단백질 기능 예측을 위한 그래프 기반 모델링

황 두 성[†] · 정 재 영[††]

요 약

단백질 상호작용 데이터는 현 생물정보학에서 기능이 알려져 있지 않은 단백질의 기능 예측에 높은 신뢰성이 있는 프로티오믹스의 계산 모델에 이용되고 있다. 단백질 기능 예측 관련 연구로는 guilt-by-association 개념을 바탕으로 대규모의 단순 2차원 단백질-단백질 상호작용 맵을 이용하고 있다. 본 논문에서는 단백질-단백질 상호작용 데이터를 이용한 그래프 기반 기능 예측 방법인 neighbor-counting, x^2-통계치 예측 모델을 살펴보고 대량의 상호작용 데이터로부터 빠른 기능예측에 효과적인 알고리즘을 제안한다. 제안하는 알고리즘은 단백질 상호작용 맵, 서열 유사성 및 경험적 전문가 지식을 이용하는 그래프 기반 모델이다. 제안된 알고리즘은 Yeast 단백질의 기능 예측을 수행하였으며, neighbor-counting, x^2-통계치 모델의 실험 결과와 비교되었다.

Graph-based modeling for protein function prediction

Doosung Hwang[†] · Jae-Young Jung[††]

ABSTRACT

The use of protein interaction data is highly reliable for predicting functions to proteins without function in proteomics study. The computational studies on protein function prediction are mostly based on the concept of guilt-by-association and utilize large-scale interaction map from revealed protein-protein interaction data. This study compares graph-based approaches such as neighbor-counting and x^2-statistics methods using protein-protein interaction data and proposes an approach that is effective in analyzing large-scale protein interaction data. The proposed approach is also based protein interaction map but sequence similarity and heuristic knowledge to make prediction results more reliable. The test result of the proposed approach is given for KDD Cup 2001 competition data along with those of neighbor-counting and x^2-statistics methods.

키워드 : 프로티오믹스(proteomics), 단백질 기능 예측(protein function prediction), 그래프 기반 모델(graph-based model)

1. 서 론

계산 모델 및 컴퓨터 하드웨어 기술의 발달이 유전체 관련 데이터의 빠른 코딩 및 관련 연구에 이점을 주었으나 대량의 유전체 데이터로부터 관련된 프로티오믹스(proteomics) 데이터의 양은 더욱 많아질 것으로 예측되며, 데이터 분석 및 이해에 있어 새로운 계산 이론의 도입 및 응용이 요구되고 있다[1]. 유전체학(genomics)에 관한 연구는 DNA로부터 유전자(gene)의 확인(identification), 발견(discovery), 서열(sequence) 등을 포함하며, 이렇게 얻어진 유전자들의 기능 이해를 통한 치유가 어려운 병과의 관계를 규명하려는 연구가 기능적 유전체학(functional genomics) 또는 프로티오믹스라 한다[2].

생명의 기본 정보인 유전체의 서열 정보가 밝혀짐에 따라 코드화 된 유전인자로부터 발현(expression)되는 단백질을 포함한 관련 분자들에 대한 정보의 이용이 가능하게 되었다. 이에따라 포스트-지놈 생명 정보학(post-genomic bioinformatics)의 궁극적인 연구 목표는 세포 내에서 발생하는 복잡한 생물학적 기능을 분자들 간의 상호작용(interaction)으로 밝히고, 분자 네트웍(molecular network)을 구성함으로써 세포 내 분자의 기능(function)을 이해하는데 있다고 한다[3,4,8]. 단백질-단백질 상호작용은 cell cycle, DNA 복제(replication) 및 전사(transcription), 대사(metabolism), 신호전달(signal transduction) 등에 중요한 역할을 함으로, 특정 단백질에 대한 생물학적 이해는 세포 내 분자들 간의 생화학적 관점에서 분자들 간의 상호작용으로부터 해당 기능을 밝히는데 선행되어야 한다. 세포 내 여러 종류의 분자들 중에서 단백질 상호작용 데이터는 고 성능 실험 기법(high-throughput technology)들이 개발·이용되어 실험으로부터 대규모(large scale) 데이터를 생산하게 되었다.

