# 특징 추출과 분석 기법에 기반한 단백질 상호작용 데이터 신뢰도 향상 시스템

이 민 수[†] · 박 승 수[††] · 이 상 호[†††] · 용 환 승[†††] · 강 성 희[††††]


## 요 약

대용량 실험으로부터 산출된 단백질 상호작용 데이터는 위양성(false positive) 데이터의 비율이 높다는 단점을 가지고 있다. 본 논문에서는 오류가 섞여있는 단백질 상호작용 데이터를 입력으로 받아 각 단백질 상호작용의 신뢰도를 검증하는 시스템을 제안하고 구현하였다. 제안 시스템은 단백질 상호작용 데이터에 상호작용의 근거로서 사용될 수 있는 다양한 생물학적 특징들에 관한 데이터를 통합하고 특징 선택 방법을 사용하여 통합된 속성들 중 위양성 여부를 판별하는데 가장 적합한 특징들을 선택한 후 데이터 마이닝 분류 알고리즘을 적용하여 대용량 실험으로부터 산출된 단백질 상호작용 데이터의 신뢰도를 평가한다. 특징 선택의 결과와 분류 기법의 성능은 데이터 특성에 매우 의존하므로, 제안 시스템에 가장 적합한 속성 부분집합과 가장 좋은 성능을 내는 분류 알고리즘을 찾기 위해 다양한 특징 선택 방법과 데이터 마이닝 분류 알고리즘들을 적용하고 그 성능을 다각적으로 비교분석 하였다. 실험 결과, 특징 선택 방법과 분류 알고리즘을 결합시킨 제안 시스템은 오류 데이터가 섞여있는 단백질 상호작용 데이터에서 실제로 상호작용하는 단백질 쌍을 골라내는 작업에 있어 기존 연구들에 비해 매우 뛰어난 성능을 보여줬다. 또한 본 연구를 통해 단백질 상호작용 데이터의 신뢰도를 검증함에 있어서 다양한 특징 선택 방법들과 분류 알고리즘들이 성능에 미치는 영향에 관해서도 정리할 수 있었다.


**키워드 : 데이터 마이닝, 분류 기법, 특징 선택, 단백질 상호작용**

# Protein-Protein Interaction Reliability Enhancement System based on Feature Selection and Classification Technique

Min Su Lee[†] · Seung Soo Park[††] · Sang Ho Lee[†††] · Hwan Seung Yong[†††] · Sung Hee, Kang[††††]


## ABSTRACT

Protein-protein interaction data obtained from high-throughput experiments includes high false positives. In this paper, we introduce a new protein-protein interaction reliability verification system. The proposed system integrates various biological features related with protein-protein interactions, and then selects the most relevant and informative features among them using a feature selection method. To assess the reliability of each protein-protein interaction data, the system construct a classifier that can distinguish true interacting protein pairs from noisy protein-protein interaction data based on the selected biological evidences using a classification technique. Since the performance of feature selection methods and classification techniques depends heavily upon characteristics of data, we performed rigorous comparative analysis of various feature selection methods and classification techniques to obtain optimal performance of our system. Experimental results show that the combination of feature selection method and classification algorithms provide very powerful tools in distinguishing true interacting protein pairs from noisy protein-protein interaction dataset. Also, we investigated the effects on performances of feature selection methods and classification techniques in the proposed protein interaction verification system.


**Key Words : Data Mining, Classification Technique, Feature Selection, Protein-Protein Interaction**

## 1. 서 론

생명과학 기술의 발달로 인해 기존에는 하나씩 수작업 실험을 통해 밝혀야 했던 생물학적 현상들이 최근에는 대용량(high-throughput) 실험 방법을 통해 엄청난 양의 결과들로 산출되고 있다.

대표적인 대용량 실험 방법으로 한 개체 안에 존재하는 모든 단백질들 간의 상호작용(protein-protein interaction, PPI)을 밝혀내는 실험을 들 수 있다. 단백질 상호작용 데이터는 각 단백질의 기능을 예측하기 위해 필수적으로 사용되

† 종신회원 : 이화여자대학교 컴퓨터학과 박사과정
†† 정 회 원 : 이화여자대학교 컴퓨터학과 교수
††† 종신회원 : 이화여자대학교 컴퓨터학과 교수
†††† 정 회 원 : 명지대학교 방목기초교육대학 교수
논문접수 : 2006년 9월 1일, 심사완료 : 2006년 12월 4일

며[1, 2], 세포 외부로부터 온 신호를 세포의 핵 안으로 전달해주는 신호전달 경로 등을 밝히는데 이용될 수 있다[3]. 한 종에 존재하는 단백질 상호작용을 스크리닝하는 대용량 실험 방법으로 Y2H(yeast two hybrid)[4,5]나 Mass Spec-trometry[6, 7] 등을 들 수 있다. 이러한 대용량의 실험들은 대량의 실험 결과를 효율적으로 생성하지만, 반면에 상호작용한다고 실험 결과로 나왔지만 실제 생체 내에서는 상호작용하지 않는 위양성(false positive) 단백질 쌍의 비율이 수작업 실험에 비해 매우 높다는 단점을 가지고 있다. 대용량 실험으로부터 산출된 단백질 상호작용 데이터의 약 50% 정도가 위양성 데이터라고 보고되고 있다[8, 9]. 세포 안에 존재하는 모든 분자 상호작용들을 밝히는 상호작용체학(interactomics)이나 이를 바탕으로 생물의 항상성 유지를 위해 세포 안에서 일어나는 프로세스들을 밝히는 대사경로학(metabol-omics)에 대한 연구는 신뢰성 있는 단백질 상호작용 데이터에 기반해서 이루어져야 하므로, 대용량 실험으로부터 산출된 단백질 상호작용 데이터에 대한 신뢰성 검증은 매우 중요한 이슈가 되었다.

본 연구에서는 오류데이터가 섞여있는 단백질 상호작용 데이터에서 실제로 상호작용하는 단백질 쌍들을 분류하기 위해 관련 생물학 데이터를 속성(attribute)로 통합하고, 이들 중 가장 관련된 속성들을 특징(feature)으로 선별한 후 데이터 마이닝 분류 기법을 적용하여 진양성과 위양성을 구분함으로써 각 상호작용 쌍을 검증하는 시스템을 소개하고 더 나아가 시스템의 성능을 최적화하는 특징 선택 방법과 분류 알고리즘의 조합을 찾는 것을 목표로 한다. 시스템 구축 및 실험 결과, 제안 시스템의 성능은 97.04%의 정확도와 96.27%의 진양성율, 97.78%의 특이성, 그리고 97.02%의 F-measure을 기록하여 기존 연구들에 비해 월등히 뛰어났다. 또한 본 연구를 통해 단백질 상호작용 검증 시스템을 구축함에 있어 적용하는 특징 추출 방법들과 분류 기법들이 어떠한 영향을 미치는지 정리할 수 있었다.

