도메인 조합 기반 단백질 상호작용 가능성 순위 부여 기법

(Protein Interaction Possibility Ranking Method based on Domain Combination)

한 동 수 † 김 홍 숙 †† 장 우 혁 †† 이 성 독 †††
(Dong-Soo Han) (Hong-Song Kim) (Woo-Hyuk Jong) (Sung-Doke Lee)

요 약 인터넷 상에 단백질 및 관련 데이터의 축적에 따라, 도메인에 기반하여 단백질의 상호작용을 계산적으로 예측하는 많은 기법들이 제안되었다. 그러나, 대부분의 기법들이 예측에서 낮은 정확도와 복수개의 단백질 쌍에 대한 상호작용 가능성들 간에 순위 정보를 제공하지 못하는 등의 한계로 인하여 실무 적용에 한계를 가지고 있다. 본 논문에서는 도메인 조합 기반 단백질 상호작용 예측 기법을 재평가하고 상호작용하는 것으로 예측되는 복수개의 단백질 쌍들에서 이들의 상호작용 가능성들 간에 순위를 부여하는 방법을 제시한다. 순위 부여 방법은 도메인 조합에 기반한 단백질 상호작용 예측 방법의 틀 내에서 확률 식을 고안하여 제시한다. 제시된 순위 부여 기법을 사용함으로써, 상호작용을 하는 것으로 예측된 단백질 쌍들간에 상호작용 가능성이 좀 더 높은 것을 구별해 낼 수 있다. 또한 순위 부여 기법의 검증 과정에서 학습에 사용된 단백질 집단의 PIP(Primary Interaction Probability)값과 일치된 PIP값을 가지는 단백질 쌍 그룹의 경우에는, 상호작용 확률과 예측 정확도 사이에 상관관계가 존재함을 확인할 수 있었다.

키워드 : 단백질-단백질 상호작용, 도메인 조합, 도메인 조합 쌍, 예측 모델, 출현 확률 행렬, 단백질-단백질 상호작용 가능성 순위 부여 기법

Abstract With the accumulation of protein and its related data on the Internet, many domain based computational techniques to predict protein interactions have been developed. However, most of the techniques still have many limitations to be used in real fields. They usually suffer from a low accuracy problem in prediction and do not provide any interaction possibility ranking method for multiple protein pairs. In this paper, we reevaluate a domain combination based protein interaction prediction method and develop an interaction possibility ranking method for multiple protein pairs. Probability equations are devised and proposed in the framework of domain combination based protein interaction prediction method. Using the ranking method, one can discern which protein pair is more probable to interact with each other than other protein pairs in multiple protein pairs. In the validation of the ranking method, we revealed that there exist some correlations between the interacting probability and the precision of the prediction in case of the protein pair group having the matching PIP(Primary Interaction Probability) values in the interacting or non interacting PIP distributions.

Key words : Protein protein interaction, Domain combination, Domain combination pair, Prediction model, AP matrix, Protein-protein interaction possibility ranking method.

1. 서 론

계산을 통한 단백질 상호작용 예측의 장점이 인식되면서 많은 다양한 예측 기법이 제안되고 있다[1-6]. 특히, 최근에는 도메인 또는 도메인 조합에 기반한 단백질 상호작용 예측 기법이 활발하게 연구되고 있다[2,7-9]. 미지의 단백질 쌍에 대해서 생물학적인 실험을 통하지 않고 상호작용 가능성을 계산을 통해 예측해 주는 기법은 분명 생물학자에게 유용한 정보를 제공해 주는 것이

† 종신회원 : 한국정보통신대학교 공학부 교수
dshan@icu.ac.kr
†† 비 회 원 : 한국정보통신대학교 공학부
dshan@icu.ac.kr
torajim@icu.ac.kr
††† 정 회 원 : 한국정보통신대학교 공학부 교수
sdlee@icu.ac.kr
논문접수 : 2004년 8월 18일
심사완료 : 2005년 6월 17일

사실이지만 다음과 같은 점에서 한계를 가지고 있다. 계산을 통해서 예측하는 기법의 정확도가 아직도 생물학자들이 사용하기에는 예측의 정확도가 충분히 높지 않다는 점이다. 이것은 예측을 위한 학습에 필요한 정확한 실험 데이타의 부족에도 큰 원인이 있지만 기존의 도메인에 기반한 단백질 상호작용 예측 기법들이 가지고 있는 문제점에도 기인하는 측면이 있다[10-14]. 최근에는 도메인 조합(domain combination)을 기반으로 한 단백질 상호작용 예측 기법도 제안되었다[8,9]. 이 기법에서는 상호작용하는 단백질 쌍들에 대한 올바른 실험적 데이타를 가지고 도메인 조합의 개념을 새롭게 도입함으로써 예측 시스템의 정확도를 개선할 수 있다는 것이 확인되었다. 그러나 복수의 단백질 쌍에서 다시 어느 쌍이 더 상호작용할 가능성이 높다는 추가적인 순위 정보는 제공하지 않고 있다.

본 논문에서는 복수 개의 단백질 쌍에 대해서 도메인 조합을 기반으로 한 단백질-단백질 상호작용 가능성 순위 부여 기법을 제안한다. 이 기법은 [8,9]에서 제시된 도메인 조합 기반 단백질 상호작용 예측 기법을 기반으로 하여 개발되었다. 복수 개의 단백질 쌍에 대한 상호작용 가능성 순위는 본 논문에서 고안한 상호작용 확률식에 의해 계산된 상호작용 확률에 의해서 결정된다. 이 순위 부여 기법을 이용하여 생물학자들은 상호작용이 있는 것으로 예측된 단백질 쌍들 중에서 상호작용이 보다 많이 예상되는 것들을 식별할 수 있다.

상호작용 가능성 순위 부여 기법의 개발에 도메인 조합 기반 접근법 사용의 적절성을 확인하기 위하여, 먼저 도메인 조합 기반 단백질 상호작용 예측 기법의 정확도를 재검증하였다. 도메인 조합 기반 단백질 상호작용 예측 기법의 재검증은 효모(yeast)에서 상호작용이 있는 것으로 알려진 단백질 쌍 집합에 대하여 상호작용이 없는 것으로 알려진 단백질 쌍 집합의 크기를 1에서 20배의 비율로 변화시키면서 민감도(sensitivity)와 특이도(specificity)의 변화 거동을 분석하는 방식으로 수행되었다. 재 검증을 통하여 상호작용이 없는 것으로 추정되는 단백질 쌍 집합 비율의 크기가 증가 할수록 예측의 정확도도 개선되었다. 효모(yeast)에 대한 데이타를 사용한 학습 집합에서 상호작용이 없는 단백질 쌍 집합의 크기가 상호작용이 있는 단백질 쌍 집합보다 20배 큰 경우에 84%의 민감도와 75%의 특이도를 보여 주었다. 이것은 도메인 조합 기반 단백질 상호작용 예측 기법이 복수의 단백질 쌍의 상호작용 가능성 순위 부여 기법 개발에 적절함을 의미한다.