단백질-단백질 상호작용 데이터 분석에 있어 2차원 상호작용 맵(interaction map)의 이용은 그래프 이론(graph theory)을 응용하여 기능이 알려지지 않은 단백질(protein with-

† 종신회원 : 단국대학교 컴퓨터과학과 교수
†† 정 회 원 : 한국전자통신연구원 선임연구원
논문접수 : 2004년 6월 26일, 심사완료 : 2005년 2월 19일

out functions)의 기능에 대한 이해의 폭을 넓히고 있다[3,4,5]. 또한 단백질 상호작용 맵은 기능이 알려지지 않은 단백질들의 기능을 상호작용 데이터를 이용한 기능 예측 및 단백질 복합체(protein complex)등 이해에 주요한 분석 도구로 인식되고 있다[6]. 계산이론의 지능형 알고리즘(artificial intelligent algorithm)들은 단백질 기능 및 세포내 소기관(localization) 분류 및 예측에 응용되었다[7]. KDD Cup 2001 및 2002에서 생물정보학 문제에서 지능형 학습 알고리즘들의 경쟁을 통한 적응성 및 응용성 여부가 도전 되어 지능형 알고리즘의 프로티오믹스 데이터의 분석에 효과가 있음을 보였다.

단백질-단백질 상호작용 데이터는 기능이 알려지지 않은 단백질의 기능을 예측하는데 유용하다. 그러므로 지금까지 알려진 단백질 상호작용 데이터에 계산이론을 적용한 상단백질의 기능 예측 모델의 개발은 단백질의 기능상 분류 및 새로운 실험에 비용을 줄여 줄 것으로 기대된다. 본 논문에서는 단백질-단백질 상호작용 데이터를 이용한 기능이 알려지지 않은 단백질의 기능 예측을 위한 계산모델에 대한 방법론에 대해 살펴보고 전문가의 경험적 지식 및 유전자의 서열 유사성(sequence similarity)을 이용한 단백질 기능 예측 알고리즘을 제안한다.

2. 관련연구

단백질 기능 예측의 모델링은 단백질을 발현한 유전인자의 서열, 표현형(phenotype, [9]) 및 구조(structure)의 서열 유사성 정보를 이용한 방법과 및 단백질-단백질 상호작용 데이터를 이용하는 방법 등이 제안되었다. 서열 유사성을 이용한 방법보다는 단백질 간의 상호작용 여부를 이용한 방법은 높은 신뢰성(reliability)을 갖는다고 알려져 있다[8]. 고성능 실험 기법(high-throughput experiment technology, [5])을 이용하게 됨에 따라 특정 유기체의 단백질-단백질 상호작용에 대한 대량의 데이터의 생성이 가능하게 되어, 실험실에서는 상호작용 데이터로부터 기능이 알려지지 않은 단백질의 기능예측에 대한 계산모델이 필요하게 되었다. 단백질 기능 예측에 적용되는 기본 개념은 "guilt-by-association" 으로 어떤 단백질이 무슨 기능을 할 것이냐에 대한 단서는 이미 기능을 알고 있는 다른 단백질과 어떻게 상호작용을 하는지 여부를 살펴봄으로써 알 수 있다는 것으로, 단백질의 상대적인 양이 아닌 단백질-단백질 상호간의 잠재적인 상호작용(potential interaction)을 살펴보는 방법을 이용한다.

그래프 기반 계산 이론의 응용은 단백질의 기능 예측 문제에서 기능이 알려지지 않은 단백질의 기능 예측에 적합한 모델이 될 수 있으며, 각 단백질의 기능 예측은 국소적 상호작용(local interaction) 데이터를 이용하기 때문에 기계학습 측면에서 인스턴트 기반 학습(instance-based learning, [10])이라고 할 수 있다. 그래프를 이용한 예측 모델의 적용은 neighbor-counting 모델[3], x^2-통계치를 이용한 모델[11], Markov random field 모델[12] 등이 있다. 또한 학습이론 기반 응용된 모델로는 SVM(support vector machine), 인공신경망(artificial neural network), Bayesian 모델의 응용 등이 제안되었다[7].

