출현 확률 행렬(Appearance Probability Matrix)로 구성하고 미지의 단백질 쌍에 대한 상호작용 가능성을 확률 값으로 제공하고자 하였다. 그러나 출현 확률 행렬에서는 모든 도메인 조합의 상호작용 기여도를 동일하다고 가정하고 있으며, 예측에 있어서도 출현 확률 행렬에 포함되어 있지 않은 도메인을 제외함으로써, 예측 가능한 단백질 쌍의 범위가 줄어드는 문제점을 지니고 있다. 여러 도메인의 협업(multi-domain cooperation) 정보를 바탕으로 단백질 상호작용을 예측하고자 한 Wang[14]의 연구에서는 단백질 상호작용에서 각 도메인 상호작용이 끼치는 영향력이 다름을 밝히고 있다.

본 논문에서는 단백질 상호작용 데이터에서 도메인 조합 쌍의 상호작용 영향력을 수치화하는 상호작용 중요도 행렬(Interaction Significance)을 제안한다. 행렬의 구축을 위해, 하나의 단백질 내에 포함된 도메인 조합을 협업 가능성에 따라 차등을 둔 Weighted 도메인 조합으로 확장하였고, 이를 바탕으로 하나의 상호작용에서 단백질 쌍이 생성할 수 있는 여러 Weighted 도메인 조합 중 실제 상호작용 주체가 될 확률을 도메인 조합 쌍의 힘(Domain Combination Pair Power, DCPPW)으로 정의하였다.

새롭게 도입된 개념은 기존 방식에서 도메인 조합 쌍 각각이 상호작용에 동일한 영향을 끼칠 것이라는 가정을 보완하여 예측 정확도의 향상에 기여한다.

본 논문에서는 단백질 상호작용 데이터로 DIP과 IntAct를 사용하였다, 또한 도메인 정보는 Pfam-A를 사용하였으며, 데이터의 통합을 위하여 UniProt의 정보를 이용하였다. 제안된 예측 방법의 검증을 위하여, 기존의 도메인 조합 기반 예측 방식과의 예측 정확도 비교를 수행하였으며, 학습집단의 크기 변화에 따른 예측 정확도의 변화 추이를 측정하였다. 그 결과, 기존의 방식에 비해 개선된 예측 결과인, 평균 sensitivity 63%, specificity 94%의 예측 정확도를 나타내었으며, 약 9배의 학습 집단 크기 증가에 대하여 specificity의 감소 없이 sensitivity는 약 40% 정도 향상되는 것을 확인하였다. 본 논문의 구성은 다음과 같다. 먼저, 2장에서 기존 도메인 조합 쌍의 개념에 대하여 간략히 설명하고 학습 집단으로 사용한 DB를 소개한다. 3장에서는 상호작용 중요도(Interaction Significance) 행렬의 구축을 Weighted 도메인 조합 및 도메인 조합 쌍 힘(Domain Combination Pair Power)을 중심으로 기술한다. 4장에서는 기존의 도메인 조합 쌍 예측과의 비교 검증 및 예측 정확도 측정 결과를 소개하고 5장을 통해 결과에 대한 분석을 한 후, 마지막 6장에서 결론을 내린다.

2. 도메인 조합(Domain Combination)과 도메인 조합 쌍(Domain Combination Pair)

도메인은 단백질을 이루는 단위로서, 단백질의 진화상에서 보존되어 있으며 고유의 기능 및 3차 구조를 가진다. 도메인은 단백질 기능의 기본 단위로 간주되고 있기 때문에, 단백질 상호작용 예측 연구에서도 이를 이용한 연구가 진행되어 왔다. 대부분의 연구에서는 단백질 쌍 각각의 도메인들이 상호작용하여 단백질 상호작용을 이룬다는 관점을 바탕으로 예측 모델을 제안하였다. 도메인 조합은 기존의 단일 도메인이 아닌 도메인 조합 쌍이 상호작용의 단위라는 가정을 바탕으로 한다. 이것은 도메인이 기본 기능 단위로 간주된다 하더라도 독립적으로 기능하지 않으며, 한 단백질 내의 이웃 도메인과 서로 영향을 주고받는다는 생각에서 비롯된다. 단백질 및 도메인을 구성하는 폴리펩타이드 폴딩은 기능과 3차 구조 형성에 있어서 주변 환경에 민감한 특성을 지닌다.

따라서, 한 도메인이 제대로 기능하기 위해서는 가장 가까운 주변 환경인 이웃 도메인의 영향을 직접 혹은 간접적으로 받을 수 밖에 없다. 또한 한 단백질 안에서 다수의 도메인이 서로 협력하여 상대 단백질 내의 도메인과 직접적인 상호작용을 할 수도 있을 것이다. 실제로 PDB에 실험적으로 3차 구조가 보고된 단백질 상호작용 쌍 중, 50여 개의 상호작용 쌍이 단일 도메인이 아닌 복수의 도메인이 협력하여 상호작용을 형성하였다는 것이 보고되었다[14].

단백질 쌍이 여러 도메인이 협력하여 상호작용을 하는 경우, 정밀한 상호작용 예측을 위해서는 각 도메인 쌍들이 단백질 상호작용에 기여하는 정도를 차등하여 반영하는 것이 필요하나, 기존에 소개된 예측 기법들에서는 계산적인 복잡도로 인하여 무시되고 있는 실정이다.

본 논문에서는 이러한 문제를 크게, 단일 단백질 내부에서 도메인 조합의 협력 정도를 계산하는 것과 이를 바탕으로 도메인 조합 쌍이 가지는 상호작용 기여 정도를 반영하는 것으로 구분하여 접근한다.

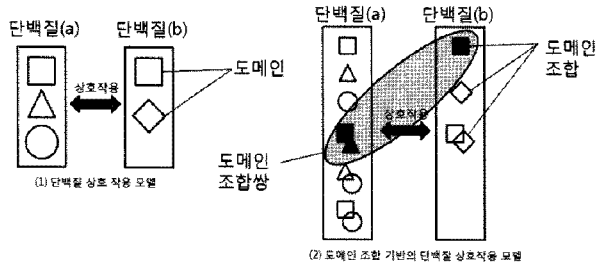


그림 1 단백질(a)와 단백질(b)의 상호작용 모델(1)과, 도메인 조합의 개념을 도입한 확장 모델(2)

다시 말해, 그림 1에서와 같이, 하나의 단백질 내에서 생성 가능한 여러 도메인 조합 각각에 대하여 기능적 협력 정도를 계산하고, 이를 통하여 단백질 상호작용 상황에서, 두 단백질이 가지는 모든 도메인 조합쌍의 상호작용 기여 정도를 구분하여 계산한다. 이는 계산학적 관점에서, 특정 feature들을 가지는 두 object의 상호작용이 있는 경우, feature들 간의 연관성과 이를 바탕으로 두 Object의 상호작용 가능성을 수치화 하는 문제로 일반화 할 수 있다. 이러한 접근 방식은 기존 방법론들과 달리 실제 단백질 상호작용을 좀더 충실하게 반영하는 강점을 가질 수 있으며, 상호작용에 있어 가장 크게 기여한 도메인 조합 쌍을 알아내는 등의 부가적 정보를 제공할 수 있는 효과가 있다. 또한, 남녀간의 결혼, 여러 증상을 가지는 환자와 질병의 관계등과 같은 여러 feature들을 가지는 두 object의 일반적 상호작용 예측으로 확장하여 적용할 수 있다.