복수의 단백질 쌍에 대한 상호작용 가능성 순위 부여 기법의 유효성을 확인하기 위하여 복수개의 단백질 쌍 집합을 그들의 *PIP*(Primary Interaction Probability) 값에 따라 몇 개의 실험 그룹으로 분할한 후, 각각의 그룹에 대하여 민감도와 특이도를 측정하였다. 일반적으로 상호작용 확률이 높은 것으로 예측된 실험 그룹에 높은 민감도와 특이도를 보였다. 이 사실은 본 논문에서 고안된 확률 식이 단백질 쌍 집합의 각 단백질 쌍들간에 상호작용 가능성 순위를 부여하는데 사용될 수 있음을 나타낸다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 도메인 조합 기반 단백질 상호작용 예측 시스템 구조에 대하여 설명하고, 3장에서는 본 논문의 상호작용 가능성 순위 부여 기법을 기술한다. 4장에서는 제안된 상호작용 가능성 순위 부여 기법의 유효성을 검증하고, 5장에서 결론을 맺는다.

2. 도메인 조합 기반 단백질 상호작용 예측 기법

도메인 조합 기반 단백질 상호작용 예측 기법은 두 단백질 간의 상호작용은 단백질 내에 존재하는 각각의 도메인 조합 간의 상호작용의 결과라는 가정에 근거하고 있다. 따라서 도메인 조합 기반 단백질 상호작용 예측 기법은 상호작용이 있는 것으로 알려진 단백질 쌍의 정보로부터 도메인 조합 간의 상호작용 가능성을 추론한다. 그리고 새로운 단백질 쌍의 상호작용 가능성은 앞서 추론된 도메인 조합 간의 상호작용 가능성에 근거하여 판단한다. 이 기법은 먼저 상호작용하는 것으로 알려진 단백질 쌍과 상호작용 하지 않는 것으로 알려진 단백질 쌍 정보를 이용하여 도메인 조합 쌍(domain combination pair 또는 *dc-pair*)의 상호작용 확률을 추측한다. 임의의 단백질 쌍에 대한 상호작용 확률은 해당 단백질 쌍에 있는 단백질을 구성하는 도메인 조합 쌍의 상호작용 효과를 계산하여 결정한다.

도메인 조합 기반 단백질 상호작용 예측 기법에서는 상호작용이 있는 것으로 알려진 단백질 쌍 집합에 있는 도메인 조합 쌍의 출현 확률 행렬과 임의로 짝지어진 상호작용이 없는 것으로 가정된 단백질 쌍 집합에 있는 도메인 조합 쌍 정보의 출현 확률 행렬을 작성한다. 그 후 각각의 출현 확률 행렬에 기반하여 각 단백질 쌍을 0과 1사이의 실수 값으로 대응시키는 확률 함수를 고안한다. 다음으로, 상호작용하는 것으로 알려진 단백질 쌍에 고안된 함수를 적용하여 얻어지는 값의 분포와 상호작용하지 않는 것으로 가정된 단백질 쌍에 고안된 함수를 적용하여 얻어지는 값을 분포를 각각 생성한다. 본 논문에서 이 실수를 *PIP* 값이라 부른다.

모든 상호작용 및 상호작용하지 않는 단백질 쌍 집합에 있는 모든 단백질 쌍에 대해 *PIP* 값을 구하는 함수가 적용 된다. 이 결과로 얻은 두 *PIP* 분포를 이용하여 미지의 새로운 단백질 쌍이 주어지면 *PIP* 값을 얻고,

얻어진 PIP 값이 어느 PIP분포에 속하게 되는지 판단하여 그 상호작용 가능성을 결정한다.

2.1 출현 확률 행렬(Appearance Probability matrix; AP matrix)

출현 확률 행렬의 생성은 다음과 같다. 주어진 단백질 쌍 집합에서, n개의 서로 다른 단백질 p_1, p_2, ..., p_n이 있을 때, 단백질의 도메인 조합은, $dc(p_1)$, $dc(p_2)$, ..., $dc(p_n)$이 되며, 이 조합의 합집합은 m개의 서로 다른 도메인 조합 dc_1, dc_2, ..., dc_m을 구성하게 되어, m-by-m AP 행렬이 생성된다. 행렬에서 원소 AP_{ij}는 주어진 단백질 쌍 집합에서 도메인 조합 $<dc_i,dc_j>$ 출현 확률을 대표한다.

AP 행렬을 만들기 위하여 먼저 가중 출현빈도(Weighted Frequency:WF) 행렬을 먼저 생성한다. 이 때 각 행과 열은 도메인 조합을 나타내며, 행렬의 각 원소 해당 행에 있는 도메인 조합과 해당 열에 있는 도메인 조합의 쌍을 나타낸다. WF 행렬에서는 주어진 단백질 쌍의 집합에서의 도메인 조합 출현 빈도가 등록된다. 원소 WF_{ab}는 도메인 조합 $<a,b>$의 가중 출현 빈도(weighted appearance frequency)이며, 다음 식 (1)에 의하여 계산된다.

$$\sum_{(p_i,q_j)} \frac{1}{|dc(p_i)| \times |dc(q_j)|} \quad (1)$$

단, (p_i,q_j)는 dc-$pair(p_i,q_j)$내에 $<a,b>$를 포함하는 단백질 쌍

즉, dc-$pair$ $<a,b>$를 포함하는 모든 단백질 쌍 $<p_i,q_j>$에서 $1/(|dc(p_i)|\times|dc(q_j)|)$값을 계산하여 더함으로써 이 식의 최종 결과가 계산된다. 식 (1)에 의해서, dc-$pair$ $<a,b>$의 잠재적인 기여 가중치가 계산된다. 가중치 부여의 의미는 상호작용 하는 단백질 쌍으로부터 얻어지는 가능한 도메인 조합 쌍의 수가 적으면 적을수록 각 dc-$pair$에 의한 상호작용에서의 기여도는 더 클 것이라는 가정에서 출발한다. WF 행렬이 생성된 후에 출현 확률 행렬(AP matrix)의 각 원소 값의 계산은 식 (2)에 따른다.

$$AP_{ij} = \frac{WF_{ij}}{\sum_{i,j} WF_{ij}} \quad (2)$$

즉, 가중 출현빈도 행렬의 모든 원소 값을 더한 값으로 원소 값을 나누어서 출현 확률 행렬을 얻는다. 이와 같이 얻어진 출현 확률 행렬의 각 원소 값은 해당 원소가 속하는 행과 열에 의해 표현되는 특정 도메인 조합이 출현할 확률을 나타내게 된다. 상호작용이 있는 것으로 알려진 단백질 쌍 그룹과 상호작용이 없는 것으로 추정되는 그룹 각각에 대해서 출현 확률을 구할 수 있으며, 이 때 얻어진 출현 확률 행렬을 AP^i, AP^r 행렬로 표시하고 이들의 공통 부분 $AP^i \cap AP^r$은 AP^c 행렬로 나타낸다. 각 행렬에 대한 자세한 정의는 다음과 같다.