Neighbor-counting(NC) 모델은 Yeast 단백질 데이터에 대한 대규모 단백질 기능 예측을 위한 방법론으로 응용되었다. 단백질-단백질 상호작용 데이터 및 밝혀진 기능을 이용 기능이 알려지지 않은 단백질에 대해 직접 상호작용이 있는 기능이 알려진 단백질의 기능 데이터를 이용하여 예측하는 방법이다[2,4,11]. 직접 상호작용(direct interaction) 단백질에 대한 기능의 빈도수를 계산하여 k 개의 큰 빈도수(k-largest frequencies)를 갖는 기능을 단백질의 기능으로 예측한다.

x^2-통계치을 이용한 기능 예측 방법론은 이웃한 false positive 상호작용 단백질에 대한 통계 데이터를 예측 모델에 이용한다. 단백질 상호작용을 밝히는 체계적 실험기술 발달은 대량의 단백질-단백질 상호작용 데이터를 짧은 시간 안에 생성시킬 수 있으나, 이 실험은 많은 false positive 데이터를 포함하게 된다. 그러나 이러한 실험으로부터 관심있는 유전인자가 생물학적 의미에서 필수인가에 대한 단서를 얻을 수 있으며 상세 실험의 진행에 있어 중요한 근거이다. 특정 유기체의 단백질-단백질 상호작용 맵의 작성은 이러한 실험의 궁극적인 목표이다[5,6]. 그러므로 단백질 상호작용을 이용한 예측 모델은 guilt-by-association 뿐만 아니라 실험에서 발생한 false positive인 상호작용 데이터를 분별할 수 있는 모델이 요구된다[14]. x^2-통계치를 이용한 기능 예측 방법론은 이웃한 false positive 상호작용 단백질에 대한 통계 데이터를 예측 모델에 이용한다. 단백질 상호작용을 밝히는 체계적 실험기술 발달을 통해 알려진 하나의 단백질은 여러 개의 기능 및 기능별 바인딩 파트너(binding partner)를 가질 수 있다는 사실을 모델에 응용한다. 간접적인 상호작용(direct or indirect interaction) 정보를 예측 계산 모델에 사용하여 단백질 상호작용 그래프로부터 위상(topology) 정보를 이용한 정해진 상호작용 깊이(depth)내에서 기능별 x^2-통계치를 이용하여 기능을 부여하는 방법이다.

살펴본 그래프 중심의 단백질 기능 예측 모델은 실험에서 밝혀진 단백질-단백질 상호작용 데이터가 예측 성능을 결정할 것으로 기대되며, 대규모 상호작용 맵, 참조 데이터 및 테스트 데이터의 효과적인 자료구조의 이용이 모델 구현에서 고려되어야 한다. 상호작용 맵의 크기는 알려진 단백질을 모두 포함하고 있어야 함으로 유기체의 단백질의 수에 따라 맵이 결정된다. 모델 응용 시 테스트 데이터에 대해 사전에 밝혀진 단백질의 속성 및 관련 정보 등이 포함된다면 예측 성능의 향상에 이점이 있을 수 있다고 기대된다.

3. 단백질 기능 예측을 위한 데이터

KDD Cup 2001 Competition[7]에서 사용된 Yeast 단백질의 기능 예측을 위한 데이터 셋은 MIPS 데이터베이스[13]로부터 만들어 졌으며 각 단백질은 6개의 속성(essential, class, complex, phenotype, chromosome, motif) 및 기능