본 논문에서 제안하는 예측 모델을 설명하기 전에 도메인 조합(Domain Combination)과 도메인 조합 쌍(Domain Combination Pair)의 개념을 설명한다. 어떠한 단백질 p가 한 개 이상의 도메인을 가지고 있다면, 이들 도메인 집합의 멱집합을 구함으로써 해당 단백질의 도메인 조합을 생성할 수 있다. 여기서 도메인 조합은 적어도 하나 이상의 도메인을 반드시 포함하는 것으로 한다. 즉, 단백질 p가 가지는 도메인 조합의 집합 DC_p는 다음과 같이 정의된다.

$$DC_p = \{dc \mid dc \subset PowerSet(domain(p))\}. \quad (1)$$

단, $domain(p)$는 단백질 p의 도메인 집합, $PowerSet()$은 도메인 집합의 멱집합에서 공집합을 제외한 집합을 생성하는 함수이다. 위의 (1)에서 공집합이 제거되므로 단백질이 n개의 서로 다른 도메인을 가지고 있다면 2^n-1개의 도메인 조합이 얻어진다. 두 단백질 p, q에서 가능한 모든 도메인 조합 쌍 집합의 정의는 (2)와 같다.

$$dc\_pair(p,q) = \{< dc_i, dc_j > | < dc_i, dc_j > \in DC_p \times DC_q \; or \; DC_q \times DC_p\},$$
$$where \; dc_i, dc_j \in DC_p \; or \; DC_q. \quad (2)$$

두 단백질 p, q가 각각 n, m개의 다른 도메인을 가지고 있을 경우, 기존의 단일 도메인 기반에서는 $n \times m$개의 도메인 쌍을 고려하는데 반해 도메인 조합 쌍 기반 방식에서는 $(2^n-1)\times(2^m-1)$개의 도메인 조합 쌍이 얻어진다.

3. 상호작용 중요도 행렬(Interaction Significance, IS Matrix) 구축

본 연구에서는 기존의 도메인 조합 쌍 모델을 바탕으로, 기존에 축적된 상호작용을 학습하여 도메인 조합 쌍의 힘(Domain Combination Pair Power, DCPPW)을 계산한다. DCPPW는 하나의 단백질 상호작용에서 특정 도메인 조합 쌍이 상호작용을 주관할 확률을 수치화 시킨 값이다. 이를 계산하기 위해서 단일 단백질 내에서 도메인간의 협업 정도를 고려하며, 도메인 조합 쌍들이 상호작용에 미치는 힘의 차이를 반영한 Weighted 도메인 조합 쌍을 정의한다. 도메인 조합의 협업 정도는 DCPPW 계산의 weight로 쓰이게 된다.

3.1 Weighted 도메인 조합 쌍(Weighted Domain Combination Pair)

단일 도메인이 기능을 위한 집약체인 것과는 달리, 도메인 조합은 항상 하나의 기능을 위해 협업한다고 볼 수 없다. 따라서 상호작용의 주체로서의 도메인 조합은 단일 도메인과 그 weight를 달리 해야 할 것이다. weight는 도메인 조합이 어느 정도 기능을 위해 협업하는지가 고려되어야 하며, 본 논문에서는 도메인 조합의 보존 정도를 측정하여 weight로 적용하였다. 단백질이 특정 기

능을 위해 발전하여 왔음을 고려할 때, 하위 기능 구조체인 도메인은 그 기능을 위해 협업하거나 영향을 주는 도메인과 함께 팀을 이루어 단백질을 형성하였을 것이다. 이러한 대전제 하에, 단백질의 발전에서 단지 서열 및 단일 도메인만이 보존되는 것이 아니라 기능을 위한 도메인 조합 또한 보존되어 왔을 가능성이 크다고 할 수 있으며, 보존이 잘된 도메인 조합은 어떠한 방식으로든 기능적 협업을 한다고 볼 수 있다. 본 논문에서는 보존 정도를 계산하기 위해 연관성 규칙의 all-confidence를 사용함으로써 조합 내 도메인 간의 상호 의존성을 측정하였다. all-confidence는 전체 단백질들 사이에서 해당 도메인 조합이 함께 나타날 확률을 측정한 것으로, Jung[10]의 연구에 따르면 도메인 조합의 분자 기능 협업도와 all-confidence는 상관관계를 가지는 것을 알 수 있다.

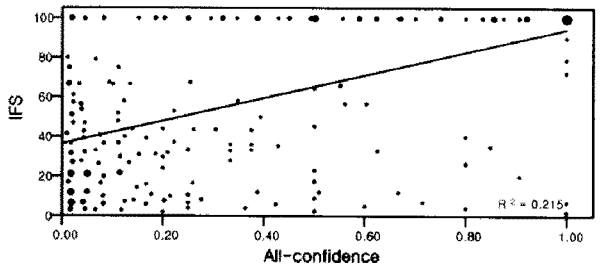


그림 2 도메인 조합의 all-confidence와 Inner Function Similarity(IFS)의 상관관계 분포

그림 2는 도메인 조합의 all-confidence와 분자 기능의 협업도의 상관관계를 분포도로 나타내고 있다. IFS는 Inner Functional Similarity로 대상 도메인 조합의 구성원이 가지는 Gene Ontology(GO) term을 분석하여, 구성원들끼리 유사한 기능을 가지는 정도를 나타내는 지표로서 사용되었으며, 각 점의 크기는 해당 도메인 조합의 수를 나타낸다. 이 분포의 피어슨 상관관계를 분석한 결과 유효수준 0.01에서 0.464의 상관관계를 가지는 것으로 나타났다. 이는 all-confidence가 높을수록 조합 내 도메인들이 유사하거나 같은 분자기능을 가진다는 것을 의미하며, all-confidence는 도메인 조합의 협업 정도의 측정에 적당함을 나타낸다. all-confidence는 다음과 같은 식으로 계산된다.