- AP^r : 상호작용이 없는 것으로 추정되는 단백질 쌍 집합으로부터 얻어지는 AP 행렬
- AP^i : 상호작용이 있는 단백질 쌍 집합으로부터 얻어지는 AP 행렬
- AP^c : $AP^i \cap AP^r$

일단, 상호작용이 있는 것으로 알려진 쌍과 없는 것으로 추정되는 AP 행렬이 얻어지면 dc-$pair$를 각각 그들이 속하는 그룹으로 분류할 수 있으며, AP^i, AP^r, AP^c 개념을 이용하여 각 범주(category)를 명명할 수 있게 된다. AP^i 행렬을 구성하는 모든 dc-$pair$는 AP^i dc-$pair$ 공간을 구성하며, 같은 방법으로, AP^r dc-$pair$ 공간, AP^c dc-$pair$ 공간이 정의된다.

2.2 상호작용 가능성 확률(Primary Interaction Probability)

본 절에서 소개하는 PIP 함수는 두 개의 출현 확률 행렬을 기반으로 고안된 식으로 미지의 단백질 쌍 $<A,B>$의 상호작용 가능성을 확률로 표시하는 확률 식으로 볼 수 있다. PIP 함수 값을 계산하기 위해서는 먼저 다음 식 (3)을 이용하여 단백질 쌍 $<A,B>$로부터 이들의 도메인 조합 dc-$pair$를 산출한다.

$$dc-pair(P_1,P_2) = \{<dc_1,dc_2> \mid <dc_1,dc_2> \in dc(P_1)\times dc(P_2)\ or\ dc(P_2)\times dc(P_1)\} \quad (3)$$

단, $dc(P) = Powerset(domain(p)) - \{\phi\}$

$domain(P)$ = 단백질 p를 구성하는 도메인의 집합

한편 dc-$pair$는 이들이 나타나는 출현 확률의 분류에 따라, 다음과 같이 dc-$pair$를 분류한다.

- $DC_c(A,B)$ = {dc-$pair$ | dc-$pair \in dc$-$pair(A,B)$이고 AP^c에 나타남}
- $DC_{r-c}(A,B)$ = {dc-$pair$ | dc-$pair \in dc$-$pair(A,B)$이고 $AP^r - AP^c$에 나타남}
- $DC_{i-c}(A,B)$ = {dc-$pair$ | dc-$pair \in dc$-$pair(A,B)$이고 $AP^i - AP^c$에 나타남}

확률 식의 값을 결정하는 기본적인 아이디어는 다음과 같은 논리에 근거하고 있다. 그림 1과 같은 한 예를 들어 설명한다. 그림에서 각각의 흰색 볼은 상호작용이 있는 단백질 쌍 집합으로부터 발견되는 하나의 도메인 조합 쌍으로 표현하고, 검은 볼은 상호작용이 없는 단백질 쌍 집합으로부터 발견되는 하나의 도메인 조합 쌍으로 표현한다. 한편, dc-$pair$ c, d는 상호작용이 있는 것으로 알려진 단백질 집단에서 빈번히 발견되며, dc-$pair$ b는 임의로 만들어진 단백질 쌍 집단에서 빈번히 발견되고, dc-$pair$ a는 상호작용이 있는 것으로 알려진 집단과 그렇지 않은 집단에서 엇비슷하게 나타나는 상황

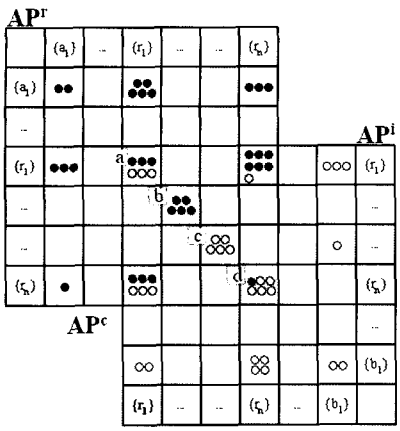


그림 1 AP 공간상에서 도메인 조합의 한 예

을 가정해보자. 이런 상황에서 두 단백질 쌍 <A,B> 와 <X,Y>로부터 도메인 조합 쌍 집합 c, d와 a, b가 각각 형성되는 경우에는 도메인 조합 쌍 집합 c, d를 형성시키는 단백질 쌍 <A,B>가 도메인 조합 쌍 집합 a, b를 형성시키는 단백질 쌍 <X,Y>보다는 상호작용을 일으킬 가능성이 높다고 추정한다. 이와 같은 논리에 의해서 단백질 쌍 <A,B>가 단백질 쌍 <X,Y>보다 상호작용을 일으킬 가능성이 높다고 판정할 수는 있지만 그 가능성을 확률적 수치로 표현하기 위해서는 별도의 수식을 준비되어야 한다.

먼저, AP^c dc-$pair$ 공간에서 $DC_c(A,B)$가 발견되었을 때의 상호작용 확률 식을 정의한다. 여기에서 확률은 $DC_c(A,B)$가 AP^c dc-$pair$ 공간에서 나타났을 때 상호작용하기 위한 단백질 쌍 <p,q>에 대한 확률을 의미한다. 다음은 주어진 단백질 쌍의 도메인 조합 쌍 집합 정보로부터 상호작용 가능성을 계산하는 확률 식 PIP에 관하여 기술한다. AP^c dc-$pair$ 공간에서 발견되는 $DC_c(A,B)$ 도메인 조합을 대상으로 상호작용 확률을 식 (4)와 같이 정의할 수 있다. 이 확률은 $DC_c(A,B)$가 AP^c dc-$pair$ 공간에서 발견될 때 단백질 쌍 <A,B>가 서로 상호작용할 확률을 의미한다. 먼저 상호작용이 일어나는 사건과 일어나지 않는 사건을 표현하기 위하여 확률 변수 X를 도입한다. 1 값은 상호작용이 일어나는 사건, 0 값은 상호작용이 없는 사건을 나타낸다. $DC_c(A,B)$) 도메인 조합을 대상으로 얻어질 확률 식 (4)는 다음과 같다.

$$P(X=1|DC_c(A,B))=$$

$$\frac{P(X=1)P(DC_c(A,B)|X=1)}{P(X=1)P(DC_c(A,B)|X=1)+P(X=0)P(DC_c(A,B)|X=0)} \quad (4)$$

여기에서, $P(X=1)$, $P(X=0)$, $P(DC_c(A,B) \mid X=1)$, $P(DC_c(A,B) \mid X=0)$의 정의는 다음과 같다.

