A Domain Combination Based Probabilistic Framework for Protein-Protein Interaction Prediction

도메인 조합 기반 단백질-단백질 상호작용 확률 예측기법

DongSoo Han[1], Jung Min Seo[2], Hong-Soog Kim[3], Woo Hyuk Jang[4]

[1,2,3,4] School of Engineering, Information and Communications University

*To whom correspondence should be addressed. E-mail: dshan@icu.ac.kr

Abstract

In this paper, we propose a probabilistic framework to predict the interaction probability of proteins. The notion of domain combination and domain combination pair is newly introduced and the prediction model in the framework takes domain combination pair as a basic unit of protein interactions to overcome the limitations of the conventional domain pair based prediction systems. The framework largely consists of prediction preparation and service stages. In the prediction preparation stage, two appearance probability matrices, which hold information on appearance frequencies of domain combination pairs in the interacting and non-interacting sets of protein pairs, are constructed. Based on the appearance probability matrix, a probability equation is devised. The equation maps a protein pair to a real number in the range of 0 to 1. Two distributions of interacting and non-interacting set of protein pairs are obtained using the equation. In the prediction service stage, the interaction probability of a protein pair is predicted using the distributions and the equation. The validity of the prediction model is evaluated for the interacting set of protein pairs in Yeast organism and artificially generated non-interacting set of protein pairs. When 80% of the set of interacting protein pairs in DIP database are used as learning set of interacting protein pairs, very high sensitivity(86%) and specificity(56%) are achieved within our framework.

Introduction

인터넷을 통한 단백질 정보의 축적[1, 2, 16]으로 말미암아, 단백질-단백질 간의 상호작용을 계산적으로 예측하는 것이 가능하게 되었다. 단백질-단백질 상호작용을 실험을 통하지 않고 계산적으로 예측함으로써 기대할 수 있는 혜택은 다양하다. 첫째로 기대할 수 있는 장점은 낮은 가격에 대량의 단백질-단백질 상호작용 예측이 가능하다는 점이다. 또한 예측된 정보를 이용하여 생물학자들은 수많은 후보 단백질 중에 실험을

This work is supported by Institute of Information Technology Assessment under Ministry of Information and Communication, Korea

하지 않고도 어떤 단백질부터 실험에 착수할 것인지에 대한 우선 순위 부여가 가능할 것이다. 또한 이 정보들은 장차 미지의 단백질의 기능 예측에 기본적인 데이터로 활용될 수 있다[14].

단백질-단백질 상호작용을 계산적으로 예측하는 데는 여러 가지 접근 방법이 있다[4, 5, 9, 11]. 가공하지 않은 단백질 서열로부터 직접 단백질-단백질 상호작용에 영향을 끼치는 요인들을 발견하고 분석하는 것이 한 가지 접근 방법이며[6], 단백질 3차 구조나 물리화학적 특성을 분석함으로써 단백질 상호작용을 예측하는 방법도 있다[3].

도메인에 기반한 단백질-단백질 상호작용 예측도 또 하나의 접근 방법이 될 수 있으며, 현재 여러 연구진들에 의하여 활발히 연구되어지고 있다[4, 11, 15]. 대부분의 도메인 기반 단백질-단백질 상호작용 예측 모델들은 단백질-단백질 상호 작용이 도메인-도메인 상호작용의 결과물이라는 추측 에서 출발한다. 이 방법들은 단백질-단백질 상호작용 데이터로부터 도메인-도메인 상호작용 정보를 추측하고, 이를 토대로 단백질의 상호작용을 예측하는 것이 일반적이다. 그리고 예전 연구들은 계산의 편의상, 단백질의 상호작용이 독립적으로 발생 하는 단일 도메인 쌍(single domain pair)의 결합에 의해 유발된다고 가정하고 있다. 그러나, 복수의 도메인들이 합동으로 단백질 상호작용에 영향을 미친다고 가정하는 것이 적절할 것으로 판단된다. 이러한 제약성을 극복하기 위하여, 본 논문에서 도메인 조합(domain combination)과 도메인 조합 쌍(domain combinations pair)의 개념을 도입한다. 도메인 조합이란 용어는 하나의 도메인 집합에서 생성 가능한 도메인 부분 집합을 의미한다. 즉, 본 논문에서 제시하는 확률 예측 모델은 단백질-단백질 상호 작용은 복수의 도메인 쌍이나 도메인 조합 간의 상호작용의 결과로 인식하며, *dc-pair*를 단백질 상호작용의 기본 단위로 해석한다.