(function)과 소기관에 대한 속성 데이터를 포함하고 있다. 그리고 단백질-단백질 상호작용(protein-protein interaction)의 데이터는 상호작용이 밝혀진 두 단백질 및 상호작용 형태(interaction type)와 expression correlation으로 주어진다. 두 단백질의 상호작용 여부는 Yeast two-hybrid와 immuno-coprecipiation의 실험들로부터 수집되었다. Yeast two-hybrid은 *in vitro* 실험으로써 물리적 상호작용(physical interaction,[5]) 의미하며, immuno-coprecipiation의 *in vitro* 실험은 유전적 상호작용(genetic interaction, [4])으로 기술된다. 주어진 문제는 계산 모델을 응용하여 Yeast 단백질의 기능이 알려진 단백질(protein with function)을 이용하여 기능이 알려지지 않은 단백질(protein without function)의 기능을 예측하는 것이다. 각 문제에 대해 예측 모델 적용 시 학습 데이터에 기능과 세포내 소기관을 같이 이용할 수 없다고 제한되었다. 주어진 단백질의 기능과 설명은 표 1에 설명되었다. 각 적용된 알고리즘들은 단백질-단백질 상호작용 및 주어진 속성들로부터 기능이 알려지지 않은 단백질의 기능을 예측하게 된다.

〈표 1〉 KDD Cup 2001 데이터의 속성

속성	값의 수	내용
Essential	4	세포가 살아 기능하는데 필수적인 단백질의 구분
Class	24	구체적인 세부 기능을 표현
Complex	52	3차원 구조의 단백질이 복합체를 만들어 갖는 4차원 구조
Phenotype	12	단백질의 표현형
Motif	236	단백질의 3차원 구조
Chromosome	17	Chromosome 번호
Function	14	단백질 기능의 분류
Localization	15	단백질이 기능을 하는 세포내의 소기관
Interaction Type	3	유전적 또는 물리적 실험에 의해 발견된 상호작용
exp_corr		서열 유사성 수치

주어진 데이터로부터 학습 데이터(training data)에 나타나는 유일한 단백질들의 수는 862개 이며, 테스트 데이터(test data)에 나타나는 수는 381개이다. 단백질의 이름을 고려하지 않고 속성 및 고유 기능을 갖은 단백질의 수는 4,345 학습 단백질 및 1,928 테스트 단백질들이 있다. 주어진 단백질의 수보다 데이터 수가 많은 이유는 하나의 단백질은 여러 개의 속성을 가질 수 있기 때문이다. 학습데이터에서 주어진 상호작용 데이터의 수는 학습에서 910개이며, 태스트에서 896개이다. 하나의 단백질이 여러 개의 기능을 가질 수 있어, 학습 데이터로부터 각 단백질은 평균 2.56개의 기능이 나타나고 있다.

4. 전문가 지식을 이용한 모델 제안

제안하는 모델은 실험상에서 나타나는 false-positive 상호작용에 대한 정량화 및 expression correlation을 이용하는 기능 예측모델이다. 상호작용에 대한 정량화는 interaction *generality*를 이용하였으며 *expression correlation*은 전이 관계(transitivity relationship)에 의한 새로운 상호작용을 추가하는데 이용한다. 제안하는 알고리즘의 설명은 다음 기호를 이용한다.

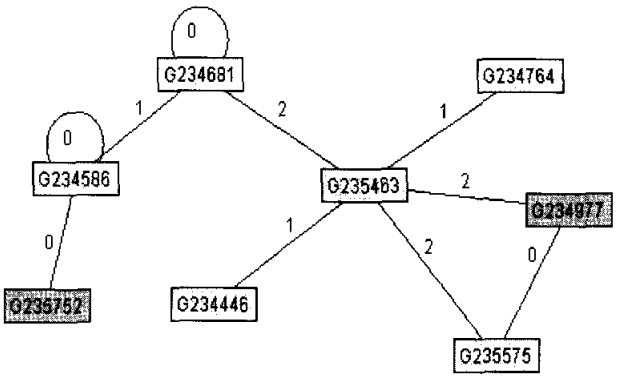
- N 주어진 단백질의 총수
- K 각 단백질이 가질 수 있는 가능한 기능의 수
- P_i i 번째 단백질, i=1,..., N
- F_i 단백질 P_i 기능 벡터 $<f_1, f_2, ..., f_K>$ 여기서 f_j는 단백질의 j-번째 기능을 나타내며 1이면 P_i는 f_j의 기능을 가지고 있으며 기능이 알려지지 않은 단백질의 경우 모든 j에 대해 f_j=?임
- T 단백질 상호작용 2차원 행렬(adjacency matrix)로서 대칭(symmetric)이며 기능에 관계없이 주어진 단백질들에 대한 상호작용의 여부를 나타냄. $T = [t_{ij}]_{NXN}$. P_i와 P_j가 상호작용을 하면 $t_{ij}=t_{ij}=1$ 그렇지 않으면 0이다.
- $NG^d(i)$ 단백질 P_i로부터 d-깊이(depth)의 직간접 상호작용을 통하여 도달 가능한(reachable) 이웃한 단백질들의 집합
- π^d_k T로부터 d-깊이(depth)상호작용을 통하여 기능 f_k을 가질 수 있는 사전(a priori) 확률로 T로부터 계산
- $n^d_i(k)$ $NG^d(i)$로부터 계산되는 이웃한 단백질 중에서 기능 f_k을 갖는 이웃한 단백질의 수($d \geq 1$)
- E 단백질들 간의 expression correlation 행렬. $E=[e_{ij}]_{N\times N}$는 P_i와 P_j의 expression correlation 값임.