$$P(X=1)=\frac{k\cdot I_{total}\cdot\sum_{i,j}(AP_I^c)_{ij}}{k\cdot I_{total}\cdot\sum_{i,j}(AP_I^c)_{ij}+(1-k)\cdot R_{total}\cdot\sum_{i,j}(AP_R^c)_{ij}},$$

$$P(X=0)=\frac{(1-k)\cdot R_{total}\cdot\sum_{i,j}(AP_R^c)_{ij}}{k\cdot I_{total}\cdot\sum_{i,j}(AP_I^c)_{ij}+(1-k)\cdot R_{total}\cdot\sum_{i,j}(AP_R^c)_{ij}},$$

$$P(DC_C(A,B)|X=1)=|DC_c(A,B)|!\prod_{\{i,j\}|(i,j)\in DC_c(A,B)\}}\frac{(AP_I^c)_{ij}}{\sum_{i,j}(AP_I^c)_{ij}},$$

$$P(DC_C(A,B)|X=0)=|DC_c(A,B)|!\prod_{\{i,j\}|(i,j)\in DC_c(A,B)\}}\frac{(AP_R^c)_{ij}}{\sum_{i,j}(AP_R^c)_{ij}}$$

이 때, $P(X=1)$는 AP^c에 존재하는 총 dc-$pair$ 공간에서 상호작용이 있는 단백질 쌍으로부터 만들어진 dc-$pair$ 공간에 있을 확률이며, $P(X=0)$은 AP^c의 도메인 조합 공간에서 상호작용이 없다고 추정되는 단백질 쌍으로부터 생성된 dc-$pair$ 공간에 있을 확률이다. I_{total}과 R_{total}은 상호작용이 있는 단백질 쌍과 상호작용이 없는 것으로 간주되고 있는 단백질 쌍의 총 개수를 각각 나타낸다. 식에서 상수 k는 자연계에서 I_{total}과 R_{total}의 비율을 나타내며 이 값을 정확하게 알 수 없으므로, 추후에 최대 가능성 추정(maximum likelihood estimation)적용을 통하여 결정한다.

$P(DC_c(A,B) \mid X=1)$는 AP^i 공간에서 $DC_c(A,B)$에 속하는 dc-$pair$ 집합이 만들어질 확률이고, $P(DC_c(A,B) \mid X=0)$는 AP^r 공간에서 $DC_c(A,B)$에 속하는 dc-$pair$ 집합이 만들어질 확률이다. $AP_I{}^c$와 $AP_R{}^c$는 각각 상호작용이 있는 dc-$pair$ 공간과 상호작용이 없는 것으로 간주되고 있는 dc-$pair$ 공간에서의 AP^c를 각각 의미한다.

한편, 동일하게, $DC_{i-c}(A,B)$ 도메인 조합을 대상으로 얻어질 확률 식은 (5)와 같다.

$$P(X=1|DC_{i-c}(A,B))=$$

$$\frac{P(X=1)P(DC_{i-c}(A,B)|X=1)}{P(X=1)P(DC_{i-c}(A,B)|X=1)+P(X=0)P(DC_{i-c}(A,B)|X=0)} \quad (5)$$

식 (5)에서, $P(X=1)$, $P(X=0)$는 $AP^{i-c} \subseteq AP^i$이므로 확률변수 X의 정의에 따라 각각 1, 0이 되어 최종적으로 얻어지는 $P(X=1|DC_{i-c}(A,B))$는 1이 된다. 참고로 $DC_{r-c}(A,B)$도메인 조합을 대상으로 얻어질 확률 식은 (6)과 같다.

$$P(X=1|DC_{r-c}(A,B))=$$

$$\frac{P(X=1)P(DC_{r-c}(A,B)|X=1)}{P(X=1)P(DC_{r-c}(A,B)|X=1)+P(X=0)P(DCr_{r-c}(A,B)|X=0)} \quad (6)$$

식 (6)에서, $P(X=1)$, $P(X=0)$은 $AP^{r-c} \subseteq AP^r$이므로 확률변수 X의 정의에 따라 각각 0, 1이 되어 최종적으로 얻어지는 $P(X=1|DC_{r-c}(A,B))$ 역시 1이 된다.

이중 상호작용은 AP^i영역에서 발생하므로, 식 (4)와

(5)를 이용하여, $DCc(A,B)$ dc-$pairs$를 갖는 (A,B) 단백질 쌍의 상호작용 가능성 확률(Primary Interaction Probability; PIP)은 다음 식 (7)에 의하여 계산된다.

$$
\begin{aligned}
&PIP(A,B)\\
&=\left(1-\frac{\|AP^c\|}{\|AP^i\|}\right)\cdot P(X=1|DC_{i-c}(A,B)+\frac{\|AP^c\|}{\|AP^i\|}\cdot P(X=1|DC_c(A,B)\\
&=\left(1-\frac{\|AP^c\|}{\|AP^i\|}\right)\cdot 1+\frac{\|AP^c\|}{\|AP^i\|}\cdot P(X=1|DC_c(A,B)\\
&=1-\frac{\|AP^c\|}{\|AP^i\|}(1-P(X=1|DC_c(A,B))) \qquad (7)
\end{aligned}
$$

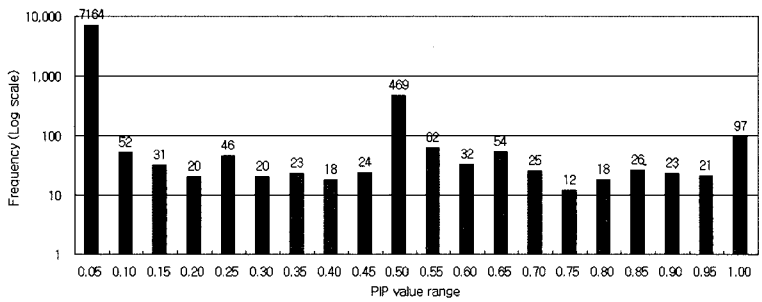
2.3 PIF 분포와 상호작용 예측

일단 두 번째 단계에서 PIP 최종식이 얻어지면, 식 (7)에 따라 상호작용이 있는 단백질 쌍과 없는 것으로 간주된 단백질 쌍 집합에 대한 PIP 값을 계산할 수 있다. 때때로, 그들을 비교하기 위하여, 분포를 정규화한다. 한편 PIP 함수는 단백질 쌍을 실수 0~1 범위 안에 대응시키는 함수의 일종으로 해석할 수 있다. 단백질 상호작용 예측 분포가 얻어지면, 이 분포에 대한 2-카테고리 분류(two-category classification) 기법 적용이 가능하다. 즉, 임의로 주어진 단백질 쌍에 대하여, 상호작용 가능성을 예측하기 위해서는 그 단백질 쌍의 PIP 값이 어느 분포에 속할지를 결정해야 한다. 2-카테고리 분류의 많은 기법이 있지만 이를 확률적으로 표현하기 위하여 단백질 쌍의 조건부 확률을 계산하여 어떤 카테고리에 속하는지를 결정하였다.