## 2. 관련 연구

### 2.1 단백질 상호작용 데이터 검증을 위한 관련 연구

대용량 실험으로부터 산출된 단백질 상호작용 데이터에 위양성 데이터가 매우 많이 포함되어있다는 사실이 보고된 후로, 이들 중 어떤 단백질 쌍이 실제로 상호작용 하는지를 검증하기 위해 다양한 방법들이 제안되고 있다. 대부분의 방법들은 단백질 상호작용과 관련 있다고 알려진 다른 생물학 데이터와의 상관성을 계산하여 각 상호작용한다고 예측된 단백질 쌍의 신뢰도로 사용하고자 하였다.

예를 들면, 단백질들의 mRNA의 발현 양상이 비슷하면 그 단백질들은 상호작용할 가능성이 높다고 밝힌 논문들에 근거하여 mRNA 발현양상을 이용하여 단백질 상호작용 데이터를 검증하는 연구가 있었다[10, 11]. 그러나 최근 연구들은 유전체 상에서의 단백질 상호작용은 서로 다른 분해율(degradation rate) 때문에 유전자 발현과 매우 약한 관계만 갖는다고 보고하고 있으므로[12, 13], mRNA 발현 양상만으로는 단백질 상호작용의 신뢰도를 평가하는데 무리가 있다.

또 다른 예로, 상호작용하는 두 단백질과 상동(homol-ogue)인 관계에 있는 다른 두 단백질이 존재하면 그 두 단백질도 서로 상호작용할 가능성이 매우 높다는 이론에 근거하여, 종들 사이에 보존되는 상동 단백질들이나 같은 종 안에서 상동 관계인 단백질들에 대한 정보를 이용하여 대용량 실험 결과로 나온 상호작용 쌍의 상동 단백질 쌍이 서로 상호작용 하는지를 찾아봄으로써 상호작용하는 단백질의 신뢰도를 검증하고자 한 연구도 있었다[14, 15]. 그러나 이 방법은 상동 단백질에 대한 정보가 알려져 있는 단백질들에만 적용할 수 있다는 제약을 가지고 있어 적용 범위(coverage)가 매우 좁다는 한계가 있다.

다른 방법으로는 같은 종에 대해서 여러 개의 대용량 분석 실험 방법들을 적용하여 산출된 데이터들을 비교하고, 여러 실험 방법을 통해 공통적으로 나타나는 상호작용 쌍들을 찾음으로써 재현성(reproducibility)에 기반해서 신뢰성있는 단백질 상호작용 데이터셋을 찾아내고자 한 연구도 있다[8]. 그러나 서로 다른 대용량 실험들은 실험 방법의 원리에 따라 서로 다른 특성을 가지는 단백질 상호작용 쌍들을 검출해 낸다. 또한 같은 기술을 사용해서 같은 종 안에 존재하는 단백질 상호작용들을 동정해보더라도 서로 다른 연구그룹에서 실험하면 공통부분이 적은 데이터들이 산출된다[4,5]. 이러한 대용량 실험 방법의 제약성 때문에 대용량의 단백질 상호작용 데이터들 사이의 교집합은 매우 작다는 한계점을 가지고 있다[8].

이러한 방법들은 각 상호작용하는 단백질 쌍의 신뢰도를 검증하기 위해 한 종류의 생물학적 속성에만 의존하기적용범위 그래서 경험적으로 단백질 상호작용과 관련 있다고 알려진 3~4가지의 생물학 특징을 통합하여 상호작용 단백질을 예측하고자 한 연구도 있었다[16-19]. 그러나 기존 연구들의 단백질 신뢰도 검증 결과들은 진양성율(true positive rate)라고도 불리는 민감도(sensitivity)는 다소 높게 나와도 그 결과의 특이성(specificity)는 매우 낮다는 즉, 실제로 상호작용하는 단백질 쌍들을 잘 찾아내지만, 반면 실제로 상호작용하지 않는 단백질 쌍들 까지도 많이 포함시켜 상호작용하는 단백질 쌍으로 분류한다는 한계점을 가지고 있었다.

본 연구에서는 상호작용하는 단백질 쌍들을 검증하기 위해 여러 관련된 생물학 데이터들을 모두 통합한 후 특징 선택 기법을 사용하여 여러 속성들 중 단백질 상호작용을 검증하기 위해 가장 적절하고 정보력 있는 속성들의 부분집합을 특징으로 선택한 후, 이 특징들에 데이터 마이닝 분류 기법을 적용하였다. 기존에는 경험적 지식에 기반하여 주관적으로 선택했던 검증 근거로 활용할 생물학적 특징들을 정보 공학적 특징 선택 방법들을 사용하여 선택하도록 하였고 다양한 분류 기법을 적용함으로써 최적의 성능과 폭넓은 적용 범위를 지원할 수 있는 검증 시스템을 구축하였다.

### 2.2 특징 선택 (Feature Selection)

'The curse of dimensionality'라는 말은 데이터의 차원이

증가할수록 데이터 분석이 급격하게 더 어려워진다는 것을 의미한다. 데이터의 차원이 증가할수록 불필요한 속성이나 값들이 많이 비어있는 속성들이 포함되어 분류 정확도나 클러스터의 품질이 매우 떨어질 수 있으며 데이터 마이닝 작업을 수행할 때 계산 복잡도가 매우 증가하여 수행 시간이나 비용 면이 매우 증가하기도 한다. 따라서 데이터 마이닝 전처리(preprocessing) 단계에서 데이터의 차원을 줄이는 특징 선택을 수행하기도 한다. 속성들의 부분집합을 선택하는 방법은 데이터가 가진 정보를 잃는 것처럼 보이지만, 중복되거나 관련 없는 속성들이 있을 경우에는 매우 유용하다.

특징 선택을 수행함으로써 생기는 장점들이 많이 있다[20-22]. 첫째, 특징 선택 과정은 데이터 안의 무의미한 속성들을 제거하고 노이즈를 줄여주기 때문에 대부분의 데이터 마이닝 분류나 군집화 알고리즘들은 특징 선택 적용을 통해 데이터의 속성의 수를 줄여준 후 적용하면 분류 성능이 향상된다. 둘째, 특징 선택 과정을 통해 사용자의 관심을 가장 관련된 변수들로 집중시켜주며 목표 항목들에 관한 해석을 보다 간결하게 만들어준다. 또한 데이터를 2~3개의 차원으로까지 줄이지는 않을 지라도, 기존보다 속성 차원이 줄어들고 속성들의 조합의 경우 수도 현격하게 줄여줌으로써 데이터 시각화도 보다 용이해진다. 마지막으로, 특징 선택하는 과정이 추가되기는 하지만, 실제 데이터 마이닝 알고리즘 적용 과정에서 요구되는 시간, 메모리의 양을 현저하게 줄여줘서 빠르고 효율적으로 예측 모델을 구축할 수 있다.