Interaction generality[11]는 단백질 상호작용 실험에서 발생되는 false-positive 상호작용을 평가하기 위한 방법으로 제안되었다. 두 상호작용 단백질 P_i와 P_j의 interaction generality $G(i,j)$는 다음과 같다.

$$G(i,j) = (|Ng(i)| + |Ng(j)| - 2) - \sum_{P_l \in Ng(i)} \delta(|Ng(l)|) + \sum_{P_m \in Ng(j)} \delta(|Ng(m)|)$$

여기서 델타함수(delta function) $\delta(x)$는 $x>1$이면 1, 그렇지 않으면 0이다. $G(i,j)$는 P_i, P_j와 직접 상호작용을 하는 단백질의 총수에 자신들을 제외한 이웃한 단백질이 직접 상호작용을 하는 단백질들의 총수를 제외한 값이 된다. 계산을 이용한 분석으로부터 interaction generality의 신뢰성이 보고되었으며 두 단백질의 상호작용에 대한 interaction generality가 작으면 두 단백질은 높은 co-expressed 된다는 사실을 밝혔다. 그러므로 유기체의 대규모 단백질 상호작용 맵의 구성은 높은 *interaction generality*를 갖는 상호작용을 제거함으로써 얻어질 수 있다. 그래프 기반 단백질의 기능

예측 모델은 신뢰할 수 있는 단백질-단백질 상호작용 맵으로부터 좋은 결과를 얻을 수 있을 것이다.



(그림 1) G234681과 G234977의 상호작용 맵

(그림 1)은 계산된 interaction generality 수치를 상호작용의 라벨로 갖는 맵을 보여주고 있다. 질의된 두 단백질은 G234681과 G234977이며 상호작용 깊이(interaction depth)는 2이다. 그림에서 G235752와 G234977은 기능이 알려지지 않은 단백질들이며 이들의 기능은 상호작용을 보이는 기능이 알려진 단백질로부터 예측될 수 있다. 자신과 상호작용(self-interaction)이 있는 상호작용에 대한 interaction generality는 0으로 계산하였다. 이유는 알려진 기능에 자신과 상호작용을 하는 단백질들에 대한 interaction generality는 기능 예측에 의미가 없다. <표 2>는 KDD Cup 2001 데이터 상호작용 데이터로부터 interaction generality의 정량화에 따른 주어진 데이터에 대한 상호작용의 수를 보여주고 있다. 약 93.74% 상호작용들은 interaction generality가 0에서 3사이에 분포되어 있다.