$$all\_conf(dc) = \frac{|\{p \mid p \in P \wedge dc \in PowerSet(domain(p))\}|}{MAX(\{i \mid \forall l(l \in PowerSet(dc) \wedge i = |\{q \mid q \in P \wedge l \in PowerSet(domain(q))\}|\})}. \quad (3)$$

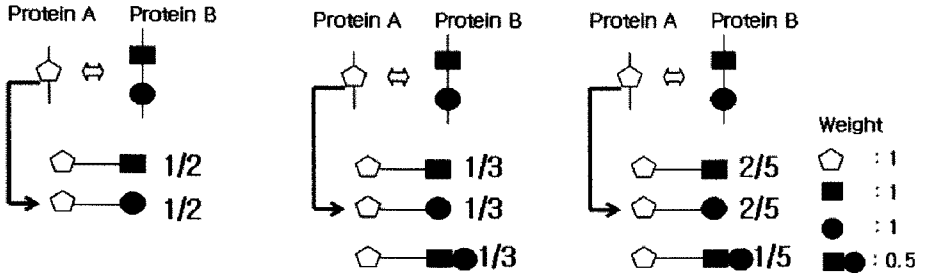
단 식에서 P는 전체 단백질 집합을 의미하며 $MAX()$는 양의 정수 집합 중 가장 큰 수를 구하는 함수이다. 즉 all-confidence는 해당 도메인 조합을 가지는 단백질의 수를, 조합의 부분 조합을 가지는 단백질의 수 중 가장 큰 수로 나눈 값이 된다. 이러한 계산을 통하여 조합을 이루는 도메인들이 전체 도메인 집합 중 얼마나 자주 함께 출현하였는지를 알아낼 수 있다. 각 도메인 조합의 weight가 마련되면, 이를 적용하여 하나의 단백질 상호작용에서 각 도메인 조합 쌍이 상호작용의 주체일 잠정적 확률을 계산한다. 하나의 상호작용 <p, q>에서 도메인 조합 dc_{ij}가 가지는 확률 식은 다음과 같다.

$$WDCP < p,q >_{i,j} = \frac{all\_conf(dc_i)}{\sum_{\forall dc_u \in DC_p} all\_conf(dc_u)} \times \frac{all\_conf(dc_j)}{\sum_{\forall dc_v \in DC_q} all\_conf(dc_v)}. \quad (4)$$

즉 도메인 조합의 단백질 내 weight는 해당 조합의 all-confidence를 전체 조합의 all-confidence의 합으로 나눈 값이 되며, 단백질 상호작용에서 조합 쌍이 상호작용의 주체일 확률은 각 조합의 단백질 내 weight의 곱이 된다. 이는 하나의 상호작용에서 상호작용이 일어났음을 확률 값 1로 보고, 이를 상호작용에서 발견되는 모든 도메인 조합 쌍이 나누어 가지는 개념이다. 그림 3은 기존의 단일 도메인을 사용한 모델과, 도메인 조합 기반 모델, 그리고 weighted 도메인 조합 기반 모델을 설명하고 있다. (a)의 기존의 단일 도메인 기반 모델에서는 상호작용 단백질 쌍에서 도메인 쌍이 두 개가 발견되므로, 각 도메인 쌍이 해당 상호작용의 주체일 확률 1을 경우의 수로 나누어 1/2이라 계산한다. (b)의 도메인 조합 기반 모델에서는 도메인 조합 쌍이 세 개이므로 각 1/3의 확률을 가진다. 여기에 각 도메인 조합의 weight를 적용 시키면, 그림 (c)와 같이 조합 쌍이 상호작용의 주체일 확률이 각각 다르게 계산된다. 단일 도메인의 weight는 항상 1일 수밖에 없으므로, 두 개 이상의 도메인을 가지는 조합의 weight는 도메인 한 개를 가지는 조합의 값보다 항상 작거나 같다. 이는 PDB에서 발견되는 복수개의 도메인으로 이루어지는 상호작용 쌍이 많지 않음을 볼 때, 타당하다고 볼 수 있다.

3.2 도메인 조합 쌍의 힘(Domain Combination Pair Power, DCPPW)

$WDCP < p,q >_{i,j}$는 학습 데이터에서 발견되는 단일 단백질 상호작용 쌍에 대하여 상호작용 확률을 1로 보고, 이를 각 도메인 조합 쌍이 단순히 나누어 갖는 값으로, 해당 도메인 조합 쌍이 실제로 상호작용을 주관할 확률 값이라고 볼 수 없다. 예를 들어 단백질 A, B, C, D가 구성하는 상호작용 쌍 집합 {<A, D>, <B, D>, <C, D>}가 있다고 가정하자. 각각의 도메인이 $domain(A)$ =



(a) 단일 도메인 기반 (b) 도메인 조합 기반 (c) Weighted 도메인 조합 기반

그림 3 단백질 상호작용에서 도메인 및 도메인 조합이 상호작용의 주체일 확률

$\{a,b\}domain(B) = \{a,c\}domain(C) = \{a,d\}domain(D) = \{e\}$ 로 구성되어 있고, 편의상 모든 weight값들을 1이라고 할 때, $WDCP < p,q >_{i,j}$ 만을 이용할 경우 도메인 조합 쌍 <{a}, {e}>와 <{b}, {e}>가 가지는 상호작용 주관 확률은 1/3로 같게 된다. 하지만 실제로는 도메인 조합 쌍 <{a}, {e}>의 상호작용 주관 확률이 <{b}, {e}>, <{c}, {e}>, <{d}, {e}>의 상호작용 주관 확률보다 높아야 함을 직관적으로 추정할 수 있는데, 그 이유는 도메인 조합 쌍 <{a}, {e}>의 출현 빈도가 다른 도메인 조합 쌍들에 비해 크기 때문이다. 전체 단백질 상호작용 데이터를 참조하여야만 도메인 조합 쌍이 상호작용을 주체할 확률 값 DCPPW를 계산 할 수 있다. 전체 데이터를 참조할 경우, 하나의 도메인 조합 쌍 $<dc_i, dc_j>$ 는 그 출현 횟수가 많을수록 단백질 상호작용에 미치는 영향이 크다고 할 수 있으며, 따라서 $<dc_i, dc_j>$ 의 힘은 ($WDCP < p,q >_{i,j}$ × 출현빈도)라고 볼 수 있다. 이때 $<dc_i, dc_j>$ 의 DCPPW는 식 (5)와 같이 계산된다.