특징 선택 과정은 속성들의 각 부분 집합을 평가하기 위한 방법(evaluation method), 특징들의 새로운 부분 집합을 생성하기 위한 탐색 전략(search strategy), 그리고 정지 기준(stop criterion)의 세 부분으로 구성된다. 속성들의 부분집합을 평가하기 위한 방법은 크게 개별 속성 평가 방법과 속성 부분집합 평가 방법으로 나눌 수 있다. 개별 속성 평가 방법은 통계적 테스트로부터 나온 p-value나 q-value, 정보 획득이나 지니 지표의 값, 또는 선형 SVM에 적용했을 때의 계수 값을 이용하여 각 속성들의 정보력을 수치 값으로 표현하여 평가하는 방법이다. 보통 개별 속성 평가 방법은 각 속성들의 정보력의 순위를 매기고 특정 값을 기준으로 특징으로 사용할 속성들을 선택한다. 이에 반해 속성 부분집합은 특정 탐색 전략에 의해 속성들의 공간을 탐색하며 선택한 속성 부분집합에 대해서 속성 간의 상관관계(correlation-based feature selection)나 속성들에 따른 목표 항목 값의 일관성(consistency-based feature selection) 등을 고려하거나 특정 분류 알고리즘에 적용했을 때의 성능(wrapper-based feature selection) 등에 기반하여 각 속성 부분집합의 정보력을 평가하고 그 값에 기반하여 특징으로 사용할 속성들의 부분집합을 선택한다.

본 논문에서는 단백질 상호작용과 관련지을 수 있는 여러 생물학적 속성들을 통합하여 만든 데이터에서 단백질 상호작용의 신뢰도 검증에 사용할 가장 관련되고 정보력 있는 속성들을 선별하기 위해, 일반적인 경험적 지식에 근거한 특징 선택 방법, 상관관계에 기반한 특징 선택 방법, 일관성에 기반한 특징 선택 방법, 그리고 의사결정 나무 알고리즘, 선형 SVM알고리즘을 래퍼 형태로 적용한 특징 선택 방법들을 사용하여 그 성능을 비교분석하였다.

### 2.3 분류 기법 (Classification Technique)

데이터 마이닝의 분류 기법은 학습 데이터를 이용해서 목표 항목 값을 분류하는 모델을 구축 한 후, 목표 항목 값을 모르는 데이터를 분류 모델에 입력하여 그의 목표 항목을 예측하는데 이용한다. 분류 기법들은 근본 원리에 따라 의사결정 나무, 예제 기반 학습(instance-based learning), 베이즈 규칙(Bayes' rule)에 기반한 기법, 함수형 분류 기법 등으로 나눌 수 있다.

의사결정 나무 기법 (Decision Tree, DT)은 범주형 값을 예측하는 데 유용하게 사용된다. 의사결정 나무 알고리즘은 분할정복(divide and conquer) 방법을 재귀적으로 사용하며 탑다운 방식으로 결정 나무를 구축하는 욕심쟁이(greedy) 알고리즘이다. 의사결정 나무 기법의 알고리즘은 기본적으로 각 노드에 들어있는 학습 데이터를 목표 항목에 기반해서 가장 잘 분별해주는 속성과 그때의 속성 값을 선택한다. 이때, 주어진 속성이 학습 데이터를 목표 항목에 따라 얼마나 잘 분류하는지를 측정하기 위해 정보 획득이라는 기준을 사용한다. 선정된 분류 속성에 따라 이 분화를 계속하여 최종 리프 노드엔 목표 항목이 있는 나무를 생성하게 된다.

예제 기반 학습의 대표적인 예로 K-최근접 이웃(K-Nearest Neighborhood, KNN) 알고리즘을 들 수 있다[24]. K-최근접 이웃 알고리즘은 정보 검색 분야에서 많이 사용되는 비모수(non-parametric) 분류 알고리즘으로 학습 데이터를 이용하여 분류 모델을 구축하는 학습 단계가 필요 없다. K라는 입력 변수를 받아 입력 쿼리 x 가 들어오는 데로 x와 가장 비슷한 K개의 학습 데이터를 찾아 그들의 목표 항목에 기반해서 x의 목표 항목을 예측한다.

베이즈 를 기반 분류 기법의 대표적인 예로 베이지안 네트워크(Bayesian Network, BN) 알고리즘과 나이브 베이즈(Naïve Bayes, NB) 알고리즘을 들 수 있다. 베이지안 네트워크 알고리즘은 다수의 속성들의 결합확률분포를 속성들 사이의 조건부 독립성에 기반해 표현하고 속성과 목표 항목 사이의 연관 관계를 계산함으로써 분류 작업에 활용한다[20]. 나이브 베이즈 알고리즘은 문서 분류 작업에 많이 활용되는 학습 알고리즘으로써 속성 간에 조건부 독립 조건을 만족한다는 전제하에 베이즈 규칙에 기반해서 관측된 샘플과 목표 항목들의 공동 확률 값을 이용하여 주어진 관측된 목표 항목들의 조건부 확률을 추정한다[25].

함수형 분류 기법으로는 Support Vector Machine (SVM), 다층 퍼셉트론 (Multilayer Perceptron, MP), 그리고 RBF망(Radial Basis Function Network, RBFN)등의 알고리즘을 들 수 있다[20]. SVM은 알고리즘 원리 자체가 두 개의 클래스를 가장 큰 여유를 두고 분류할 수 있는 초평면

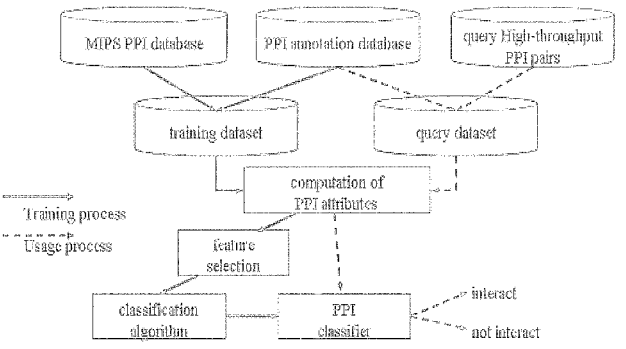
(hyperplane)을 찾는 것을 목적으로 하고 있기 때문에, 이진 분류 문제에서 매우 탁월한 성능을 보여준다. 다층 퍼셉트론 알고리즘은 패턴 인식이나 데이터 분류에 많이 사용되는 생물학적 신경 시스템에 착안한 정보처리 패러다임이다. 입력 값에 따른 분류 항목들의 규칙을 인식하기 위해 수많은 뉴론들이 학습 데이터에 따라 결정된 가중치 값들에 의해 연결되어있는 하나의 결합체 구조로 구성된다. 전형적인 다층 신경망의 은닉층 뉴런은 시그모이드 작동함수를 갖고 있고 입력의 가중합을 취한다. 이에 비해 RBF망은 입력의 가중합을 취하지 않으며, 은닉 뉴런의 작동함수는 가우시안(Gaussian) 함수를 사용한다. RBF 망은 전형적인 다층 순방향 네트워크보다 빠른 수렴속도, 보다 적은 오차와 높은 신뢰성을 보여주기도 한다.

본 논문에서는 단백질 상호작용 검증 시스템에 가장 적합한 분류 알고리즘을 찾기 위해, 의사결정 나무 기법의 C4.5 알고리즘[23], 예제 기반 학습 기법의 K-최근접 이웃알고리즘, 베이즈 규칙 기반 기법의 나이브 베이즈와 베이지안 네트워크, 그리고 함수형 분류 기법의 SVM과 다층 퍼셉트론, RBF망 알고리즘 들을 적용하였다.