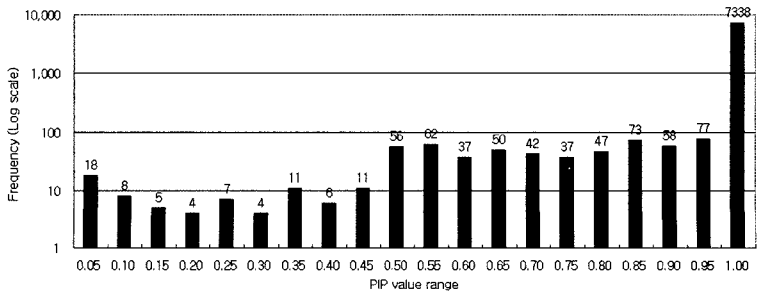
3. 상호작용 가능성 순위 부여 기법(Interaction Possibility Ranking Method)

2장에서 얻어진 PIP 식을 곧바로 특정 단백질 쌍의 상호작용 확률로 간주하는 것을 생각해 볼 수도 있다. 하지만 얻어진 PIP 최종 식 (7)을 상호작용이 있는 단백질 쌍과 없는 것으로 간주된 단백질 쌍 집합에 적용하여 각 집단 별로 0과 1사이에 분포하는 PIP 값 분포를 보면 상호작용이 있는 단백질 쌍의 경우에도 자주 낮은 PIP 값을 갖는 경우도 있고 상호작용이 없는 것으로 추정되는 단백질 쌍의 경우에 있어서도 높은 PIP 값을 갖는 경우가 자주 발견된다(그림 2 참조).

이것은 하나의 단백질 쌍 상호작용 확률을 직접 PIP 값에 이용하는 것 보다 상호작용 및 상호작용하지 않는 PIP 값 분포를 기반으로 한 다른 확률 식을 고안해야 함을 의미하고 있다. 이 절에서 제안한 기법은 상호작용 및 상호작용하지 않는 PIP 분포내의 특정한 분포 영역에서 나타나는 PIP값 확률을 이용하기 위함이다. 다시 말하면, 단백질 쌍 (A,B)에서 PIP 값이 계산되고 그 상호작용 확률 값은 PIP 분포내에서 나타나는 PIP 값 확률의 계산에 의해서 결정된다. 따라서 PIP 값을 곧바로 단백질 쌍의 상호작용 확률로 간주하기 보다는 PIP



(a) 상호작용이 없는 것으로 추정되는 단백질 쌍에 대한 PIP 값의 분포



(b) 상호작용이 있는 단백질 쌍의 PIP 분포

그림 2 단백질 쌍의 PIP 분포(log scale)

식을 두 개의 집단에 적용하여 집단 별로 *PIP* 값 분포를 얻고 얻어진 분포를 기준으로 주어진 특정 단백질 쌍이 특정 분포에 어떤 확률로 속할지를 결정하는 것이 적절하다. 본 절에서는 특정 단백질 쌍이 특정 분포에 속할 확률을 계산하는 기법을 개발하고 이를 바탕으로 복수의 단백질 쌍에 대한 상호작용 가능성을 순위 관계로도 사용할 수 있도록 한다.

본 절에서는 도메인 정보는 알려져 있지만 상호작용에 관한 실험 정보는 보고 되지 않은 두 단백질의 상호작용 가능성을 확률 값으로 나타내는 방안을 제시한다. 이를 위해서 먼저 대상이 되는 두 단백질 *A*, *B*의 *PIP* 값 *PIP*(*A*,*B*)를 계산한다. 두 단백질 *A*, *B*의 상호작용 확률 값은 계산된 *PIP*(*A*,*B*) 값이 학습에 사용된 전체 단백질 쌍에서 동일한 *PIP* 값을 갖는 단백질 쌍이 존재하는 경우와 그렇지 않은 경우로 나누어서 계산한다.

3.1 학습에 사용된 전체 단백질 쌍 내에 *PIP*(*A*,*B*)와 동일한 *PIP* 값을 갖는 단백질 쌍이 있는 경우

만약 상호작용 또는 상호작용하지 않는 *PIP* 값 분포에서 *PIP*(*A*,*B*)와 동일한 *PIP* 값을 갖는 단백질 쌍이 있을 경우 상호작용 확률의 계산은 다소 간단하다. 이를 식 (8)에 나타내고 있다.

$$P(X=1|p=PIP(A,B))=\frac{P(X=1)P(p=PIP(A,B)|X=1)}{P(X=1)P(p=PIP(A,B)|X=1)+P(X=0)P(p=PIP(A,B)|X=0)} \quad (8)$$

여기에서, $P(X=1)$, $P(X=0)$, $P(p=PIP(A,B)|X=1)$, $P(p=PIP(A,B)|X=0)$는 각각

$$P(X=1)=\frac{k\cdot\sum_{i=1}^{m} freq_i^x}{k\cdot\sum_{i=1}^{m} freq_i^x+(1-k)\cdot\sum_{i=1}^{n} freq_i^y},$$

$$P(X=0)=\frac{(1-k)\cdot\sum_{i=1}^{n} freq_i^y}{k\cdot\sum_{i=1}^{m} freq_i^x+(1-k)\cdot\sum_{i=1}^{n} freq_i^y},$$

$$P(p=PIP(A,B)|X=1)=\frac{freq^x_{PIP(A,B)}}{\sum_{i=1}^{m} freq_i^j},$$

$$P(p=PIP(A,B)|X=0)=\frac{freq^y_{PIP(A,B)}}{\sum_{i=1}^{n} freqy_i^y}$$

으로 정의 된다.

$$P(X=1\,\|\,p-PIP(A,B)|\le\frac{w}{2})=\frac{P(X=1)P(|p-PIP(A,B)|\le\frac{w}{2}|X=1)}{P(X=1)(|p-PIP(A,B)|\le\frac{w}{2}|X=1)+P(X=0)P(|p-PIP(A,B)|\le\frac{w}{2}|X=0} \quad (9)$$

$$P(X=1\,\|\,p-PIP(A,B)|\ge\frac{w}{2})=PIP(A,B) \quad (10)$$

식 (8)에서 $P(X=1)$는 총 단백질 쌍 수 중 상호작용이 있다고 알려진 쌍의 비율을 나타내며, $P(X=0)$는 총 단백질 쌍 수 중에 상호작용이 없다고 추정되는 쌍의 비율을 나타낸다. $freq_i^x$와 $freq_i^y$는 각각 PIP_i^x를 가지는 샘플이 상호작용이 알려진 집합에 출현하는 빈도와 PIP_i^y를 가지는 샘플이 상호작용이 없다고 추정되는 집합에 출현하는 빈도이다. 또한, 상수 *k*는 식 (4)에서 쓰여지는 것과 동일하다. $P(p=PIP(A,B)|X=1)$는 상호작용이 있는 모 집단 내에서 확률변수 p 값이 *PIP*(*A*,*B*)가 될 확률을 의미한다. 마찬가지로, $P(p=PIP(A,B)|X=0)$는 상호작용이 없다고 추정되는 모 집단 내에서 확률변수 p 값이 *PIP*(*A*,*B*)가 될 확률을 의미한다. 경우에 따라서 *PIP*(*A*,*B*)값을 갖는 샘플이 모집단 내에 존재하지 않는 수도 있어, 이 경우에는 사전에 정의된 범위 안에 존재하는 p 값을 대신 사용한다. 만약 *PIP* 분포내에서 충분한 *PIP* 값이 있다면 식 (8)에 의해 계산된 상호작용 확률은 상당히 확실성이 있을 것이고, 그렇지 않은 경우는 계산된 상호작용 확률 값의 이용에 있어서 신중함을 기해야 할 것이다.