$$DCPPW < p,q >_{i,j} = \frac{< dc_i, dc_j > \text{의 힘}}{< dc_i, dc_j > \text{의 힘} + < dc_i, dc_j > \text{가 아닌 } dc\_pair(p,q) \text{의 힘}}. \quad (5)$$

식 (5)는 특정 상호작용 안에서 도메인 조합 쌍 $<dc_i, dc_j>$ 이 상호작용의 주체일 확률은 그 조합의 힘을 해당 상호작용에서 발견되는 다른 도메인 조합 쌍의 힘과 상대적인 비율로 측정함을 의미한다. 즉 전체 데이터를 참조하므로 하나의 상호작용에서의 $<dc_i, dc_j>$ 가 상호작용 주체일 확률을 통계적으로 구할 수 있으며, 이를 공식화하면 식 (6)과 같이 표현할 수 있다.

$$DCPPW < p,q >_{i,j} = \frac{WDCP < p,q >_{i,j} \times |I\_pair(dc_i, dc_j)|}{\sum_{<dc_u, dc_v> \in dc\_pair(p,q)} WDCP < p,q >_{u,v} \times |I\_pair(dc_u, dc_v)|}. \quad (6)$$

단 $I_pair(dc_u, dc_v)$는 도메인 조합 쌍 $<dc_i, dc_j>$ 를 가지는 모든 단백질 상호작용 쌍을 말하며 다음과 같이 정의된다.

$$I\_pair(dc_i, dc_j) = \{< p_u, q_v > | < dc_i, dc_j > \in dc\_pair(p_u, q_v)\}. \quad (7)$$

3.3 상호작용 중요도 행렬(Interaction Significance, IS Matrix) 구축 및 단백질 상호작용 예측

IS matrix는 전체 상호작용 데이터에서 학습된 도메인 조합 쌍의 최종적인 상호작용 기여도를 저장한다. matrix의 행과 열은 전체 단백질에서 발견되는 도메인 조합이며, symmetric matrix이다. 각 원소는 상호작용 데이터에서 도메인 조합 쌍이 발견될 때에 그 쌍의 기여도를 저장하므로 또한 sparse matrix의 형태를 띤다. 도메인 조합 쌍 $<dc_i, dc_j>$ 에 해당하는 IS matrix의 원소 값 $IS <dc_i, dc_j>$ 은 다음의 식으로 정의된다.

$$IS(dc_i, dc_j) = \frac{\sum_{\forall (p_u, p_q) \in I\_pair(dc_i, dc_j)} DCPPW < p_u, q_v >_{i,j}}{|I\_pair(dc_i, dc_j)|}. \quad (8)$$

즉, $IS <dc_i, dc_j>$ 는 $<dc_i, dc_j>$ 가 발견되는 모든 단백질 상호작용 쌍 $I\_pair\,(dc_i, dc_j)$ 에서 각각 구한 $DCPPW < p_u, q_v >_{i,j}$ 값들의 평균값이다.

미지의 단백질 쌍의 상호작용 확률은 각 단백질의 도메인 쌍 또는 도메인 조합 쌍들의 상호작용 여부와 밀접한 관계를 가지고 있다. 본 예측 시스템에서는 단백질 상호작용이 가지는 여러 도메인 조합 쌍들 간의 상호작용들 중에 적어도 하나의 유효한 상호작용이 존재 할 경우 단백질 상호작용이 일어난다고 보고 예측을 진행하였다. 기존 예측 시스템의 AP matrix[13,14]는 각 원소들의 전체 합이 1로 두 개의 행렬(상호작용 하는 단백질 쌍과 상호작용하지 않는 단백질 쌍에 대한 행렬)을 서로 비교하여 상호작용 확률을 예측하였다. 본 연구에서 새롭게 제안하는 상호작용 중요도 행렬의 각 원소 값들은 해당 도메인 조합 쌍이 단백질 상호작용 사이에서 가질 수 있는 힘을 반영한 출현 확률을 가지고 있다. 앞선 과정을 통해 얻어진 상호작용 중요도 행렬을 기반으로 미지의 단백질 쌍 <p, q>에 대한 확률 예측 식은,

$$IP(p,q) = 1 - \prod_{(dc_i, dc_j) \in dc\_pair(p,q)} (1 - IS(dc_i, dc_j)) \quad (9)$$

로 정의할 수 있다. 여기서 $IP(p, q)$는 단백질 p, q가 상호작용 할 확률이며, 이것은 단백질 p, q의 도메인 조합 쌍들 중 적어도 하나의 상호작용이 일어날 확률과 같다. 상호작용 중요도 행렬의 각 원소 값들은 도메인 조합 쌍의 상호작용이 일어날 확률이므로 여러 도메인 조합 쌍 중 적어도 하나의 상호작용이 일어날 확률은 1에서 모든 도메인 조합 쌍의 상호작용이 일어나지 않을 확률을 빼는 것과 같다. 예를 들어 단백질 p, q 쌍이 가지는 도메인 조합 쌍 각각의 확률이 X_1, X_2, X_3 라고 하면, 모든 도메인 조합 쌍의 상호작용이 일어나지 않을 확률은 $(1-X_1)(1-X_2)(1-X_3)$이다. 그러므로 적어도 하나의 도메인 조합 상호작용이 일어날 확률 $IP(p,q)$ 의 값은 $1-(1-X_1)(1-X_2)(1-X_3)$가 된다.

4. 검 증

검증을 위해 단백질 상호작용 데이터는 DIP과 IntAct를 사용하였으며 UniProt ID를 통하여 통합하였다. 각 단백질의 도메인 정보는 해당하는 UniProt ID를 통하여 Pfam-A에서 추출하였다. 본 논문에 사용된 DIP의 데이터는 2008년 10월 14일 릴리즈 버전으로 총 단백질 수 20,442개를 포함한 57,330의 상호작용을 가지고 있으며, IntAct는 2008년 11월 17일 릴리즈 버전으로 55,036개의 단백질로 이루어진 115,311개의 상호작용을 포함하고 있다. 두 데이터베이스 사이에는 총 9,764개의 중복된 상호작용이 발견되어 최종 통합된 단백질 상호작용은 총 162,877 쌍이다. 이 가운데, S. cerevisiae 종의 65,902개의 상호작용을 대상으로 도메인 보유여부를 살펴본 결과 최종적으로 45,385개의 사용 가능한 단백질 쌍을 추려내었다.