<표 2> KDD Cup 2001에서 interaction generality(IG)

IG	상호작용 수	분포율(%)
0	939	54.98
1	405	78.69
2	168	88.52
3	89	93.74
4	38	95.96
5	46	98.65
6	19	99.77
7	3	99.94
8	1	100.00
합	1,708	

(그림 2)는 단백질 기능 예측을 위한 제안된 알고리즘이다. *TrainSet*은 기능이 알려진 단백질 와 그의 기능의 쌍 (P_i,F_i)으로 P_i는 단백질의 이름, F_i는 단백질 P_i의 밝혀진 기능 벡터이다. *TestSet*은 기능이 알려지지 않은 단백질과 그의 기능의 쌍 $(Pj,?)$으로 표현된다. T는 단백질-단백질 상호작용을 나타내는 2차원 인접 행렬(adjacency matrix), gth는 interaction generality의 값, *eth*는 상호작용을 하는 두 단백질의 correlation coefficient는 나타내는 2 차원 인접 행렬이다. 여기서 l은 l-번째 높은(l-largest) 빈도수를 위한 값이며 d는 기능 예측에 고려되는 직-간접 상호작용의 d-깊이(d-depth)내 속한 단백질을 포함시키기 위한 값이다. 제안된 알고리즘은 병목(threshold) 파라미터 *gth, eth, l, d* 등의 변화에 따라 기능 예측에 고려되는 단백질의 수를 제한하게 된다.

```
Annotate_Function( TrainSet, TestSet, T, gth, eth, l, d)
// TrainSet is a hash table where each entry is presented by a pair of (Pi,Fi)
// TestSet is a list entry of proteins Pj for protein of unknown function
// T is a symmetric interaction matrix
// gth is the threshold value for interaction generality
// eth is the threshold value for expression correlation
// l is the value for choosing l largest values with gth and eth
// d is the interaction depth
while TestSet.size() > 0 do
  1. Select a set of neighbors of Pj in TestSet
     • Q=<q1,q2,...,qK> where qm=0 for all m
     • Fj=<f1,f2,...,fK> where fm=0 for all m
     • compute Ngd(j)
  2. Compute the function frequency of Pj
     • for Pi ∈ Ngd(j) do
         if G(i,j)≤gth and EC(i,j)>eth then
           increment qm by 1 if Pi has fm=1
  3. Annotate functions to Fj
     • decide l largest functions f'1,f'2,...,f'l using Q
     • assign <f'1,f'2,...,f'K> to Fj
  4. Add (Pj, Fj) to TrainSet
  5. Remove Pj from TestSet
```

(그림 2) 단백질 기능 예측을 위한 제안된 알고리즘

*TestSet*의 P_j에 대해 d-깊이내 상호작용을 하는 TrainSet의 단백질 $P_i \in Ng^d(j)$로부터 기능 빈도 벡터 F_j를 계산한다. 만약, P_i와 P_j의 interaction generality가 *gth*보다 작고, expression correlation이 *eth*보다 작으며 F_j의 계산에서 P_i는 제외된다. F_j로부터 l-largest 기능만이 Pj의 기능으로 예측하고 (P_j,F_j)는 *TrainSet*으로 옮겨지고 *TestSet*에서는 제거된다. 알고리즘의 수행은 *TestSet*에 속한 단백질이 없거나 예측되는 *TestSet*의 단백질이 존재하지 않으면 예측을 종료한다.

<표 3> KDD Cup 데이터의 neighbor-counting 모델의 실험 결과[1)]

l	1	2	3	4	5	6	7
TP	535	641	712	729	733	738	738
TN	3,356	3,265	3,114	3,005	2,943	2,899	2,868
FP	355	249	178	161	157	152	152
FN	122	213	364	473	535	579	610
No	336	336	336	336	336	336	336
Acc(%)	89.08	89.42	87.59	85.49	84.16	83.26	82.55

5. 실 험

<표 3>은 KDD Cup 데이터에서 neighbor-counting 모델의 실험 결과를 보여주고 있다. 예측 성능은 confusion ma-

1) TP:true positive, TN:true negative, FP:false Positive, FN:false negative
Acc = (TP+TN)/(TP+TN+FP+FN) * 100.0

trix로 계산되었으며 상호작용을 하는 이웃한 단백질의 기능별 빈도수 l의 값을 1에서 7까지 변화시켜 예측 성능을 확인하였다. 테스트 데이터로부터 직-간접적인 상호작용을 갖는 336개의 단백질에 대해 기능 예측이 가능하였으며 l=2에서 89.42%의 가장 좋은 예측 율을 보이고 있다. 이러한 실험 결과는 KDD Cup 데이터에서 각 단백질이 약 2~3개의 기능을 가지고 있다는 것을 반영한다고 있다. l 값이 커짐에 따라 예측 성능이 저하되고 있다.