그 후에는, 상호작용이 있는 단백질 쌍 집합에서 각각 *dc-pair*의 출현 빈도를 세어서 출현 확률 배열 구조에 저장한다. 이 배열을 토대로 단백질-단백질 상호 작용 확률 예측 모델을 구축한다. 본 논문에서 사용한 접근 방법에서는 도메인 쌍에 대한 정보가 *dc-pair* 정보 안에 포함되어 있으므로, 종래의 도메인 쌍에 기반한 방법에 비하여 더욱 포괄적이다. 또한, 종래의 기술은 주로 scoring system에 기반하여 단순히 score 값을 제공하는데 반해서, 본 방법은 상호작용 가능성에 대한 확률 값을 제시한다는 점에서 좀 더 현실적이라고 할 수 있다. 또한 기존의 방법은 단백질 상호작용이 있는 것으로 보고된 단백질 쌍의 집합만을 사용하는 데 반하여, 본 예측 틀은 임의의 상호 작용이 없는 것으로 추정되는 단백질 집합(non-interacting set)에 대한 정보도 같이 사용한다.

예측 모델의 유효성은 효모(yeast)에서 상호 작용이 있는 것으로 알려진 단백질 쌍 집합과 상호작용이 없는 것으로 추정되는 단백질 쌍 집합을 대상으로 검증하였다. DIP 데이터베이스[16, 17]의 상호 작용이 있는 것으로 알려진 단백질 쌍 집합의 80%를 학습 집단으로 사용했을 때, 제안된 예측 시스템은 매우 높은 sensitivity(86%)와 specificity(56%)를 보여 주어 제안된 예측 시스템의 유용성을 입증하였다.

본 논문은 다음과 같이 구성되어 있다. 먼저, related work에서는 단백질-단백질 상

호작용 예측에 관한 관련 연구를 상술한다. prediction framework에서는 예측 시스템 구조를 자세히 설명하고, validation에서는 예측 시스템의 유효성 검증 결과를 기술하며, conclusion에서는 결론을 내리기로 한다.

Related Work

그 동안 많은 다양한 단백질 상호작용 예측 방법들이 제안되어 왔다. Wojcik[15]는 *H. pylori*의 서열 유사도(sequence similarity)와 상호작용하는 도메인(interacting domains) 정보를 이용하여, 상호작용 지도(interaction map)를 작성하였으며, 이 지도를 바탕으로 다른 종인 *E. coli*에서의 interaction map을 예측하는 방법을 제시하였다. Marcotte[9]는 지놈(genome) 정보를 이용하여 단백질의 기능을 예측하는 방법을 제시하였으며, domain fusion method[5, 9]를 고안하였다.

Park[13]은 효모(yeast)의 지놈(genome)과 PDB(Protein Data Bank)[2]를 조사하여, 진화학적으로 관련 있는 도메인 사이 상호작용이라는 관점에서 단백질 도메인 간의 상호작용을 검토하여, SCOP(Structural Classification of Proteins)[10] 패밀리의 상호 작용 타입을 구분하였다.

일반적으로 단백질 구조와 서열의 기본 단위로서 도메인이 알려져 있으며, SCOP, CATH, FSSP와 같은 다양한 분류 시스템에서 도메인의 개념이 사용되고 있다[10, 12, 8]. Bock[3]은 단백질-단백질 상호작용을 예측하기 위하여, 단백질 서열의 물리화학적 특성을 사용하는 방법을 제안하였으며, 아미노산 서열의 물리화학적 특성을 조사하여 DIP[16, 17]의 상호작용 단백질 쌍으로부터 support vector machine을 이용하여 단백질 상호작용을 예측하였다.

Deng[4]은 pfam(http://www.sanger.ac.uk/Software/Pfam/index.shtml) 데이터베이스에 정의된 도메인을 이용하여, 도메인 쌍 간의 상호작용 확률을 추정하였다. Deng은 maximum likelihood estimation을 적용하여, 관측된 단백질-단백질 상호작용과 일치하는 상호작용 도메인을 추론하였다. Ng[11]는 DIP 데이터베이스와 protein complex, Rosetta Stone sequence 등의 데이터를 종합하여 도메인-도메인 상호작용을 유추하였으며, InterDom이라는 데이터베이스를 개발하였다(http://interdom.lit.org.sg.). Goffard[7]은 단백질 상호 작용 유추를 위한 웹기반 서버인 IPPRED를 개발하였으며, 만일 protein A와 B 간의 상호작용을 알고 싶은 경우에, 기존에 상호작용이 알려진 protein C, D와의 유사성을 조사한 다음 C와 D와 A, B가 각각 유사성이 있는 경우에 protein A와 B 사이에 서로 상호작용이 있다고 유추하였다.