## 3. 단백질 상호작용 데이터 검증 시스템

본 장에서는 우리가 제안하는 특징 추출과 분류 기법을 적용한 단백질 상호작용 데이터 검증 시스템에 대해 설명한다. 제안 시스템은 대용량 실험에 의해 추정된 상호작용하는 단백질 데이터셋에서 진양성 데이터와 위양성 데이터를 구분하기 위해 데이터 전처리(data preprocessing) 단계에서 단백질 상호작용의 근거로 사용될 다양한 생물학 데이터를 통합하고, 이들 중 단백질 상호작용 검증을 위한 최적의 근거로 사용될 수 있는 특징 부분집합을 찾기 위해 특징 선택 방법을 데이터에 적용한다. 그 후, 데이터 마이닝 분류 알고리즘을 이용하여 단백질 상호작용 데이터의 신뢰성을 검증하는 판별 모델을 구축한다.

(그림 1)에서 실선으로 된 화살표는 분류기를 생성하기 위한 학습 과정을, 점선으로 된 화살표는 구축된 분류기에 기반해서 대용량 실험으로부터 산출된 데이터를 분류하는 과정을 나타낸다. 학습 과정에서는 우선 MIPS 데이터베이스로부터 가져온 단백질 상호작용 쌍들과 단백질 상호작용에 관한 명세 정보를 이용하여 분류 모델 학습에 이용될 단백질 쌍 리스트로 구성된 데이터를 만든다. 구축된 학습 데이터에 단백질 상호작용의 근거로 사용될 수 있는 생물학적 의미를 가지는 값들을 계산하여 여러 속성들을 더한다. 그 후, 실제로 상호작용하는 단백질 쌍과 상호작용하지 않는 단백질 쌍을 가장 잘 골라줄 수 있는 근거로 사용될 생물학적 특징들을 선택하기 위해 특징 선택 방법을 적용하고, 그 결과에 분류 알고리즘을 적용하여 단백질 상호작용 신뢰성 검증을 위한 분류기를 구축한다. 새로운 대용량 실험으로부터 보고된 단백질 상호작용 쌍이 입력으로 들어오면 명세




(그림 1) 대용량 실험으로부터 나온 단백질 상호작용 데이터를 분류하기 위한 시스템 구조

데이터베이스를 이용하여 각 단백질 쌍마다 속성 값들을 계산한 후, 학습 과정을 통해 구축된 분류기를 이용하여 각 쌍의 진위성 여부를 검증하게 된다.

본 시스템에서 학습에 사용되는 단백질 상호작용 데이터셋은 목표 항목이 양성인 데이터와 음성인 데이터들로 구분할 수 있다. 데이터 마이닝 결과의 품질은 데이터 소스의 품질과 매우 밀접하게 연관되어 있으므로 데이터셋을 어디로부터 모으고 어떻게 구성할 것인가는 매우 중요한 문제이다. 특히 학습에 사용될 표본 데이터셋은 모집단의 분포를 반영하며 특정 현상을 표현하는데 있어 치우침이 없어야 한다.

공개된 단백질 상호작용 데이터베이스들의 대부분은 대용량 실험 결과로 나온 데이터나 다른 방법들에 의해 추정된 단백질 상호작용 데이터들을 많이 포함하고 있어 그 신뢰성이 매우 떨어지는 반면, MIPS (Munich Information Center for Protein Sequence)[27,28] 공개 데이터베이스는 전문가의 수작업을 거쳐 엄선한 단백질 상호작용들로 구성되어 있기 때문에 신뢰성이 매우 높은 데이터로 여겨진다. 본 논문에서 양성 데이터로 사용한 실제로 상호작용하는 단백질들의 쌍들은 MIPS 단백질 상호작용 데이터베이스[28]에서 효모(Yeast) 종에 대한 단백질 복합체(complex) 데이터를 다운로드 받아 이를 이진 단위로 나누어서 만든 진양성 상호작용 8,250 쌍으로 구성되어 있다.

양성 단백질 상호작용 데이터와는 달리, 음성 단백질 상호작용 데이터는 존재하지 않는다. 각 단백질 쌍에 대해 가능한 모든 환경에서 두 단백질은 항상 상호작용하지 않는다라는 실험을 수행한다는 것은 현실적으로 불가능하며 이론상으로도 상호작용에 대한 증거가 부족하다는 것이 두 단백질이 상호작용하지 않는다는 것을 의미하지 않으므로, 상호작용하지 않는 단백질 쌍인 음성 데이터는 양성 데이터처럼 명확하게 정의할 수가 없기 때문이다. 그래서 우리는 Jansen *et al.*(2003) 논문[16]을 참고하여 양성 단백질 상호작용 데이터에서 사용된 단백질들을 이용하여 가능한 모든 종류의 단백질 쌍을 만들고, 각 단백질의 세포 내 위치 정보를 활용하여 그들 중 세포 기관을 4개의 단위로 나누었을 때 서로 만날 수 없는 위치에 존재하여 상호작용 할 수 없는 단백질들을 선택하여 음성 데이터셋을 만들었다. Jansen

〈표 1〉 검증 근거로 사용된 생물학적 속성들

| | 속성 | 설명 | 개수, 비율 |
|---|---|---|---|
| 1 | Absolute mRNA expression | 상호작용하려면 비슷한 양으로 존재해야 하므로 mRNA 발현량이 비슷한지 계산 | 15,998, 97.0% |
| 2 | mRNA co-expression | 같은 복합체 안에 있는 단백질들은 같이 발현하는 경향이 있는 보고에 근거하여 두 단백질의 발현 프로파일의 상관도 계산 | 15,752, 95.5% |
| 3 | Marginal essentiality | 한계적 필수성(marginal essetiality)이 높으면 두 단백질은 상호작용할 가능성이 높다는 가설에 근거하여 이를 계산 | 15,169, 92.0% |
| 4 | MIPS functional similarity | 같은 생물학적 프로세스에 있는 단백질들이 서로 상호작용할 가능성이 높다는데 근거하여 MIPS funcat 명세를 바탕으로 두 단백질이 공유하는 단백질 기능 카테고리를 찾고 그 카테고리의 특이성 계산 | 12,023, 72.9% |
| 5 | Absolute protein abundance | 상호작용하려면 비슷한 양으로 존재해야 하므로 단백질 양이 비슷한지 계산 | 9,834, 60.0% |
| 6 | GO functional similarity | 같은 생물학적 프로세스에 있는 단백질들이 서로 상호작용할 가능성이 높다는데 근거하여 Gene Ontology명세를 바탕으로 두 단백질이 공유하는 단백질 기능 카테고리를 찾고 그 카테고리의 특이성 계산 | 9,468, 57.4% |
| 7 | Co-regulation | 같은 전사인자에 의해 조절되는 단백질들이 상호작용할 가능성이 높다는 이론에 근거하여 co-regulation에 대한 데이터에 근거하여 계산 | 4,132, 25.0% |
| 8 | Interlogs in another organism | 종 간에 상호작용을 매핑한 interlog에 기반하여 계산 | 3,754, 22.8% |
| 9 | Co-essentiality | 두 단백질이 하나의 복합체에 존재한다면 둘 다 필수적인 단백질이거나 둘 다 필수적이 아니거나 한다는 근거하에 두 단백질 모두 필수적인지, 아닌지에 근거하여 값을 계산 | 3,918, 6.1% |
| 10 | Phylogenetic profiles | 상호작용하는 단백질들 사이에서 같이 진화한 것들이 보고되고 있으므로 진화 프로파일에 근거하여 계산 | 986, 6.0% |
| 11 | Gene neighborhood | 기능적으로 관련된 유전자들이 유전체 상에서 가까이 위치하는 경향이 있으므로 유전체 상에서의 거리 계산 | 314, 1.9% |
| 12 | Rosetta stone | 같은 패스웨이나 복합체 상에 존재하는 단백질들은 다른 종에서도 그렇게 존재하므로 이를 계산 | 115, 0.7% |
| 13 | Threading scores | 단백질 3차원상의 접힘을 예측하는데 사용되는 threading정보를 이용하여 상호작용 가능성 계산 | 109, 0.7% |
| 14 | Synthetic lethality | 개별적으로는 필수적이 아니더라도 같은 패스웨이나 복합체 상에 존재한다면 공동으로 두 유전자를 빼버렸을 때 세포가 죽으므로 이를 계산 | 97, 0.6% |
| 15 | Gene cluster(operon method) | 단백질 제조를 제어하는 유전자들의 단위인 같은 오페론에 속하는지 여부를 계산 | 4, 0.0% |
| 16 | Co-evolution scores | 단백질 종(family)사이에서의 같이 진화한 정도를 계산 | 2, 0.0% |