3.2 학습에 사용된 전체 단백질 쌍 내에 *PIP*(*A*,*B*)와 동일한 *PIP* 값을 갖는 단백질 쌍이 없는 경우

만약 상호작용 또는 상호작용하지 않는 *PIP* 값 분포에서 *PIP*(*A*,*B*)와 동일한 *PIP* 값을 갖는 단백질 쌍이 없을 경우에는 모든 가능한 도메인 조합 쌍에서 *PIP* 값은 얻어진 단백질 *A* 및 *B* 형식으로 될 수 있다. *PIP* 분포에서 하나의 *PIP* 값으로 *PIP*(*A*,*B*)에 가까운 값이 선택되며, 선택된 *PIP* 값은 단백질 쌍 (A,B)의 상호작용 확률 계산에 우선적으로 사용된다. 이 *PIP* 값은 *PIP*(*A*,*B*)로부터 미리 정해진 거리 내에 있을 것이다. 여기에서 거리는 사용자 또는 시스템에서 사용하는 우선적 값에 의해 결정된다. 일단 *PIP* 값이 결정되면 상호작용 확률 계산을 위해서 3.1 절 식이 사용된다.

이런 시도의 이론적 배경으로는 단백질 쌍 (*A*,*B*)에 의해서 구성된 도메인 또는 도메인 조합 쌍은 단백질 *A*, *B*의 상호작용 자극에 의해서 정해진 규칙으로 실행된다는 사실을 기본으로 하기 때문이다. 그렇기 때문에 도메인 조합 쌍의 *PIP* 값은 도메인 조합 쌍을 포함하는 하나의 단백질 쌍의 특성을 반영 할 것이다. 또한 *PIP*(*A*,*B*)에 단백질 쌍(*A*,*B*)을 반영하기 위하여 원래의 *PIP* 값에 가까운 *PIP* 값을 선택한다.

만약 위의 과정에 의해서 동일한 *PIP* 값이 발견되지 않을 경우에 *PIP* 분포의 가장 가까운 값 $PIP_N(A,B)$로부터 *PIP*(*A*,*B*)까지의 거리에 의해서 다음과 같은 두가지 경우를 생각한다. 각 경우의 상호작용 확률 계산을 위하여 다른 기법을 개발하고 사용한다.

- 경우 1 : $|PIP(A,B)-PIP_N(A,B)|<\delta$: 이 경우는 k-nearest-neighbor estimation 기술에 근접한 기술

을 사용한다. 단백질 쌍 (A,B)의 주어진 PIP 값에서 실현 가능한 윈도우 크기 w'는 집합이고, 만약 범위 내의 상호작용 단백질 쌍의 번호가 k를 넘으면 실행한다. 만약 상호작용 단백질 쌍의 번호가 k 이내 이면 윈도우 크기 w'는 범위 내에서 더 많은 상호작용 단백질 쌍에 근접하도록 증가되어 진다. 이 과정은 상호작용 단백질 쌍의 번호가 k를 넘을 때까지 반복되며, 그 시점에서 사용된 윈도우 크기 w는 마지막이 된다. 윈도우 크기 w가 결정된 후 단백질 쌍 (A,B)의 상호작용 확률은 식 (9)에 의해 계산된다. 범위 표기법을 제외하고 식 내의 형태는 식 (8)과 비슷하다. 식 (9)에서 $P(X=1)$은 전체 단백질 쌍의 상호작용 단백질 쌍 비율이고, $P(X=0)$은 전체 단백질 쌍의 상호작용하지 않는 단백질 쌍 비율이다. $freq^x_i$ 는 상호작용 단백질 쌍의 집합에서 값 PIP^x_i를 갖는 샘플 수 이다. $Freq^y_i$는 단백질 쌍의 상호작용하지 않는 집합 내 값 PIP^y_i를 갖는 샘플 수 이다. $P(|\ p-PIP(A,B)\ | \le w/2 \mid X=1)$은 단백질 쌍의 상호작용 집합 내에서 변수 p가 $PIP(A,B)-w/2$와 $PIP(A,B)+w/2$ 범위에 있을 때의 확률이다. $P(|\ p-PIP(A,B)\ | \le w/2 \mid X=0)$은 단백질 쌍의 상호작용하지 않는 집합 내에서 변수 p가 $PIP(A,B)-w/2$와 $PIP(A,B)+w/2$ 범위에 있을 때의 확률이다.

- 경우 2 $|\ PIP(A,B)-PIP_N(A,B)\ | > \delta$: 이 경우는 상호작용 확률은 $PIP(A,B)$ 값과 PIP 값 자신이 상호작용 확률이 되는 것에 의해서 결정되어지며, 식 (10)으로 표현된다.

4. 검증(Validation)

제안한 예측 시스템의 검증을 위하여, 다음과 같은 2개의 단백질 쌍 데이타를 준비하였다. 상호작용이 알려진 단백질 쌍 집합은 DIP 데이타베이스(http://dip.doe-mbi.ucla.edu)[15]의 효모(yeast)에서 총 15,174개의 상호작용이 보고된 단백질 쌍을 준비하였다. 검증에는 상호작용이 있는 것으로 보고된 15,174개의 단백질 쌍 중에서 도메인 정보를 모두 가지고 있는 단백질 쌍은 약 7500여개로서 이것이 검증에 사용되었다. 단백질에 대한 도메인의 정보는 PDB(http://www.ebi.ac.uk/proteome/)[16]에서 추출하였다.