Prediction Framework

Domain Combination and Domain Combination Pair

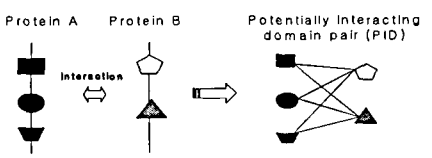
본 논문에서 제안하고 있는 예측 모델을 설명하기 전에 도메인 조합(Domain Combination)과 도메인 조합 쌍(Domain Combination Pair)의 개념을 설명한다. 기술의 편의상 Domain Combination은 간단히 *dc*, Domain Combination Pair는 *dc-pair*로 표기하기로 한다. 어떤 단백질 p가 복수의 도메인을 가지고 있다면, 도메인 조합은 단백질 p의 도메인 집합으로부터 만들어질 수 있는 모든 가능한 도메인 그룹이 된다. 여기서 그룹은 적어도 하나의 도메인을 반드시 포함하는 것으로 한다. 즉, 단백질 p의 모든 가능한 도메인 조합의 집합은 다음과 같이 정의된다.

$$dc(p) = PowerSet(domain(p)) - \varnothing \quad (1)$$

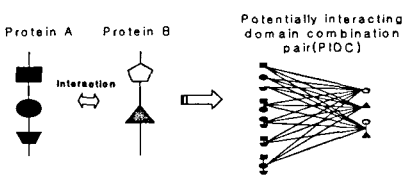
여기에서, *domain(p)*는 protein *p*의 도메인 집합을 나타낸다. 식 (1)에서 공집합이 제거되므로 단백질이 *n*개의 서로 다른 domain을 가지고 있다면, $2n-1$개의 도메인 조합이 얻어진다. 본 논문에서 제시하는 예측 모델에서는 도메인 조합 쌍(domain combination pair)을 단백질 상호작용의 기본 단위로 간주하며, 동일 단백질 안의 하나 이상의 복수의 도메인 조합 쌍이 연합하여 단백질 상호작용에 영향을 주는 것으로 가정한다. 두 단백질 *p*, *q*에서 모든 가능한 도메인 조합 쌍의 집합의 정의는 다음과 같다.

$$dc\text{-}pair(p,q) = \{<dc_1,dc_2>| <dc_1,dc_2> \in dc(p) \times dc(q) \text{ or } dc(q) \times dc(p) \text{ where, } dc_1, dc_2 \in dc(p) \text{ or } dc(q)\} \quad (2)$$

두 단백질 *p*, *q*가 각각 *n*, *m*개의 다른 도메인을 가지고 있을 경우, $2n$-1, $2m$-1개의 서로 다른 *dc-pair*를 얻게 된다. 그림 1(b) 에서는 각각 3개와 2개의 도메인을 갖는 단백질이 상호작용하는 경우, 잠재적인 상호작용 *dc-pair*를 보여주고 있다.



(a)



(b)

그림 1 도메인 쌍에 기반한 기존의 예측 접근 방법(a)과 도메인 조합에 기반한 새로운 예측 방법(b)

또한, 기존 방법에서 사용했던 도메인 쌍 기반 방법과 도메인 조합 쌍 기반 접근 방법의 차이점을 보여준다. 그러나, 이것은 두 단백질 간의 상호작용이 관찰된 경우에라도 어떤 *dc-pair*들이 상호작용을 일으키는 결정적인 역할을 담당하는 지에 관한 충분한 정보를 제공하지는 않는다. 향후 인터넷을 통한 상호작용 단백질 쌍의 정보가 축적되면, 중요한 *dc-pair*를 추출하는 것이 가능할 것으로 예상된다.

The Big Picture

본 논문에서 제안된 예측 시스템은 크게 두 과정으로 구성된다. 그림 2는 본 예측 시스템의 전체 구조를 보여준다. 첫 번째 과정은 예측을 준비하는 과정이며 두 번째 과정에서는 예측을 수행하는 과정이다.