*et al.*(2003) 논문에서는 가능한 음성 데이터 2백70만여 개를 모두 학습 과정에 사용하였다. 그러나 모델 구축을 위해 사용할 샘플은 모집단의 분포도를 최대한 반영해야 하므로 우리는 위양성 데이터의 분포가 전체의 50% 정도로 예측된다는 논문[8,9]에 근거하여 이들 중 8,250개의 쌍을 랜덤하게 선택하여 음성 데이터셋을 만들어 사용하였다. 이런 과정을 통해 양성 8,250개 쌍과 음성 8,250개 쌍을 포함한 총 16,500개의 단백질 쌍으로 구성된 데이터를 제안 시스템 구현 및 성능평가에 사용하였다.

### 3.1. 상호작용 단백질들의 생물학적 특징 계산

단백질 상호작용 검증을 위해 그 동안 여러 논문들에서 언급하거나 사용했던 단백질 상호작용과 관련된 속성들을 모아서 상호작용 근거 속성들로 사용하였다[16,17]. 각 속성마다 가지고 있는 사례의 개수도 다르고 통합하는 과정에서 단백질 상호작용 데이터셋과의 교집합 정도도 각각 달랐다. <표 1>에 각 속성의 이름과 사용하게 된 생물학적 배경을 정리하고 각 속성이 전체 16,500개의 데이터 중 그 값이 존재하는 데이터는 몇 개인지 그리고 전체 데이터에서의 비율이 어느 정도인지에 따라 정렬하였다.

16가지의 속성들은 각각 계산하는 방법은 다르지만 대체적으로 비슷한 생물학적 특성에 기반한 속성들끼리 묶어볼 수 있다. 속성 1, 2번은 두 단백질들을 생성하기 위해 발현되는 mRNA 양들과 관련된 속성들이며, 속성 4번과 6번은 상호작용하는 단백질은 비슷한 기능을 수행한다는 점에 착안하여 단백질 기능 관련 온톨로지 (GO molecular function [19], MIPS funcat [20])를 이용하여 두 단백질의 기능이 얼마나 비슷한지를 측정한 속성들이다. 속성 3, 9, 14번은 두 단백질이 같은 단백질 복합체(complex)나 생물학적 경로(pathway)의 구성원이라면 그 복합체가 생명유지에 있어 필수적인지 아닌지에 따라 두 단백질은 같은 필수성(essentiality) 값을 가진다는 이론에 근거한 속성들이다. 속성 5번은 상호작용하려면 서로 절대적인 단백질 양이 비슷해야 한다는 이론을 반영한 속성들이다. 속성 7, 11, 15번은 두 단백질에 대한 정보를 가지고 있는 유전자들이 같은 전사인자에 의해 조절되거나 유전체 상에서 거리가 가까우면

서로 비슷한 기능을 수행할 가능성이 높고 기능을 수행하는 과정에 상호작용이 일어날 수 있다는 점을 이용한 속성들이다. 속성 8, 10, 16번은 진화 구조 기반해서 단백질 상호작용이 서로 다른 종 사이의 상동관계에서도 보존된다는 성질을 이용한 속성들이다.

### 3.2 특징 선택과 분류 알고리즘

3.1절에서 정리한 생물학적 속성들 중 목표 항목 즉, 해당 단백질 쌍이 상호작용 하는지 안 하는지 여부와 가장 관련된 속성 부분집합을 특징 선택 방법을 사용하여 선별하고 이들을 단백질 상호작용 데이터 검증의 근거로 사용한다. 특징 선택 알고리즘에 따라 선택된 특징 집합이 달라지며, 적용하는 도메인이나 데이터셋에 따라 최적의 성능을 보이는 분류 알고리즘이 달라진다. 따라서 우리는 여러 가지 특징 선택 알고리즘들과 분류 알고리즘들을 조합하여 제안 시스템에 적용하고 그들의 성능을 비교분석 함으로써, 제안 시스템에 가장 적합한 특징 선택 알고리즘과 분류 알고리즘을 선택하였다.

2.2절에서 설명한 특징 선택 기법들 중 속성 공간 탐색 전략은 양방향 최적 우선 탐색 (bi-directional best-first search)을 사용하고, 속성 부분집합을 평가하는 기준은 상관성 기반 특징 선택 방법, 일관성 기반 특징 선택 방법, 의사결정 나무 C4.5 알고리즘을 사용한 래퍼 기반 특징 선택 방법, 선형 SVM 알고리즘을 사용한 래퍼 기반 특징 선택 방법을 적용하여 보았다. 각 특징 방법들의 적용으로 만들어진 데이터에 C4.5 알고리즘[23], K값이 3인 K-최근접 이웃 알고리즘, 베이지안 네트워크 알고리즘, 나이브 베이즈 알고리즘[25], SVM 구축에는 RBF 커널과 회귀 모델을 적용한 순차최소최적화 (sequential minimal optimization, SMO) 알고리즘[26], 역전파 메커니즘을 이용한 다층 퍼셉트론 알고리즘, 그리고 기저함수를 구하기 위해 K값이 2인 K-평균 군집화(K-means clustering)와 로지스틱 회귀(logistic regression)를 이용한 RBF망 알고리즘[20]들을 사용하여 단백질 상호작용 검증을 위한 분류기를 구축하고 이들의 성능을 비교 분석 하였다.

## 4. 실험 결과와 논의

단백질 상호작용 데이터베이스는 Microsoft SQL 2005를 이용하여 구축되었고, 데이터 전처리 과정은 Microsoft Visual Studio .Net에 기반해서 구현되었다. 다양한 분류 알고리즘에 기반하여 생성된 상호작용하는 단백질 쌍의 신뢰도를 체크하는데 사용될 분류기들은 Weka 3.4.8 버전을 이용하여 구축하였다[26].