반면에, 상호작용이 없다고 추정되는 단백질 쌍은 효모 유기체의 도메인 정보를 가진 임의적 쌍에서 인위적으로 생성되었다. 아직까지 상호작용 하지 않는 단백질 쌍에 대한 공표된 정보가 없다. 상호작용하지 않는 단백질 쌍이 준비되었을 때 도메인 정보가 알려진 단백질 쌍 집단에서 상호작용이 있는 것으로 보고된 단백질 쌍 집단을 제거하는 방식을 통하여 임의로 생성되었고, 상호작용이 없는 단백질 쌍의 같은 개수(7,500)가 초기 계산에 사용되었다. 현재까지 모든 단백질에 대한 상호작용이 밝혀진 것이 아니므로, 이상의 방법을 통해서, 상호작용이 없다고 추정되는 집단 안에 상호작용이 있는 단백질 쌍이 완전히 제거되지는 않을 것이다. 그러나, 만일 전체 단백질 쌍 공간 안에 상호작용하는 단백질 쌍이 아주 드물다고 추측한다면, 본 예측 모델에서 사용된 상호작용이 없다고 추정되는 집단으로도 충분할 것이며, 입증 결과가 이러한 방법으로 상호작용이 없다고 추정되는 집단을 생성하고 사용하는 것이 적절하다는 것을 보일 것으로 예상된다. 이상의 방법으로 2개의 집단을 준비한 후, 각각을 학습 집단과 검증 집단으로 나누었다.

이전의 도메인 조합기반 단백질 상호작용 예측 기술 검증방법에 의해서 검증한 결과는 단지 86%의 민감도(sensitivity)과 56%의 특이도(specificity)을 보여주었다[8]. 한편, 자연계에는 상호작용하는 단백질 쌍보다 상호작용하지 않는 단백질 쌍이 더 많이 존재하는 것이 보통이기 때문에 본 논문에서는 상호작용하지 않는 단백질 쌍 집합의 크기를 점점 증가시켜 가며 재 평가를 시도하였다.

표 1 평가집합 크기 비율의 변환에 따른 민감도와 특이도

Size ratio	3	5	10	15	20
Sensitivity	96.44	93.95	90.79	89.97	84.36
Specificity	37.37	42.48	60.64	54.15	75.00

표 1은 상호작용하지 않는 단백질 쌍 집합의 크기 비율이 변화함에 따른 각 그룹의 민감도와 특이도를 나타내고 있다. 여기에서 상호작용하지 않는 단백질 쌍 집합의 크기가 상호작용하는 단백질 쌍 집합의 20배일 경우 높은 84%의 민감도와 75%의 특이도를 나타내었다. 이것은 도메인 조합 기반 단백질 상호작용 예측 기술이 상호작용 가능성 순위 부여 방법을 개발하기 위해 사용될 수 있음을 의미한다.

제안된 상호작용 가능성 순위 부여 방법의 유효성을 평가하기 위하여 각 1590개와 1589개의 상호작용과 상호작용하지 않는 단백질 쌍을 우선적으로 조사하였다. 우선 상호작용 및 작용하지 않는 단백질 쌍 집합 내에서 PIP 분포의 PIP 값에 일치하는 단백질 쌍 비율을 측정하였다. 표 2에 그 결과를 보여준다. 표 2에서 상호작용하지 않는 단백질 쌍 집합의 크기가 상호작용하는 단백질 쌍 집합 크기의 20배 일 때 단백질 쌍의 상호작용 및 상호작용 하지 않는 각각의 평가 집합은 약 50% 및 45%의 PIP 일치도를 나타내고 있다. 상호작용하지 않는 단백질 쌍에서 전체 학습 집합의 크기가 증가하면

단백질 쌍 의 *PIP* 값 일치 비율도 증가 한다. 또한, 전체 학습 집합의 크기가 증가함과 함께 단백질 쌍의 *PIP* 값 일치 비율은 약간의 변화만 있음을 알 수 있다. 이것은 단백질 쌍 학습 집합의 총 크기가 증가해도 상호작용 단백질 쌍의 학습 집합 크기는 변하지 않기 때문이다.

표 2 *PIP* 분포 내 평가 단백질 쌍의 *PIP* 값 일치 정도

	3.0	5.0	10.0	15.0	20.0
I	721	733	727	851	780
Hit ratio(%)	45.35	46.10	45.72	53.52	49.06
II	362	467	529	540	643
Hit ratio(%)	22.77	29.37	33.27	33.96	40.44
Total	1083	1200	1256	1391	1423
Hit ratio(%)	37.79	39.59	40.48	45.93	45.16

I: 상호작용 쌍의 수, II: 상호작용하지 않는 쌍의 수

표 3 상호작용 및 상호작용하지 않는 단백질 쌍의 수와 상호작용 확률

Interaction Probability	Number of interacting protein-pairs	Number of non-interacting protein-pairs
0.0 - 0.2	48(3.0%)	884(59.3%)
0.2 - 0.4	146(9.2%)	35(2.4%)
0.4 - 0.6	202(12.7%)	415(27.9%)
0.6 - 0.8	55(3.5%)	26(1.7%)
0.8 - 1.0	1139(71.6%)	130(8.7%)
Total	1590(100.0%)	1490(100.0%)

Note: 평가에 사용된 상호작용 및 상호작용하지 않는 단백질 쌍 크기의 비율은 15

표 3에서는 상호작용 가능성 순위 부여 기법을 확인하기 위하여 순위 부여 기법에 의해서 결정된 상호작용 확률의 각 범위에서 발견된 단백질 쌍의 수를 나타내었다. 상호작용 및 상호작용하지 않는 평가 그룹의 크기는 1590과 1490이었다. 상호작용하는 평가 그룹의 큰 단백질들은 높은 상호작용 확률 범위 내에서 관찰되었고, 상호작용하지 않는 평가 그룹의 큰 단백질들은 낮은 상호작용 확률 범위 내에서 관찰되었다. 그러나, 상호작용 평가 그룹 내의 몇몇 단백질 쌍들은 낮은 상호작용 확률을 갖는다. 이것은 순위 부여 기법이 신중히 적용되어져야 함을 의미하며, 상호작용 확률이 중간(0.4-0.6)에 있을 때 순위 부여 기법의 적용에 있어서 더 많은 신중을 기울여야 함을 나타낸다. 그럼에도 불구하고 본 논문에서 제안한 순위 부여 기법은 어느 정도의 상당한 확실성이 있음을 결론으로 내릴 수 있다.

5. 결 론

본 논문에서는 복수의 단백질 쌍에 있어서 도메인 조합 기반 상호작용 가능성 순위 부여 기법을 제안하였다. 그리고 기술적 평가 구축 및 그 유효성을 증명하였다. 평가 과정에서 도메인 조합 기반 단백질-단백질 상호작용 기법은 재평가 되었고 예측 정밀도는 학습 상호작용 및 상호작용하지 않는 단백질 쌍 집합의 비율의 변화에 의해서 상당한 증가가 있었음을 확인하였다. 본 기법의 예측 능력은 인터넷 상에서 더 많은 단백질 상호작용 데이타가 공개되고 축적되어 질때 그 빛을 발할 것으로 확신한다.