예측 준비 과정은 다시 세 개의 단계를 포함한다. 첫 번째 단계에서는 상호작용이 있는 것으로 알려진 단백질 쌍 집합과 상호작용이 없는 것으로 추정되는 단백질 도메인 쌍 집합으로부터 각각 도메인 조합 정보와 그 출현 빈도를 추출한다. 이 정보들은 출현 확률 배열(Appearance Probability matrix;AP matrix)라고 불리는 배열 구조에 저장된다. 두 번째 단계에서는 AP matrix를 기반으로 단백질-단백질 상호 작용 예측식을 정의한다. 이 확률식은 미 정의된 상수를 포함하게 되며 maximum likelihood estimation 적용을 통하여 결정한다. 마지막 세 번째 단계에서는 상호작용이 있는 것으로 알려진 단백질 쌍 집합과 상호작용이 없는 것으로 추정되는 단백질 도메인 쌍 집합의

PIP(Primary Interaction Probability) 값 분포를 얻게 된다.

두 번째 과정에서는 첫 번째 과정에서 얻어진 분포에 기초하여, 단백질-단백질 상호작용을 예측하는 또 다른 확률식이 정의되며, 이 확률식을 이용하여 확률을 계산한다. 이 확률은 단백질-단백질 상호작용을 예측하는 최종 확률이다. 각 단계의 세부 사항은 다음 절에서 설명한다.

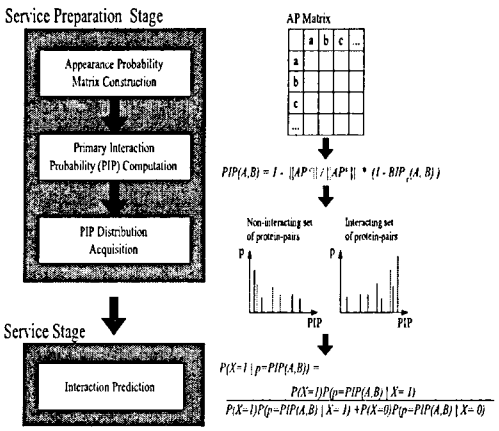


그림 2 본 예측 틀의 전체 구조

AP Matrix

예측 준비 과정의 첫 과정으로서 출현 빈도 배열을 생성한다. 주어진 단백질 쌍 집합에서, n개의 서로 다른 단백질 $\{p_1,p_2,\ldots, p_n\}$이 있을 때, 단백질의 도메인 조합은, $\{dc_1,dc_2,\ldots,dc_m\}$이 되며 $dc(p_1)$, $dc(p_2)$, $dc(p_n)$의 합집합은 m개의 서로 다른 도메인 조합 $\{dc_1,dc_2,\ldots, dc_m\}$을 구성하게 되어, $m\times m$ AP matrix가 생성된다. 배열에서 원소 AP_{ij}는 주어진 단백질 쌍 집합에서 도메인 조합 $<dc_i, dc_j>$ 출현 확률을 대표한다.

AP matrix를 생성하기 위하여, 먼저 WF (Weighted Frequency) 배열을 먼저 생성한다. 이 때 각 열과 행은 도메인 조합을 나타내며, 배열의 각 원소는 *dc-pair*를 나타낸다. WF matrix에서는, 주어진 단백질 쌍의 집합에서의 도메인 조합 출현 빈도가 등록된다. WF matrix에서 원소 WF_{ab}는 도메인 조합 $<a, b>$의 weighted appearance frequency를 가지게 되며, 다음 식에 의하여 계산된다.

$$\sum_{\substack{\forall(p_i,q_j)\text{ such that}\\ <a,b>\in dc\text{-pair}(p_i,q_j)}} \frac{1}{|dc(p_i)|\times|dc(q_j)|} \quad (3)$$

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 최종결과가 계산된다. 식 (3)에 의해서, *dc-pair* $<a, b>$의 잠재적인 기여 가중치가 계산된다. 가중치 부여의 의미는 상호 작용하는 단백질 쌍으로부터 얻어지는 가능한 도메인 조합 쌍의 수가 적으면 적을수록, 각 *dc-pair*에 의한 상호작용에서의 기여도는 더 클 것이라는 가정에서 출발한다. *dc-pair*의 출현 빈도에 가중치를 주는 방법에 대하여 다른 많은 방법이 있겠지만, 이 논문에서는 이것에 관한 더 이상의 논의는 생략하기로 한다.