공개된 단백질 상호작용 DB에서 UniProt ID를 통하여 통합이 가능하며, 도메인 정보가 알려진 상호작용 쌍만을 추려낸 결과 약 30% 정도만이 사용 가능하였다. 최종 선별된 단백질 상호작용들이 제안하는 방법과 같은 도메인 기반의 예측 방식에서 어느 정도 유효성을 가지는지 알아보기 위하여, 최종 선별된 단백질 상호작용을 10%씩 추출하면서 새롭게 추가되는 도메인 패턴의 개수를 측정하였다. 그 결과 그림 4와 같이, 약 70% 정도(PPI 개수 32000여개)부터 새롭게 추가되는 도메인의 개수는 현저하게 감소하는 추세를 보였다. 이로 미루어 볼 때, 우리가 사용한 단백질 상호작용 데이터는 상호작용 도메인 패턴을 상당부분 포함하고 있는 것으로 판단된다. 그러나, 실험에 사용하는 상호작용 데이터가 많아 진다면 상호작용 패턴의 출현 빈도 등에 있어 좀 더 명확한 차이를 찾을 수 있어 예측 결과의 향상을 기대할 수 있을 것이다. 좀더 많은 수의 데이터를 사용하는 것은 단백질의 유사성 비교 혹은 서열상의 유사 도메인 추출등과 같은 방법으로 개선할 수 있으나 본 논문에서는 이를 향후 과제로 남긴다.

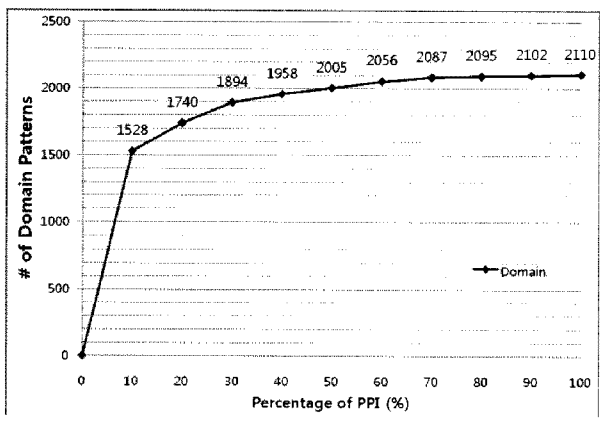


그림 4 단백질 쌍의 축적에 따른 서로 다른 도메인 패턴의 개수 변화 추세 그래프

단백질 상호작용 예측 결과의 검증을 위해 전체 단백질 상호작용 쌍들 중 80%를 학습 집단으로 사용하였고, 나머지 20%를 상호작용하는 단백질 쌍의 검증 집단으로 사용하였다. 또한 상호작용하는 단백질 쌍의 검증 집단과 동일한 수의 무작위로 생성된 단백질 쌍의 집단을 상호작용하지 않는 집단의 검증을 위해 사용하였다. 검증 결과는 예측결과를 통계적으로 측정할 수 있는 대표적인 지표인 Sensitivity와 Specificity를 통해 나타낸다. Sensitivity는 실제로 상호작용하는 데이터를 사용해 예측하였을 경우 시스템이 얼마나 Positive의 결과를 보이는 지를 평가하며, 다음과 같은 식으로 나타낼 수 있다.

$$sensitivity = \frac{number\ of\ True\ Positives}{number\ of\ True\ Positives + number\ of\ False\ Negatives}.$$

Specificity 는 상호작용하지 않는 데이터를 예측할 경우 예측 시스템이 얼마나 Negative의 결과를 보이는 지를 측정하며, 다음과 같은 식으로 나타낼 수 있다.

$$specificity = \frac{number\ of\ True\ Negatives}{number\ of\ True\ Negatives + number\ of\ False\ Positives}.$$

검증 시에는 상호작용 확률 IP 값의 threshold를 0에서 1사이로 0.1씩 증가 시키면서 정확도를 분리하여 측정하였다. 이 때, threshold 보다 높은 확률 값을 가지면 해당 단백질 쌍은 상호작용이 존재한다고 예측하며, 낮은 확률 값을 가질 경우 상호작용이 존재하지 않는다고 예측한다. 표 1은 본 논문에서 제안하는 상호작용 중요도 행렬을 이용한 예측 정확도를 보여준다.

표 1에서 sensitivity는 최고 63.42%에서 threshold의 증가에 따라 크게 하강하는 경향을 보였으며, specificity는 threshold의 증가에 따라 다소 상승하기는 하였

표 1 상호작용 중요도 행렬을 이용한 *S. cerevisiae*의 예측 정확도

Threshold	Sensitivity (%)	Specificity (%)
0.0 초과	63.42	94.38
0.1 이상	62.64	94.92
0.2 이상	61.86	95.32
0.3 이상	60.00	95.78
0.4 이상	57.31	96.89
0.5 이상	54.40	97.40
0.6 이상	43.85	98.10
0.7 이상	30.17	98.66
0.8 이상	25.11	98.95
0.9 이상	18.87	99.12

으나 전반적으로 94%에서 99% 사이의 안정된 정확성을 나타냈다. Threshold가 0.9인 경우, IP값이 0.9 이상인 것들만 상호작용 한다고 판단하기 때문에 대부분의 비 상호작용 단백질 쌍은 상호작용이 없다고 판단되어 specificity가 증가하고, 0.9 아래에 위치한 상호작용 단백질 쌍은 상호작용이 없는 것으로 잘못 판단되기 때문에 sensitivity는 떨어지고 있는 것을 볼 수 있다.

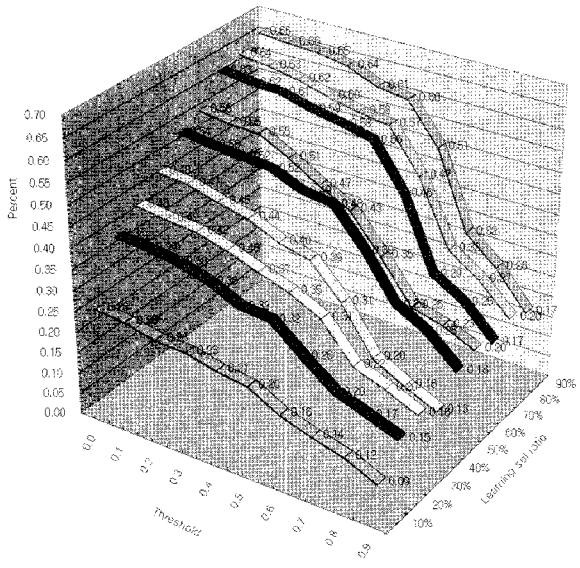
한편, 본 논문의 상호작용 예측 시스템은 학습을 통하여 예측을 수행하는 방식으로, 학습집단의 축적은 예측 정확도의 향상을 기대할 수 있다. 이를 확인하기 위해 고정된 수의 테스트 집단(300개의 상호작용 및 비 상호작용 단백질 쌍)을 준비하고 학습 집단의 크기를 전체 상호작용 쌍의 10%에서 90%까지 변화시키며 예측 정확도의 추이를 검증한 결과 표 2와 같은 예측 정확도의 변화를 확인하였다. 학습 집단의 크기가 전체 단백질 쌍의 10%일 경우, sensitivity는 약 26%에서 학습집단의 크기가 늘어남에 따라 급격히 상승하였으며, 70%를 기점으로 점차 완만하게 증가하며 90% 크기의 학습집단에서 약 65%의 정확도를 보였다. 이에 반해 Specificity는 학습 집단의 사이즈 변화에 크게 관계없이 약 94% 이상의 정확도가 전체 threshold 영역에서 나타남을 확인하였다.