제안된 기법과 시스템의 기여도는 다음과 같이 요약할 수 있다. 첫째, 본 예측 시스템을 이용하여, 생물학자로 하여금, 많은 비용과 시간이 소요되는 단백질 상호작용 실험을 통하지 않고 단백질 상호작용에 대해서 시간과 비용 측면에서 획기적인 기여를 할 것으로 기대된다. 구체적으로는 본 예측 기법에서 제공하고 있는 복수의 단백질 쌍의 상호작용 가능성 순위 부여기법은 생물학자가 자신의 단백질 상호작용 실험을 설계할 때 매우 유용하게 활용될 수 있을 것이다. 둘째, *PIP* 값과 분포들은 인터넷 상에 이미 공개된 단백질 상호작용 데이타 집단에 포함된 오류 데이타를 찾아 바로잡는 데 있어서도 유용하게 활용될 수 있을 것으로 기대된다. 셋째, 본 예측 시스템에서 사용한 계산적 방법에 의한 단백질 상호작용 예측은 단시간 내에 대규모 단백질 쌍에 대해서 상호작용 가능성을 예측할 수 있어 이를 기반으로 대규모 단백질 상호작용 네트워크 구성이 용이하고 다시 이를 기반으로 수많은 단백질 중에서 중요한 단백질을 추정하고 검증하는 데 활용할 수 있을 것으로 기대된다[17]. 넷째, 본 시스템은 미지의 단백질에 대한 기능을 추정하는 것과 같은 단백질 동정(identification)시에 기본적인 계산적 접근방법으로 활용될 수 있다. 다섯째, 본 연구에서 제안하고 있는 예측 시스템은 생물학자들이 그들의 연구 분야에서 유사한 경우를 만났을 때 참고 모델로 이용될 수 있다. 향후에는 효모에 기반하여 구축된 본 시스템을 단백질의 상동관계 검색(homology search) 기능 등을 활용하여 쥐와 인간과 같은 다른 종의 단백질 집단에 대해서도 확장 적용할 수 있는 방안을 모색할 계획이다[18]. 현재 시스템의 프로토타입이 인터넷 상(http://silver.icu.ac.kr:8080/torajim/index.html)에서 공개되어 있어서 많이 활용될 수 있을 것이다.

참 고 문 헌

[1] J. R. Bock and D. A. Gough, Prediction of protein-protein interaction from primary structure, Bioinformatics, 17, 455-460, 2001.

[2] J. Park, M. Lappe and S. A. Teichmann, Map ping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J. Mol. Biol., 307, 929-938, 2001.

[3] E. Sprinzak and H. Margalit, Correlated sequence-signatures as markers of protein-protein interaction. J. Mol. Biol., 311, 681-692, 2001.
[4] J. Wojcik and V. Sch?chter, Protein-Protein interaction map inference using interacting domain profile pairs. Bioinformatics, 17 Suppl., S296-S305, 2001.
[5] A.J. Enright and C.A. Ouzounis, Chapter 33: Protein-Protein Interactions-A Molecular Cloning Manual, Cold Spring Harbor Laboratory Press, Cold spring Harbor, NY, 2002.
[6] S. Ng, Z. Zhang and S. Tan, Integrative approach for computationally inferring protein domain interactions. Bioinformatics, 19, 923-929, 2003.
[7] M. Deng, S. Metah, F. Sun and T. Chen, Inferring Domain-Domain Interactions from ProteinProtein Interactions, Genome Research, 12, 1540-1548, 2002.
[8] D. Han, H. Kim, J. Seo, and W. Jang. Domain Combination based Probabilistic Framework for Protein-Protein Interaction Prediction. Genome Informatics, 14: 250-259, 2003.
[9] 한동수, 서정민, 김홍숙, 장우혁, 도메인 조합 기반 단백질-단백질 상호작용 확률 예측 툴, 정보과학회 논문지 : 컴퓨팅의 실제, 10권 4호, 299-304, 2004.
[10] A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 247, 536-540, 1995.
[11] L. Holm, and C. Sander, The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res., 24, 206-210, 1996.
[12] F. M. G. Pearl, D. Lee, J. E. Bray, I. Sillitoe, A. E. Todd, A. P. Harrison, J. M. Thornton and C. A. Orengo, Assigning genomic sequences to CATH. Nucleic Acids Research, 28, 277-282, 2000.
[13] N. Goffard, V. Garcia, F. Iragne, A. Groppi and A. de Daruvar, IPPRED: Server for Proteins Interactions Inference. Bioinformatics, 19, 903-904, 2003.
[14] S. Dohkan, A. Koike and T. Takagi, Prediction of Protein-Protein Interactions Uisng Support Vector Machines, Fourth IEEE Symposium on Bioinformatics and Bioengineering, 576-583, May, 2004.
[15] I. Xenarios, E. Fernandez, L. Salwinski, X. J. Duan, M. J. Thompson, E. M. Marcotte and D. Eisenberg, DIP: The Database of Inter acting Proteins: 2001 update. Nucleic Acids Res., 29, 239-241, 2001.
[16] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni and F. Servant, The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res., 29, 37-40, 2001.
[17] A. J. Enright, I. Iliopoulos, N. C. Kyrpides and C. A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events. Nature, 402, 86-90, 1999.
[18] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates and D. Eisenberg, Detecting protein function and protein-protein interactions from genome sequences. Science, 285, 751-753, 1999.

한 동 수

1989년 서울대학교 계산통계학과(학사) 1991년 서울대학교 계산통계학과(석사) 1996년 일본 교토대학교 정보공학과(박사). 1996년 4월~1996년 7월 일본 NEC C&C 중앙연구소 연구원. 1996년 9월~1997년 10월 ㈜현대정보기술 정보기술연구소 책임연구원. 1997년 11월~현재 한국정보통신대학교 공학부 부교수

김 홍 숙

1994년 서강대학교 컴퓨터학과(학사) 1996년 서강대학교 컴퓨터학과(석사) 1996년 3월~1998년 2월 현대정보기술(주) 정보기술연구소 선임연구원. 2003년 한국정보통신대학교 공학부(박사). 2001년 2월~2004년 9월 엔솔테크㈜ 기술연구소 BIT S/W개발실장. 2005년 4월~현재 한국전자통신연구원 이동통신연구단 선임연구원

장 우 혁

2003년 충남대학교 컴퓨터공학교육학과(학사). 2005년 한국정보통신대학교 공학부(석사). 2005년 2월~현재 한국정보통신대학교 공학부. Bioinformatics and Information Management track 박사과정

이 성 독

1988년 전북대학교 전자공학과(학사) 1991년 전북대학교 전자공학과(석사) 2002년 일본 토호쿠대학교 정보과학연구과(박사). 1991년 5~1993년 7월 군산대학교 전기공학과 조교. 2002년 4~2003년 3월 일본 토호쿠대학교 전기통신연구소 연구원. 2003년 8월~2005년 7월 한국정보통신대학교 공학부 계약교수. 2005년 8월~현재 한국정보통신대학교 공학부 연구조교수