식 (3)을 $\{<A,B>, <A,C>, <B,C>\}$로 주어진 단백질 쌍 집합의 각 단백질의 도메인이 $domain(A) = \{a_1,a_2\}$, $domain(B) = \{b_1\}$, $domain(C) = \{a_1,c_1\}$로 구성되어 있는 예에 적용하면, 도메인 조합 $\{<\{b_1\},\{a_2\}>\}$은 *dc-pair* (A,B)에서만 출현하므로 $WF_{\{b1\},\{a2\}}$의 값은 $1/(|dc(B)|\times|dc(A)|)$가 된다.

$dc(A) = \{\{a_1\},\{a_2\},\{a_1,a_2\}\}$, $dc(B) = \{\{b_1\}\}$, $|dc(A)| = 3$, $|dc(B)| = 1$이므로, $1/(|dc(B)|\times|dc(A)|)$의 값은 1/3이 된다. 이런 방법으로 WF matrix 모든 요소들을 계산할 수 있다. WF matrix가 생성된 후에 AP matrix의 각 원소 값의 계산은 다음 식에 따른다.

$$AP_{ij}=\frac{WF_{ij}}{\sum_{i,j} WF_{ij}} \quad (4)$$

이와 같이 얻어진 배열의 각 원소 값은 특정 도메인 조합이 해당 공간에서 출현할 확률을 나타내게 된다. 상호작용이 있는 것으로 알려진 단백질 쌍 집합과 상호작용이 없는 것으로 추정되는 단백질 도메인 쌍 집합에 대하여 각각의 AP matrix를 얻을 수 있다. 두 배열의 상당한 부분이 서로 겹쳐지지만, 형태가 꼭 일치할 필요는 없다. 그리고 상호작용이 있는 것으로 알려진 단백질 쌍 그룹과 상호 작용이 없는 것으로 추정되는 그룹 각각에 대해서 출현 확률을 구할 수 있으므로 각각의 출현 확률 배열을 AP^i, AP^r 배열로 표시하고 이들의 공통 부분 $AP^i \cap AP^r$은 AP^c 배열로 나타낸다.

Primary Interaction Probability

두 번째 단계에서는 첫 과정에서 얻어진 두 개의 출현 확률 배열을 기반으로, 상호작용을 모르는 단백질 쌍 <*A,B*>에 대한 확률을 예측하는 확률식이 정의되며, 이 확률식에 포함되는 미지의 상수가 결정된다. 먼저, 단백질 쌍 <*A,B*>로부터 식 (2)를 이용하여, 이들의 도메인 조합 *dc-pair*들을 산출한다. 많은 *dc-pair*들이 만들어지며, 다음과 같이 *dc-pair*를 분류한다

- $DC_c(A,B)$ = {*dc-pair* | *dc-pair* ∈ *dc-pair* (*A,B*) and appears in AP^c *dc-pair* space}
- $DC_{r-c}(A,B)$ = {*dc-pair* | *dc-pair* ∈ *dc-pair* (*A,B*) and appears in $AP^r - AP^c$ space}
- $DC_{i-c}(A,B)$ = {*dc-pair* | *dc-pair* ∈ *dc-pair* (*A,B*) and appears in $AP^i - AP^c$ space}

그림 3은 AP^i, AP^r 공간에서 *dc-pair*(*A,B*)가 만들어질 때, 각 원소들이 어느 카테고리에 속하는지를 보여준다. *dc-pair*(*A,B*)의 각 원소들은 특수 기호(*, Δ, ×)로 표시된다.

AP^c *dc-pair* 공간에서 발견되는 *DCc*(*A,B*) 도메인 조합을 대상으로 상호작용 확률식을 다음과 같이 정의할 수 있다. 이 확률은 *DCc*(*A,B*)가 AP^c *dc-pair* 공간에서 발견될 때 단