5. 고 찰

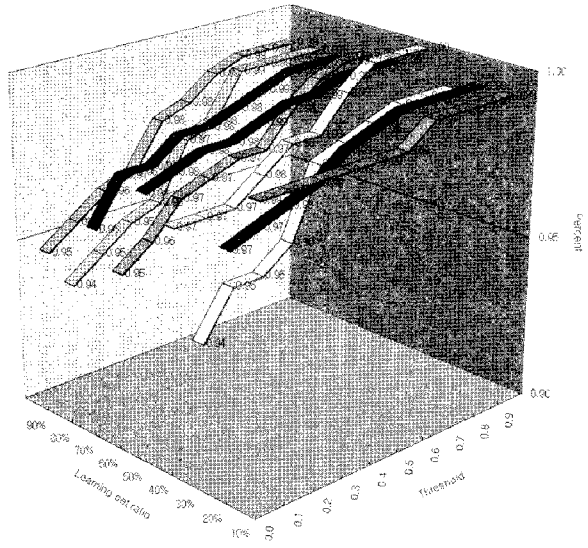
Han's 그룹의 기존 예측 방법[13,14]에서는 테스트 집단의 도메인이 AP matrix에 존재하는 경우(Hit)만 예측 결과를 제시하고 있어 본 시스템과의 직접적인 비교는 무리가 있다. 이에, 기존 방법에서 Hit하지 않는 경우의 sensitivity와 specificity를 50%로 가정하고 학습집단의 상호작용 단백질 쌍과 비 상호작용 단백질 쌍의 ratio중 가장 높은 예측 정확도를 보이는 ratio의 결과로 비교를 수행하였다. AP matrix 기반의 단백질 상호작용 예측은 ratio 10.0에서 평균 40.48%의 hit을 보였으므로, 이 때의 sensitivity 78.73%, specificity 95%를 비교대상으로 한다. hit하지 않은 59.52%의 예측 정확도를 50%의 랜덤 수준으로 가정하면, sensitivity와 specificity는 각각 약 61%, 68%가 계산된다. 반면, 본 논문에서 제안하는 방법을 적용한 결과 sensitivity는 약 63%, specificity는 약 94%로 각각 2%, 26% 상승한

표 2 학습 집단의 크기 변화에 따른 threshold 구간별 예측 정확도 검증 결과

Threshold		10%	20%	30%	40%	50%	60%	70%	80%	90%
0.0	sensitivity	0.26	0.39	0.43	0.48	0.54	0.56	0.63	0.64	0.66
	specificity	0.99	0.97	0.94	0.97	0.98	0.95	0.96	0.94	0.95
0.1	sensitivity	0.25	0.39	0.42	0.47	0.53	0.55	0.62	0.63	0.66
	specificity	0.99	0.97	0.95	0.97	0.98	0.96	0.97	0.95	0.95
0.2	sensitivity	0.23	0.38	0.42	0.45	0.52	0.55	0.61	0.62	0.65
	specificity	0.99	0.98	0.95	0.97	0.98	0.97	0.97	0.95	0.96
0.3	sensitivity	0.23	0.36	0.40	0.44	0.52	0.51	0.59	0.60	0.64
	specificity	0.99	0.98	0.96	0.98	0.98	0.97	0.98	0.96	0.96
0.4	sensitivity	0.21	0.32	0.37	0.40	0.49	0.47	0.58	0.58	0.61
	specificity	0.99	0.99	0.98	0.98	0.99	0.97	0.98	0.97	0.98
0.5	sensitivity	0.20	0.32	0.35	0.39	0.48	0.43	0.56	0.57	0.60
	specificity	0.99	0.99	0.98	0.98	0.99	0.97	0.98	0.98	0.98
0.6	sensitivity	0.16	0.26	0.31	0.31	0.39	0.35	0.46	0.47	0.51
	specificity	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99
0.7	sensitivity	0.14	0.20	0.22	0.20	0.29	0.25	0.29	0.32	0.32
	specificity	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99
0.8	sensitivity	0.12	0.17	0.20	0.16	0.25	0.23	0.25	0.27	0.26
	specificity	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99
0.9	sensitivity	0.09	0.15	0.16	0.13	0.18	0.20	0.17	0.20	0.17
	specificity	0.99	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99



(a) Sensitivity



(b) Specificity

그림 5 300개의 테스트 단백질 쌍에 대한 학습 집단 크기에 따른 상호 작용 예측 정확도 그래프

예측 정확도를 보여준다. 이는, 단백질 각각이 포함하고 있는 도메인들 간의 협업 확률 고려와 단백질 쌍에 대하여 도메인 조합들이 상호작용에 미치는 영향을 차등하여 계산한 것이 주요한 것으로 보인다.

검증 결과에서는, 실험으로 밝혀진 단백질 상호작용 쌍들이 포함하는 도메인 조합은 보존된 일정한 패턴을 가지고 있음을 알 수 있다. 테스트로 사용한 비 상호작용 쌍은, 단백질을 임의로 쌍을 만들어 사용한 것으로, 학습집단의 도메인 패턴과 거의 겹치지 않아 낮은 IP 값을 보여주게 된다. 반면, 단백질 상호작용 쌍은 학습 집단에서 포함하는 도메인 조합 패턴이 테스트 집단에서도 재 발견될 확률이 단백질 쌍을 임의로 조합한 비 상호작용 집단보다 높다. 그림 5에서 sensitivity는 학습 집단의 크기 증가에 따라 급격히 상승하고 있으며, 이는 곧 학습 집단의 크기가 증가함에 따라 테스트 집단과 겹치는 도메인 패턴이 많아지며, 이러한 패턴은 임의로 생성한 단백질 쌍에서는 거의 발견되지 않음을 암시한다. 또한, sensitivity의 상승 곡선이 완만해 지는 것은, 학습 집단의 크기 증가가 더 이상 많은 수의 새로운 도메인 조합을 포함하지 않는 다는 것을 의미한다. 반면, 임의로 조합한 비 상호작용 집단의 도메인 패턴은 매우 다양하게 나타날 수 있으며, 학습 집단의 증가가 테스트 집단이 가지고 있는 도메인 패턴을 충분히 포함할 만큼의 영향을 끼치지 못하고 있다.