〈표 4〉 KDD Cup 데이터의 x^2-통계치 모델의 실험 결과

l	1	2	3	4	5	6	7
TP	222	432	545	669	706	751	786
TN	3,429	3,234	3,005	2,631	2,346	2,143	1,745
FP	628	428	305	181	144	99	99
FN	89	284	513	887	1,172	1,375	1,738
No	336	336	336	336	336	336	336
Acc(%)	83.59	83.93	81.27	75.55	69.97	66.25	57.94

<표 4>는 x^2-통계치를 응용한 단백질 기능 예측의 실험 결과이다. Neighbor-counting 방법과 마찬가지로 336개의 직-간접적인 상호작용이 존재하는 336개의 단백질에 대해 기능 예측이 가능하였으나 예측 성능은 l의 값이 1, 2인 경우를 제외하고 3 이상의 경우 예측 율이 급격하게 감소되고 있다. 이러한 l 값의 변화에 따른 성능 저하는 모델에서 x^2 통계치의 가정은 단백질의 기능에 대한 확률분포의 부적합성에 있다고 예상된다.

제안된 알고리즘은 IG의 값을 1부터 5까지 가변시키면서 실험을 하였으며, 표 5는 IG가 1과 5인 경우의 예측 성능을 보여준다. 표 5에서 IG가 1인 경우에는 251개의 단백질 기능이 예측되었고, IG값이 증가하여 5인 경우에는 단백질의 수가 증가하여 304개의 단백질 기능 예측이 가능하였다. expression correlation 데이터의 이용은 전이 관계를 적용하며 0.5 이상 되는 경우에만 예측을 하도록 하였다. Neighbor-counting 방법에 비해 작은 예측율의 증가를 보이고 있으나 잘못 예측 분류한 오류 율(FN 및 FP의 비율)에서 제안된 방법의 경우 대부분의 경우 작게 나타나고 있다. 이는 모델에서 서열 유사성 정보를 이용한 새로운 상호작용과 interaction generality를 포함시켜 기능 예측에 대한 제한적 적용에서 발생한다.

〈표 5〉 KDD Cup 데이터의 제안하는 모델의 실험 결과

IG=1	1	2	3	4	5	6	7
TP	435	493	527	537	540	541	541
TN	2,487	2,448	2,362	2,297	2,283	2,271	2,260
FP	238	180	146	136	133	132	132
FN	103	142	228	292	307	319	330
No	251	251	251	251	251	251	251
Acc(%)	89.55	90.13	88.54	86.85	86.51	86.18	85.84

IG=5	1	2	3	4	5	6	7
TP	543	602	644	652	654	654	654
TN	3,006	2,966	2,859	2,824	2,812	2,811	2,811
FP	144	184	291	326	338	339	339
FN	259	200	158	150	148	148	148
No	304	304	304	304	304	304	304
Acc(%)	89.10	90.28	88.64	87.96	87.70	87.68	87.68