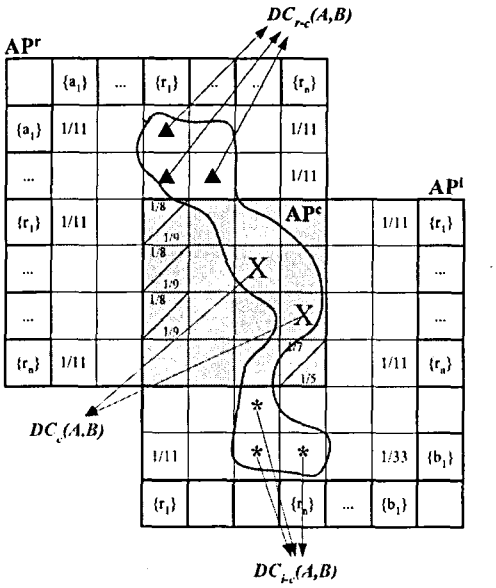


그림 3 도메인 조합 카테고리

단백질 쌍 <*A, B*>가 서로 상호작용할 확률을 의미한다. 상호작용이 일어나는 사건과 일어나지 않는 사건을 표현하기 위하여 확률 변수 X를 도입하였다. 1 값은 상호작용이 일어나는 사건, 0 값은 상호작용이 없는 사건을 나타낸다.

$$P(X=1 \mid DC_c(A,B)) = \frac{P(X=1)P(DC_c(A,B)|X=1)}{P(X=1)p(DC_c(A,B)|X=1)+P(X=0)P(DC_c(A,B)|X=0)} \quad (5)$$

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

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

$$P(X=0) =$$

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

$$P(DC_c(A,B)\mid X=1)=$$

$$|DC_c(A,B)|!\cdot\prod_{\{i,j\}(i,j)\in DC(A,B)\}}\frac{(AP_I^c)_{ij}}{\sum_{i,j}(AP_I^c)_{ij}},$$

$$P(DC_c(A,B)\mid X=0)=$$

$$|DC_c(A,B)|!\cdot\prod_{\{i,j\}(i,j)\in DC(A,B)\}}\frac{(AP_R^c)_{ij}}{\sum_{i,j}(AP_R^c)_{ij}}$$

이 때, $P(X=1)$는 AP^c에 존재하는 총 *dc-pair* 공간에서 상호작용이 있는 단백질 쌍으로부터 만들어진 *dc-pair* 공간을 나타내며, $P(X=0)$은 AP^c의 도메인 조합 공간에서 상호작용이 없다고 추정되는 단백질 쌍으로부터 생성된 *dc-pair* 공간을 나타낸다. 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 maximum likelihood estimation 적용을 통하여 결정한다. $P(DCc(A,B)\mid X=1)$는 AP^i 공간에서 $DCc(A,B)$에 속하는 *dc-pair* 집합이 만들어질 확률이고, $P(DCc(A,B)\mid X=0)$는 AP^r 공간에서 $DCc(A,B)$에 속하는 *dc-pairs* 집합이 만들어질 확률이다. AP_I^c와 AP_R^c는 각각 상호작용이 있는 *dc-pair* 공간과 상호작용이 없는 것으로 간주되고 있는 *dc-pair* 공간에서의 AP^c를 각각 의미한다. 동일하게, $DCi\text{-}c(A,B)$ 도메인 조합을 대상으로 얻어질 확률식은 다음과 같다.

$$P(X=1\mid DC_{i\text{-}c}(A,B))=\frac{P(X=1)P(DC_{i\text{-}c}(A,B)\mid X=1)}{P(X=1)p(DC_{i\text{-}c}(A,B)\mid X=1)+P(X=0)P(DC_{i\text{-}c}(A,B)\mid X=0)} \quad (6)$$

위 식에서, $P(X=1)$, $P(X=0)$는 각각 1, 0이 되어 최종적으로 얻어지는 확률은 1이 된다. 식 (5), (6)을 이용하여, $DC_c(A,B)$ *dc-pairs*를 갖는 (A,B) 단백질 쌍의 상호작용 가능성 확률(Primary Interaction Probability;PIP)은 다음과 같다.

$$\mathrm{PIP(A,B)}=1-\frac{\|AP^c\|}{\|AP^i\|}(1-P(X=1\mid DC_c(A,B)) \quad (7)$$