### 4.1 특징 선택 결과 및 논의

본 연구의 결과를 기존 연구와 비교하기 위해 Jansen *et al.* (2003) 논문에서 경험적 지식에 근거하여 사용한 4가지 속성도 사용하여 보았다[16]. 각 특징 선택 방법에 기반하여 선별된 속성들은 <표 2>와 같이 방법에 따라 선택된 속성의 개수 및 종류가 다르게 나타났다. MIPS functional similarity는 모든 특징 선택 방법에서 선별되었으며, mRNA co-expression와 GO functional similarity는 4가지 방법에서 선택되었고, Marginal essentiality는 3가지 특징 방법에서 선택되었다. 이 네 가지 생물학적 속성들은 기본 원리가 서로 다른 특징 선택 방법들에 의해 반복적으로 선택되었으므로, 단백질 상호작용 검증에 근거로서 매우 유용하게 사용할 수 있는다는 것을 알 수 있다. 빈번하게 선택된 속성들은 대부분이 적용 범위(coverage)가 넓은 속성들이었으며, 학습 데이터에서의 적용 범위가 2% 미만인 속성들은 거의 선택이 안되었다. 적용범위가 2% 미만이지만 일관성 기반 특징 선택 방법에 의해 선택된 synthetic lethality는 값을 가지고 있는 97개의 예제들 중 95개(97.9%)가 양성 데이터이고 나머지 2개(2.1%)만이 음성 데이터이어서 데이터에의 적용 범위는 매우 작지만 일관성이 높은 속성으로 채택되었다. 이와 같이 일관성 기반 특징선택은 속성 값들이 어느 클래스에 치중해있는 지에 기준하여 속성을 선택하는 경향을 보였다. 반면, 래퍼형 특징 추출 방법들은 값의 적용 범위가 큰 속성들을 선택하는 경향을 보였다.

<표 2> 특징 선택 방법에 따라 선택된 속성 부분집합들

| 분류 | 개수 | 선택된 속성들 |
|---|---|---|
| 경험적 지식 (Jansen et al.) | 4 | mRNA co-expression, MIPS functional similarity, GO functional similarity, Co-essentiality |
| 상관성 | 6 | mRNA co-expression, MIPS functional similarity, GO functional similarity, Co-regulation, Phylogenetic profiles, Interlogs in another organism |
| 일관성 | 9 | mRNA co-expression, MIPS functional similarity, GO functional similarity, Co-essentiality, Absolute mRNA Expression, Marginal essentiality, Absolute protein abundance, Phylogenetic profiles, Synthetic Lethality |
| 래퍼 C4.5 | 4 | mRNA co-expression, MIPS functional similarity, GO functional similarity, Marginal essentiality |
| 래퍼 SVM | 3 | MIPS functional similarity, Absolute mRNA expression, Marginal essentiality |

### 4.2 분류 기법 적용 결과 및 논의

각 분류기의 성능은 10-fold cross validation을 사용해서 비교 분석하였다. 10-fold cross validation은 사용 가능한 데이터들을 서로 독립적인 10개의 데이터셋으로 나누어서 10번에 거쳐 하나씩 테스트 데이터로 나머지 아홉 개는 학습 데이터로 사용하며 성능평가를 반복하는 방법이다. 시스템 평가기준을 무엇으로 삼느냐에 따라 분류 결과 해석이 달라질 수 있으므로, 우리는 시스템의 성능을 여러 면에서 평가하기 위해 다양한 성능 평가 기준을 사용하였다. 분류

〈표 3〉 정확도에 기반한 특징 선택과 분류 알고리즘 조합의 성능 평가

| | 래퍼-DT | 래퍼-SVM | 일관성 | 상관성 | 경험적 | 전체 | 평균 |
|---|---|---|---|---|---|---|---|
| DT | 92.01 | 86.97 | 92.24 | 90.85 | 91.18 | 92.24 | 90.92 |
| KNN | 71.98 | 67.23 | 50.96 | 50.00 | 61.15 | 50.00 | 58.55 |
| NB | 93.65 | 84.82 | 93.71 | 93.38 | 93.78 | 93.75 | 92.18 |
| BN | 91.41 | 84.77 | 91.84 | 91.15 | 90.75 | 91.95 | 90.31 |
| MP | 95.41 | 82.36 | 96.52 | 96.42 | 95.53 | 97.04 | 93.88 |
| RBFN | 96.01 | 95.22 | 96.02 | 96.38 | 95.56 | 96.63 | 95.97 |
| SVM | 92.73 | 89.02 | 95.10 | 92.67 | 92.76 | 92.92 | 92.53 |
| 평균 | 90.46 | 84.34 | 88.06 | 87.26 | 88.67 | 87.79 | |

결과가 얼마나 정확했는지에 대한 평가 기준은 실제 목표 항목과 결과로써 예측된 목표 항목을 비교한 진양성(True Positive, TP), 위양성(False Positive, FP), 위음성(False Negative, FN), 진음성(True Negative, TN) 데이터의 비율에 기반해서 계산된다.

본 연구에서 사용된 평가 기준은 다음과 같다. 정확도(accuracy)는 전체 데이터셋 중 정확한 목표 항목으로 분류된 데이터의 비율이고 (식 1), 진양성율 (true positive rate, TP-rate)는 실제 해당 목표 항목에 해당하는 데이터들 중 정확하게 예측된 데이터들의 비율을 의미하며 정보 검색 분야의 재현율(recall), 의료정보학 분야의 민감도(sensitivity)와 같은 의미로 사용된다 (식 2). 위양성율 (false positive rate, FP-rate)는 목표 항목에 속하지 않는 데이터 중에서 그 목표 항목으로 잘못 예측된 데이터의 비율을 뜻하며 (식 3), 1에서 위양성율을 뺀 값을 의료정보학 분야에서는 실험 또는 진단 방법의 특이성(specificity)을 평가하는 척도로 사용한다 (식 4). 정밀도(precision)는 특정 목표 항목으로 예측된 데이터들 중 실제로 그 목표 항목에 해당하는 데이터의 비율을 의미한다 (식 5). F-measure는 정밀도와 재현율의 조화평균을 구함으로써 두 평가 기준을 혼합시킨 성능 평가 척도이다 (식 6). 시스템의 분류 성능을 평가함에 있어 서로 다른 시각을 가지고 있는 이 평가 기준들에 기반하여 각 분류기의 성능을 비교분석 하였다 <표4~8>. 각 표에서 행은 분류 알고리즘의 종류를 열은 특징 선택 방법들을 나타낸다. 마지막 열은 각 분류 알고리즘들의 평균 성능을, 마지막 행은 각 특징 선택 방법들의 평균 성능을 나타낸다.

$$\text{Accuracy}=(TN+TP)\circ 100/(TP+FP+FN+TN) \quad \text{(식 1)}$$

$$\text{TP-rate}=(TP)\circ 100/(TP+FN) \quad \text{(식 2)}$$

$$\text{FP-rate}=(FP)\circ 100/(FP+TN) \quad \text{(식 3)}$$

$$\text{Specificity}=1-(\text{FP-rate}) \quad \text{(식 4)}$$

$$\text{Precision}=(TP)\circ 100/(TP+FP) \quad \text{(식 5)}$$

$$\text{F-measure}=2\circ \text{Precision}\circ \text{Recall}/(\text{Precision}+\text{Recall})=(2TP)\circ 100/(2TP+FP+FN) \quad \text{(식 6)}$$