학습 집단의 증가에 따라 sensitivity가 상승하지만 65%에 그치고 있는 것은 몇 가지로 해석할 수 있다. 첫째, 학습 집단 자체에 포함되어 있는 오류이다. 본 연구에서 사용한 단백질 상호작용 데이터는 High-throughput 방식을 사용한 것이며, 일반적으로 많은 오류가 포함되어 있다고 알려져 있다. 둘째, 학습 집단 크기의 부족이다. 현재 밝혀진 단백질 상호작용 쌍이 포함하고 있는 도메인이 충분하지 않을 수 있으며, 좀더 많은 수의 상호작용 집단이 추가 된다면 예측 정확도의 향상이 이루어 질 수 있다. 4장의 검증 결과를 통해 학습 집단의 크기가 증가 할수록 예측 정확도는 향상됨을 보였으며, 이로 미루어 볼 때, 더 이상 sensitivity가 증가 하지 않는 정도의 학습 집단 크기가 가장 이상적이라고 할 수 있다. 마지막으로, feature의 부족을 들 수 있다. 단백질 상호작용에는 다양한 요소가 작용할 수 있으며, 도메인만을 고려한 본 논문과 같은 방식이 어느 정도 한계를 가질 수 있다. 그러나, 정밀한 실험을 수행하는 데에는 많은 시간적, 비용적 노력이 필요하며, 프로그램 상에서 쉽게 수행할 수 있는 방식은 실제 실험의 대상을 크게 줄여 준다는 점에서 전처리 과정으로 의미를 가지게 된다. 또한, 제안하는 방식의 예측 결과를 살펴 보았을 때, 도메인 조합은 단백질 상호작용에서 의미 있는 역할을 하고 있음을 강하게 암시하고 있으며, 기타 다른 feature들을 추가한 새로운 방식의 기초로 활용될 수 있다. 실제로 S. cerevisiae 의 경우, 도메인은 약 2000개 정도가 발견되며, 그 중 두 개만을 선택하는 단일 도메인 상호작용의 경우만 하더라도 $_{2000}C_2 \approx 2{,}000{,}000$ 정도의 조합 생성이 가능하다. 이는 단지 단일 도메인 조합의 경우이며, 10개 정도의 도메인을 선택하는 경우는 무한에 가깝다. 이런 상황에서 실험적으로 밝혀진 단백질 상호작용 쌍 약 35,000 개 사이에서 도메인 조합 쌍이 중복되는 것은 단백질 상호작용에 있어 도메인 조합에 일정한 패턴이 나타남을 강하게 암시한다.

이러한 관점에서, 0이 아닌 값을 가지는 모든 단백질

쌍은 실험으로 밝혀진 단백질 쌍에서 나타난 도메인 패턴을 최소한 하나 이상 가지고 있음을 의미하고 있다. 따라서 본 논문에서 계산하고 있는 IP 값이 정확히 단백질 상호작용 확률을 나타내는 가에 대해서는 향후 추가적인 검증이 필요하다. 이는, 단백질 쌍에서, 높은 weight를 부여한 도메인 조합이 실제 단백질 쌍에서 상호작용 역할을 담당했는지를 확인하고 그 결과를 IP value의 크기에 따라 분석함으로써 가능하다.

본 논문에서 개발한 내용은 단일 어플리케이션 형태로 구현되어 웹을 통해 공개되어 있다(http://code.google.com/p/prespi/downloads/list). 프로젝트 사이트에서는 소스 코드와 학습에 사용된 DB 및 사용 설명서가 함께 제공되고 있으며, GNU의 Lesser General Public License (LGPL)을 따른다. 그림 6은 정상적으로 학습을 마친 후 예측을 수행한 경우 생성된 출력 파일의 예시이다.

출력은 해당 단백질의 Pfam-A도메인을 보여주고, 각 도메인 조합 쌍의 weight 및 DCPPW의 계산 결과가 포함되어 있으며 최종적으로 단백질 상호작용 확률을 기재한다.

6. 결 론

본 논문에서는 기존의 도메인 조합 기반 출현 확률 행렬 기반의 단백질 상호작용 기법을 수정 보완한 상호작용 중요도 행렬 기반의 예측 방식을 고안하였다. 출현 확률 행렬 기반의 예측에서는 단백질 상호작용 쌍이 생성할 수 있는 도메인 조합 쌍의 중요도가 동일하다고 가정하였다. 반면, 새롭게 제안하는 상호작용 중요도 행렬에서는 도메인 조합 쌍이 상호작용에서 실제로 영향을 미칠 확률을 차등 계산하도록 하였다. 이를 위하여, 도메인들이 진화과정에서 서로 특정한 기능을 위하여 협업할 확률을 계산하고, 이를 바탕으로 하나의 단백질 쌍 내에서 생성할 수 있는 도메인 조합 쌍 각각의 중요도를 수치화 하여 상호작용 중요도 행렬을 완성하였다.

새롭게 구성된 상호작용 중요도 행렬을 적용한 결과, 기존의 시스템과 비교하여 예측 정확도가 향상됨을 발견할 수 있었다. 또한, 학습집단의 축적에 따라 점차적으로 예측 정확도가 상승함을 보임으로써, 향후 실험데이터의 증가에 따른 예측 시스템의 꾸준한 성능향상 가능성을 확인하였다.

한편, 기존의 방식에서는 단백질 상호작용 쌍뿐만 아니라, 비 상호작용 쌍에 대해서도 출현 확률 행렬을 생성하여 학습에 사용하였다. 그러나, 본 연구에서는 도메인 조합 쌍이 생성될 수 있는 경우의 수가 매우 크다는 점과, 이에 반해 실험적으로 밝혀진 단백질 상호작용 쌍의 개수가 비교적 적음에도 불구하고 중복되는 도메인 조합 쌍이 다수 발견됨에 주목하고, 학습에서 비 상호작용 쌍의 사용을 배제하였다.