6. 결 론

본 논문에서는 생물정보학 분야에서 그래프 이론을 기반으로 한 단백질 기능 예측방법을 살펴보았고, 각 접근 방법에 대한 향후 연구 방향 및 개선 방향을 제시하였다. 현재까지 대부분의 단백질 기능 예측은 *in vitro* 실험으로부터 얻어진 대규모 단백질 상호작용 데이터를 이용하였지만, 데이터의 신뢰성에 대해 의문시 되어왔다. 그래서 단백질 기능 예측 모델의 정확도를 높이기 위해 관련 유전인자 및 단백질의 기능상 관련 속성만 고려하는 모델들이 높은 기능 예측 율을 보장하기 위해 연구되어야 한다. 제안된 단백질 기능 예측 모델은 그래프 이론에 기반하며, 단백질의 서열 유사성 정보 및 경험적 지식을 이용하고 있다. 제안된 단백질 기능 예측 모델은 KDD Cup Yeast 단백질 데이터 셋에 대해 neighbor-counting 모델 및 x^2-통계치 모델과 비교되었다. 향후 연구 과제로는 단백질 기능 예측에 필요한 풍부한 데이터 셋에 대한 연구가 필요하며, 생물정보학의 전문가들과 협력을 통하여 표준화된 데이터 셋을 준비함으로써 계산모델의 객관적 평가 방법에 대한 연구의 병행이 요구된다. 앞으로 보다 나은 정확성을 제공하기 위해 밝혀진 단백질에 대한 기능별 속성 및 생물 전문가의 기능 판별 지식 등을 이용한 그래프 기반 예측 알고리즘의 성능 개선에 대한 연구가 필요하다.

참 고 문 헌

[1] N. M. Luscombe et al., What is bioinformatics? An introduction and overview, International Medical Informatics Association Yearbook, p 83-100, 2001

[2] B. Schwikowski et al, A network of protein-protein interactions in yeast, Nature Biotechnology, p 1257-1261, no 3, vol 8, 2000

[3] P. Baldi et al., Bioinformatics: The Machine Learning Approach, The MIT Press, 2003

[4] M. Fellenberg et al., Integrative Analysis of Protein Interaction Data, Intelligent Systems for Molecular Biology, AAAI Press, p 152-161, vol 8, 2000

[5] T. Ito et al., Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins, Proceedings of the National Academy of Sciences, p 4569-4574, vol 97, 2000

[6] C. L. Tucker et al., Towards an understanding of complex

protein networks, TRENDS in cell biology, p 102-106, no 3, vol 11, 2001

[7] J. Cheng et al, KDD Cup 2001 Report, SIGKDD Exploration, p 47-64, vol 3, 2001

[8] S. Oliver, Guilt-by-association goes global, Nature, p 601-603, vol 403, 2002

[9] T. Mitchell, Machine Learning, McGraw Hill, 1997

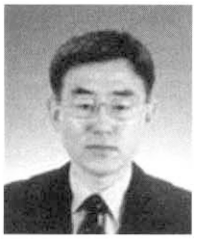
[10] H. Hishigaki et al., Assessment of prediction accuracy of protein function from protein-protein interaction data, yeast, p 523-531, vol 18, 2001

[11] A. Clare and R. D. King, Machine learning of functional class from phenotype data, Bioinformatics, p 160-166, vol 18, 2002

[12] Minghua Deng, Shipra Mehta, Ting Chen, Fengzhu Sun, Predictions of protein function using protein-protein interaction data, The first IEEE Computer Society bioinformatics conference, CSB2002, 2002

[13] MIPS Yeast Data, http://mips.gsf.de/proj/yeast/

[14] R. Saito et al., Interaction generality, a measurement to assess the reliability of a protein-protein interaction, Nucleic Acids Research, p 1163-1168, no 5, vol 30, 2002



황 두 성

e-mail : dshwang@dankook.ac.kr

1985년 충남대학교 계산통계학과 졸업(학사)

1990년 충남대학교 대학원 계산통계학과(석사)

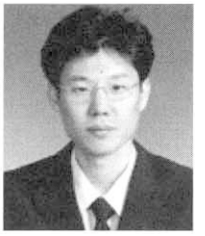
2003년 Wayne State University, Computer Science(박사)

1990년~1991년 국토개발연구원 연구원

1991년~1998년 전자통신연구소 연구원

2003년~현재 단국대학교 컴퓨터과학과

관심분야 : 데이터 마이닝(data mining), 머신 학습(machine learning), 바이오인포매틱스(bioinformatics)



정 재 영

e-mail : jjy72@etri.re.kr

1999년 경북대학교 전자공학과(공학사)

2001년 경북대학교 전자공학과(공학석사)

2001~현재 한국전자통신연구원 미래기술연구본부 연구원

관심분야 : 바이오인포매틱스(bioinformatics), 프로티오믹스(proteomics)