PIP 분포와 상호작용 예측

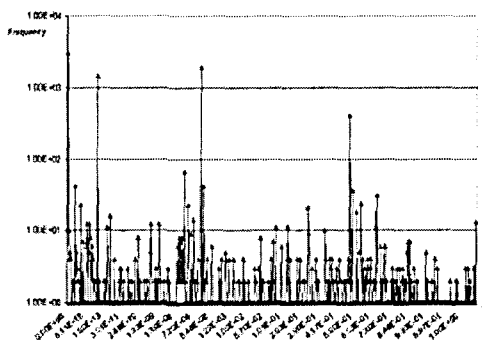
일단 두 번째 단계에서 PIP 최종식이 얻어지면, 식 (7)에 따라 상호작용이 있는 단백질 쌍과 없는 것으로 간주된 쌍 집합에 대한 PIP 값을 계산할 수 있다. 모든 PIP 값이 계산되면 PIP 분포를 얻을 수 있고, 두 집합을 비교하기 위하여, 분포를 정규화한다. 한편 PIP 함수는 단백질 쌍을 실수 0~1 범위 안에 투사시키는 함수의 일종으로 해석할 수 있다. PIP 분포가 얻어지면, 이 분포에 대한 2-카테고리 분류(two category classification) 기법 적용이 가능하다. 즉, 임의로 주어진 단백질 쌍에 대하여, 그들이 상호작용을 할 가능성을 예측하기 위해서는 그 단백질 쌍의 PIP 값이 어느 분포에 속할지를 결정해야 한다.

Validation

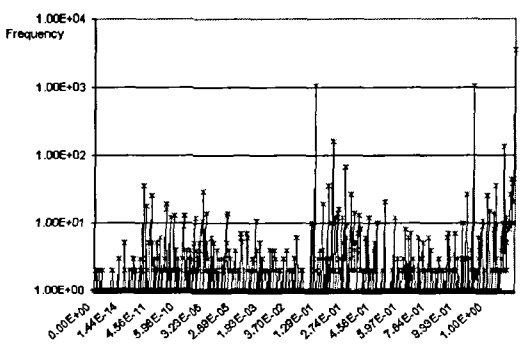
제안된 예측시스템의 검증을 위하여, 다음과 같은 2개의 단백질 쌍 데이터를 준비하였다. 상호작용이 알려진 단백질 쌍 집합은 DIP의 yeast에서 총 15,174개의 상호 작용이 보고된 단백질 쌍을 준비하였다.

상호작용이 없다고 추정되는 단백질 쌍은 도메인 정보가 알려진 단백질 쌍 집단에서, 상호작용이 알려진 단백질 쌍 집단을 제거하는 방식으로, 임의로 생성되었다. 입증의

편의를 위하여, 상호작용이 없는 것으로 추정된 단백질 쌍의 경우에는 상호작용이 보고된 단백질 쌍과 같은 15,174개의 단백질 쌍을 준비하였다. 그럼에도 불구하고, 이 집단 안에 상호작용이 있는 단백질 쌍이 완전히 제거된 상태는 아니다. 하지만 만일 전체 단백질 쌍 공간 안에 상호작용하는 단백질 쌍이 아주 드물다고 추측한다면, 본 예측 모델에서 사용된 상호작용이 없다고 추정되는 집단으로도 충분할 것으로 예상된다. 이상의 방법으로 2개의 집단을 준비한 후, 각각을 학습 집단과 검증 집단으로 나누었다. 학습 집단으로 상호작용이 있는 것으로 알려진 전체 단백질 쌍의 80%를 사용했을 때, 12861*12861 크기의 AP^1와 14470*14470 크기의 AP^2이 생성되었다. 그림 4는 두 집단을 대상으로 한 PIP 값의 분포를 보여 주고 있다.



(a)



(b)

그림 4 상호작용이 없다고 추정되는 단백질 쌍에 대한 PIP 값의 분포 (a)와 상호작용이 있는 단백질 쌍의 PIP 분포(b) (log scale)

각 집단의 PIP 값은 0~1 사이에 중복되어 위치한다. 그러나 상호작용이 보고된 집단의 PIP 값들은 대부분 1 가까이 있으며 상호작용이 없다고 추정되는 집단의 PIP 값은 0 주위에 위치한다. 이것은 PIP 값이 상호작용이 보고된 집단과 상호작용이 없다고 추정되는 집단을 나누는 좋은 분류자(classifier)가 됨을 나타낸다. PIP 값의 분포를 다양한 2-카테고리 분류(2-category classification)방식을 적용하여 분류할 수 있다. 본 논문에서는 예측 모델의 유효성을 검사하기 위하여, hybrid classification을 고안하였으며, 에러 확률식은 다음과 같다.

$$P(e) = \sum_{\{i,j|PIP_i^x = PIP_j^y\}} Min[p_i^x, p_j^y], \qquad (8)$$

$$P_i^x = \frac{freq_i^x}{\sum_{i=1}^{m} freq_i^x}, \quad P_j^y = \frac{freq_i^y}{\sum_{j=1}^{m} freq_i^y}$$