예측 시스템에서는 0이 아닌 IP 값을 가지는 모든 단백질 쌍을 상호작용의 가능성이 있는 것으로 판별하고 있다. 그러나, IP 값의 차이가 실제 단백질 상호작용의 신뢰도를 가늠하는지에 대해서는 추가적인 검증이 필요하다. 이는 PDB와 같은 단백질 상호작용 시의 실제 도

```
==============================================================================================================
Domain of Q12222 : PF00069
Domain of P29509 : PF00070 PF07992 PF00070:PF07992
weight of PF00069: 1.0
weight of PF00070: 1.0
>dc_value of PF00069 and PF00070 : 0.27854494352969433
weight of PF00069: 1.0
weight of PF07992: 1.0
>dc_value of PF00069 and PF07992 : 0.27854494352969433
weight of PF00069: 1.0
weight of PF00070:PF07992: 0.835694074630737
>dc_value of PF00069 and PF00070:PF07992 : 0.23277835882611866
>Expression<
1- (1- 0.27854494352969433)(1- 0.27854494352969433)(1- 0.23277835882611866)
>>>IP value of (Q12222, P29509) is 0.600663131691052
==============================================================================================================
Domain of Q06708 : PF02985
Domain of P47135 : PF00076 PF00806 PF00806:PF00076
weight of PF02985: 1.0
weight of PF00076: 1.0
>dc_value of PF02985 and PF00076 : 0.9077029905272139
weight of PF02985: 1.0
weight of PF00806: 1.0
>dc_value of PF02985 and PF00806 : 0.47688624345427666
weight of PF02985: 1.0
weight of PF00806:PF00076: 0.00351493852213025
>dc_value of PF02985 and PF00806:PF00076 : 2.2789802085323532E-4
>Expression<
1- (1- 0.9077029905272139)(1- 0.47688624345427666)(1- 2.2789802085323532E-4)
>>>IP value of (Q06708, P47135) is 0.9519137199418065
==============================================================================================================
```

그림 6 PreSPI-beta-1.0 의 단백질 상호작용 예측 예시: 단백질의 도메인 조합 정보, 각 도메인 조합의 weight, 각 도메인 조합 쌍의 DCPPW, 최종 IP 계산 결과를 포함하고 있음

메인 바인딩 정보와, IP 값에 따른 단백질 쌍 내부의 도메인 조합 쌍 중요도 순위를 비교함으로써 검증이 가능하며, 이는 향후 과제로 남긴다.

참 고 문 헌

[1] Marcotte, E., Pellegrini, M., Ng, H., Rice, D., Yeates, T., Eisenberg, D., "Detecting protein function and protein-protein interactions from genome sequences," *Science*, 285, pp.751-753, 1999.

[2] Szilagyi A, Grimm V, Arakaki A, Skolnick J, "Prediction of physical protein-protein interactions," *Phys Biol.*, 2, S1-S16, 2005.

[3] Sprinzak, E., Margalit, H., "Correlated sequence-signatures as markers of protein-protein interaction," *J. Mol. Biol.*, 311, pp.681-692, 2001.

[4] Deng, M., Mehta, S., Sun, F., Chen, T., "Inferring domain-domain interactions from protein-protein interactions," *Genome Res.*, 12, pp.1540-1548, 2002.

[5] Chen, L., Wu, L., Y. W., Zhang, X., "Inferring protein interactions from experimental data by association probabilistic method," *Proteins*, 62, pp.833-837, 2006.

[6] Liu, Y., Liu, N., Zhao, H., "Inferring protein-protein interactions through high-throughput interaction data from diverse organisms," *Bioinformatics*, 21, pp.3279-3285, 2005.

[7] Dohkan, S., Koike, A., Takagi, T., "Support vector machines for predicting protein-protein interactions," *Genome Inform*, 14, pp.502-503, 2003.

[8] Riley, R., Lee, C., Sabatti, C., Eisenberg, D., "Inferring protein domain interactions from databases of interacting proteins," *Genome Biology*, 6,R89, 2005.

[9] Moza, B., Buonpane, R., Zhu, P., Herfst, C., Rahman, A., McCormick, J., Kranz, D., Sundberg, E., "Long-range cooperative binding effects in a T cell receptor variable domain," *Proc Natl Acad Sci*, 103, pp.9867-9872, 2006.

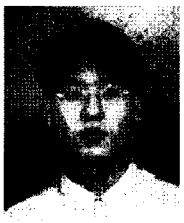
[10] S.H. Jung, H.Y. Hur, D. Kim, D.S. Han, "Identification of Conserved Domain Combinations in S. cerevisiae Proteins," *Bioinformatics and Bioengineering*, pp.14-20, 2007.

[11] J. Brodie and I. J. McEwan, "Intra-domain communication between the nterminal and DNA-binding domains of the androgen receptor: modulation of androgen response element DNA binding," *Journal of Molecular Endocrinology*, 34, pp.603-615, 2005.

[12] N. B. E. Ronne and K. Dano, "Domain interplay in the urokinase receptor," *J. Biol. Chem.*, 217(37), 22, pp.885-22 894, 1996.

[13] Han, D., Kim, H., Jang, W., Lee, S., Jung, S., "PreSPI: a domain combination based prediction system for protein-protein interaction," *Nucl Acids Res*, 32, pp.6312-6320, 2004.

[14] R.S. Wang, Y. Wang, L.Y. Wu, X.S. Zhang, L. Chen, "Analysis on multi-domain cooperation for predicting protein-protein interactions," *BMC Bioinformatics*, 8, pp.2-20, 2007.



장 우 혁

2003년 충남대학교 컴퓨터공학교육학과(학사). 2005년 한국정보통신대학교 공학부(석사). 2005년 2월~현재 한국과학기술원 정보통신공학과 박사과정

정 석 훈

2004년 한동대학교 전산학과(학사). 2007년 한국정보통신대학교 전산학과(석사) 2007년 2월~현재 한국과학기술원 정보통신공학과 박사과정

정 휘 성

2007년 한국정보통신대학교 전산학과(학사). 2007년 8월~현재 한국과학기술원 정보통신공학과 석사과정

현 보 라

2007년 한국정보통신대학교 전산학과(학사). 2009년 한국과학기술원 정보통신공학과(석사). 2009년 8월~현재 삼성전자㈜ 연구원

한 동 수

1989년 서울대학교 전산학과(학사). 1991년 서울대학교 전산학과(석사). 1996년 일본 교토대학교 정보공학(박사). 1992년 삼성전자(주) 연구원. 1996년 일본 NEC C&C 연구소 연구원. 1997년 현대정보기술(주) 정보기술연구소 책임연구원. 2009년 한국정보통신대학원 공학부 교수. 2009년 3월~현재 한국과학기술원 전산학과 교수