<표 3>는 정확도를 기준으로 하여 특징 선택과 분류 알고리즘 조합에 대한 성능 평가를 정리한 것이다. 단백질 상호작용 데이터의 신뢰도 검증에 있어 다층 퍼셉트론 알고리즘과 특징 선택 작업 없이 전체 생물학적 속성들을 모두 활용 했을 때 검증 판별의 정확도가 97.04%로 매우 높게 나타났다. 의사결정 나무알고리즘의 경우에는 모든 속성을 사용한 경우와 일관성에 기반한 특징 선택 방법을 적용했을 경우 모두 판별 정확도 92.24%로 좋은 성능을 보였다. K-최근접 이웃 알고리즘의 경우는 모든 속성을 사용한 경우와 상관성 기반 특징 선택 방법을 적용했을 때 정확도가 50%로 매우 낮았으며 분류의 근거로 사용하는 특징 수가 많을수록 판별 정확도가 감소하는 경향을 보였다. 또한 K-최근접 이웃 알고리즘은 의사결정 나무를 이용한 래퍼 기반 특징 선택 방법을 사용했을 경우 71.98%로 분류 정확도가 올라가는 경향을 보여주었다. 나이브 베이즈 알고리즘은 경험적으로 선택한 속성들을 사용할 경우에 가장 좋은 정확도를 보여주었고, 베이지안 네트워크는 전체 속성을 다 사용할 경우에 가장 정확도가 높았다. SVM 알고리즘은 일관성 기반 특징 선택 방법과 함께 사용했을 경우에 정확도가 95.10%로 높게 나왔다. 전반적으로 가장 좋은 성능을 보여준 알고리즘들은 다층 퍼셉트론과 RBF망 알고리즘이었다. RBF망 알고리즘은 특징 선택 방법에 상관 없이 평균 95.97%의 정확도를 유지하여 가장 견고한(robust) 알고리즘으로 드러났으며, 그 다음으로는 다층 퍼셉트론 알고리즘이 평균 93.88%의 정확도를 보여주었다. 반면 K-최근접 이웃 알고리즘은 평균 58.55%의 정확도로 가장 오분류 비율이 많은 것으로 나타났다. 특징 선택 방법에 따른 검증 정확도를 살펴보면, 의사결정 나무 알고리즘을 이용한 래퍼 기반 특징 선택 방법이 평균 90.46%의 정확도를 보여주어 전반적으로 특징 선택에 가장 좋은 것으로 나타났으며, 그 다음이 경험적 특징 선택 방법 (88.67%), 일관성 기반 특징 선택 방법(88.06%)이 평균 정확도가 높은 것으로 나타났다.

대부분의 분류 시스템이나 기존의 단백질 상호작용 데이터 검증에 관한 연구에서는 진양성율(민감도)에 기준하여 성능을 평가하는 경우가 많다. (식 2)를 보면 진양성율은 분류 결과의 성능을 진양성 데이터와 위음성 데이터의 개수를 사용하여 평가한다. 즉, 실제로 상호작용하는 단백질 쌍들 중 몇 %가 제대로 예측되었는지를 계산한다. 그렇기 때문에, 분류 및 판별 결과의 성능을 측정함에 있어 위양성 데이터, 즉 실제로는 상호작용하지 않는 단백질 쌍이 분류기를 통해 서로 상호작용하는 단백질 쌍으로 판별되어진 경우들은 무

〈표 4〉 F-measure에 기반한 특징 선택과 분류 알고리즘 조합의 성능 평가

| | 래퍼-DT | 래퍼-SVM | 일관성 | 상관성 | 경험적 | 전체 | 평균 |
|---|---|---|---|---|---|---|---|
| DT | 91.92 | 87.29 | 92.10 | 90.69 | 91.39 | 92.10 | 90.92 |
| KNN | 77.23 | 73.11 | 66.93 | 66.67 | 38.00 | 66.67 | 64.77 |
| NB | 93.89 | 86.54 | 93.95 | 93.65 | 93.52 | 93.99 | 92.59 |
| BN | 91.31 | 85.19 | 91.86 | 91.02 | 90.53 | 91.97 | 90.31 |
| MP | 95.31 | 84.26 | 96.50 | 96.37 | 95.61 | 97.02 | 94.18 |
| RBFN | 96.09 | 95.19 | 96.08 | 96.38 | 95.45 | 96.62 | 95.97 |
| SVM | 92.74 | 89.27 | 95.08 | 92.70 | 92.75 | 92.93 | 92.58 |
| 평균 | 91.21 | 85.84 | 90.36 | 89.64 | 85.32 | 90.19 | |

시하게 된다. 예를들면, 정확도가 50%에 불과했던 K 최근접 알고리즘에 상관성에 기반한 특징 선택 방법을 적용했을 경우와 전체 속성들을 모두 사용했을 경우들은 진양성율이 100%로 계산되었다. 그런데 두 경우에 대한 위양성율도 100%라고 계산되었다. 이렇게 진양성율과 위양성율이 모두 100%라는 상충된 성능평가 결과의 원인은 해당 두 분류기들에서 모든 단백질 상호작용 데이터들을 실제로 상호작용하는 쌍이라고 판별했기 때문이다. 그래서 진양성율 계산 방법에 따르면 실제 상호작용하는 쌍들을 해당 분류기들이 다 찾아주었기 때문에 진양성율이 100%, 위양성율 계산법에 따르면 실제로 상호작용하지 않는 쌍들 모두 상호작용하는 쌍들로 분류되었기 때문에 위양성율이 100%, 그리고 이들의 정밀도도 (식 4)에 따라 0%가 되는 것이다. 진양성율, 위양성율, 정밀도와 같이, 분류 결과를 정리한 진양성, 위양성, 위음성, 진음성 중 두 가지만 사용하여 성능을 평가하는 것은 상황에 따라서 공정한 평가방법이 아닐 수도 있다. 따라서 분류 결과는 진양성, 위양성, 위음성, 진음성 모두를 사용하여 계산하는 정확도나 진음성만 제외한 나머지를 사용하여 성능을 계산하는 F-measure와 같은 평가 방법을 사용하여 비교하는 것이 공정하며, 경우에 따라서는 진양성율과 정밀도를 같이 사용하거나, 진양성율과 위양성율을 같이 사용하는 등, 서로 상충관계에 있으면서 서로 보완이 되는 성능 평가 기준들을 두 개 이상 함께 제시하며 성능을 평가하는 것이 바람직하다.