고로, 에러 확률 $P(e)$는 두 집단 간의 PIP 값이 중복되는 경우가 적을수록 감소한다. 본 모델의 유효성을 테스트하기 위하여 베이즈 규칙(*Bayes rule*)을 사용하여 sensitivity와 specificity를 측정하였다. 상호작용이 알려진 단백질 쌍과 없다고 추정 되는 단백질 쌍을 검증집단으로 이용하여, 3번 반복으로 테스트하여 보았으며, 상호작용이 알려진 단백질 쌍 전체 수 중 80%의 단백질 쌍을 학습 집단으로 사용하였을 때 86%의 sensitivity와 56%의 specificity가 얻어졌다. 현재 보고된 상호 작용이 있다고 보고된 단백질 쌍 집단과 본 논문에서 임의

로 생성된 상호작용이 없다고 추정되는 단백질 쌍 안에는 실험적인 에러 데이터가 포함되어 있을 수 있으므로, 이 점을 고려한다면, 이 결과로 본 예측 모델이 유효하다고 결론지을 수 있다.

Conclusion

본 연구에서는 단백질-단백질 상호작용을 예측하는 확률 시스템을 제안하였으며, 유효성 테스트를 실시하였다. 제안된 확률 틀에서는 단백질의 상호작용 기본 단위로서 *dc-pair*를 채택하였으며, 확률식 PIP은 단백질 쌍을 실수 0~1 범위에 투사시킴으로써, 그 분류 능력이 증명되었다.

인터넷을 통한 단백질 상호작용 데이터가 축적될수록 본 예측 틀의 예측 능력은 더 향상될 것이라 기대된다. 제안된 예측 틀의 효과는 4가지로 요약할 수 있다. 첫째로, 본 예측 틀을 이용하여, 생물학자로 하여금, 많은 비용과 시간이 소요되는 단백질 상호작용 실험을 통하지 않고 단백질 상호작용에 대해서 시간과 비용 측면에서 획기적인 기여를 할 것으로 기대된다. 둘째 본 틀과 같은 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. 셋째 본 시스템은 미지의 단백질에 대한 기능을 추정하는 것과 같은 단백질 동정(identification)시에 기본적인 계산적 접근방법으로 활용될 수 있다. 넷째 본 연구에서 제안하고 있는 예측 틀은 생물학자들이 그들의 연구 분야에서 유사한 경우를 만났을 때 참고 모델로 이용될 수 있다. 향후에는 쥐와 인간과 같은 다른 종의 단백질 집단에 본 예측 틀을 적용할 예정이다. 다음 단계에는 단백질 상호작용 네트워크 구축이나 예측된 상호작용 데이터에 기반한 시각화(visualization)를 통하여 생물학자들이 객관적으로 유용한 단백질 정보를 손쉽게 추출할 수 있도록 할 계획이다.

References

[1] 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, 2001, 37-40

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, The Protein Data Bank. *Nucleic Acids Res.*, 28, 2000, 235-242

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

[4] M. Deng, S. Metah, F. Sun and T. Chen, Inferring Domain-Domain Interactions from Protein-Protein Interactions. *Genome Research*, 12, 2002, 1540-1548

[5] 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, 1999, 86-90

[6] 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

[7] N. Goffard, V. Garcia, F. Iragne, A. Groppi and A. de Daruvar, IPPRED: Server for Proteins Interactions Inference. *Bioinformatics*, 19, 2003, 903-904

[8] L. Holm, and C. Sander, The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.*, 24, 1996, 206-210

[9] 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, 1999, 751-753

[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, 1995, 536-540

[11] S. Ng, Z. Zhang and S. Tan, Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19, 2003, 923-929

[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 Res.*, 28, 2000, 277-282

[13] 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, 2001, 929-938

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

[15] J. Wojcik and V. Schächter, Protein-Protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17 Sup pl., 2001, S296-S305

[16] I. Xenarios and D. Eisenberg, Protein interaction databases. *Curr. Opinion in Biotechnology*, 12, 2001, 334-339

[17] I. Xenarios, E. Fernandez, L. Salwinski, X. J. Duan, M. J. Thompson, E. M. Marcotte and D. Eisenberg, DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res.*, 29, 2001, 239-241