<표 4>에서 볼 수 있듯이 F-measure로 성능 평가를 한 결과는 정확도에 기준한 성능 평가 결과와 비슷하게 K-최근접 이웃 알고리즘의 단점을 잘 드러내고 있으며, 정확도에 기준한 결과와 일치하게 특징 선택 방법들 중 의사결정 나무를 사용한 특징 선택을 적용했을 경우(평균 91.21%)의 F-measure가 전반적으로 높았고, 알고리즘을 기준으로 봤을 때에도 RBF망(평균 95.97%)의 F-measure가 전반적으로 좋게 나타나 정확도와 같은 결과를 보여주었다. F-measure를 기준으로 봤을 경우에도 특징 선택 과정 없이 다층 퍼셉트론을 사용하는 경우가 97.02%로 가장 성능이 좋게 나타났다. 하지만 특징 선택을 적용하면 거의 최고의 성능을 가지는 단백질 상호작용 검증 시스템을 보다 빠르게 구축할 수 있으므로, 모델 구축 시간을 단축시키기 위해서는 특징 선택을 적용하는 것이 매우 효과적일 것이다.

### 4.3 관련 연구와의 성능 비교

단백질 상호작용을 경험적 생물학적 근거를 사용하여 검증하고자 한 기존 연구들과 본 제안 시스템의 성능을 비교해보면, Deng *et al.* (2003)의 경우에는 상호작용하는 두 단백질 사이의 유전자 발현 상관성과 단백질 기능을 근거 자료로 하여 Maximum likelihood estimation을 사용하여 단백질 상호작용을 검증하였고 성능을 진양성율 70%, 특이성 70%로 보고하였다[18]. Jansen *et al.* (2003)에서는 본 논문과 같은 양성 데이터에 음성 데이터를 2백70만여 개를 모두 학습 데이터로 사용하여 앞서 언급한 4가지 경험적 지식에 근거한 생물학적 속성과 베이지안망 알고리즘을 사용하여 75%의 진양성율을 나타내는 분류 예측 모델을 구축했다고 발표하였다[16]. 또한 Patil, *et al.* (2005)에서는 단백질 도메인 정보와 기능상 유사성, 상동성에 관한 성질을 사용하여 베이지안망을 이용하여 89.9%의 진양성율과 62.8%의 특이성을 보여준 시스템을 발표하였다[19]. 각 시스템마다 사용한 데이터가 다르고 적용한 알고리즘들의 설정들이 다르므로 직접적인 비교는 불가능하나, 본 제안시스템의 신중한 학습 데이터 선정과 근거로 사용할 생물학적 특성들의 모음, 다양한 특징 선택 방법들과 분류 알고리즘들의 적용, 그리고 치우침 없는 성능 평가를 위한 다각적인 성능 평가 등을 미루어 볼 때, 본 논문에서 전체 생물학적 속성들과 다층 퍼셉트론을 이용한 단백질 상호작용 신뢰성 검증 시스템은 97.04%의 정확도와 96.27%의 진양성율, 97.78%의 특이성, 그리고 97.02%의 F-measure 성능을 나타내어 제안 시스템이 기존 연구들보다 매우 좋은 성능을 보여줌을 명확히 보여주었다.

## 5. 결론 및 향후 연구

본 연구에서는 대용량 실험 결과로 나온 단백질 상호작용 데이터의 신뢰도를 검증하기 위해 관련 생물학 데이터를 속성으로 통합하고 이들 중 관련있는 속성들을 특징 선택 기법을 이용하여 선별한 후 데이터 마이닝 분류 기법을 적용하여 진양성과 위양성을 구분하는 분류 시스템을 제안하였고, 이 시스템에 적용할 수 있는 다양한 특징 선택 방법과 분류 알고리즘들의 조합에 따른 성능을 다각적으로 비교 분석하였다. 효모 종에 제안 시스템을 적용해본 결과, 생물학자들의 경험적 지식을 반영한 속성들을 통합하고 특징 선택

과 분류 알고리즘을 조합·적용하며 그 성능을 비교분석 함으로써 기존의 연구들에 비해 오염된 단백질 상호작용 데이터에서 참인 상호작용 데이터들을 매우 높은 정확도로 분류하는 다층 퍼셉트론 알고리즘을 이용한 단백질 상호작용 검증 시스템을 구축할 수 있었다. 또한 적용한 특징 선택 방법과 분류 알고리즘에 따라서 시스템 성능이 달라 시스템 구현 시에 어떠한 특징 선택 방법과 분류 알고리즘을 적용할 것인지를 결정하는 일이 매우 중요함을 보여주었다. K-최근접 이웃 알고리즘과 같은 예제 기반 학습 기법들은 목표항목과 관련 없는 속성에 매우 민감하므로 특징 선택을 미리 수행해줄 경우 성능이 많이 향상되었다. 알고리즘 내부에 속성별 가중치를 자동으로 결정해주는 메커니즘이 있는 RBF망이나 신경망 같은 알고리즘의 경우에는 특징 선택 과정이 오히려 분별 성능을 떨어뜨릴 수도 있으나, 적절한 특징 선택 방법을 사용할 경우에는 시스템 성능은 거의 비슷하게 유지하면서도 모델 학습 비용을 단축시키는 측면에서 큰 효과를 거둘 수 있었다.

상호작용체학과 대사경로학 관련 연구는 신뢰성있는 상호작용 데이터로부터 시작해야 한다. 분류 알고리즘을 사용해서 상호작용하는 단백질들의 신뢰도를 추정하는 우리의 연구는 이런 목적에서 볼 때 매우 의미있고 유용하게 사용될 수 있을 것이다.

## 참 고 문 헌


[1] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global Protein Function Prediction from Protein-Protein Interaction Networks". Nature Biotechnology, Vol.21, pp.697-700, 2003.

[2] M. P. Samanta and S. Liang, "Predicting Protein Functions from Redundancies in Large-scale Protein Interaction Networks". PNAS, Vol.100, pp.12579-12583, 2003.

[3] M., A. Steffen, Petti, J. Aach, P. D'haeseleer, and G. Church, "Automated Modelling of Signal Transduction Networks". BMC Bioinformatics, Vol.3, pp.34-44, 2002.

[4] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, et al. "A Comprehensive Analysis of Protein-Protein Interactions in Saccharomyces Cerevisiae". Nature, Vol.403, pp.623 - 627, 2000.

[5] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, et al. "A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome". PNAS, Vol.98, pp.4569 - 4574, 2001.

[6] A. C. Gavin, M. Bosche, R. Krause, et al. "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes". Nature, Vol.415, pp.141 - 147, 2002.

[7] Y. Ho, A. Gruhler, A. Heilbut, et al. "Systematic Identification of Protein Complexes in Saccharomyces Cerevisiae by Mass Spectrometry". Nature, Vol.415, pp.180 - 183, 2002.

[8] C. von Mering, R. Krause, B. Snel, M. Cornell, et al. "Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions". Nature, Vol.417, pp.399-403, 2002.

[9] E. Sprinzak, S. Sattath and H. J. Margalit. "How reliable are experimental protein-protein interaction data?" Molecular Biology, Vol. 327, pp.919-923, 2003.

[10] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenber, "ProteinInteractions: Two Methods for Assessment of the Reliability of High Throughput Observations". Molecular and Cellular Proteomics, Vol.1, pp.349-356, 2002.

[11] H. Ge, Z. Liu, G. M. Church, and M. Vidal, "Correlation between Transcriptome and Interactome Mapping Data from Saccharomyces Cerevisiae". Nature Genetics, Vol.29, pp.482-486, 2001.

[12] R. Jasen, D. Greenbaum and M. Gerstein, "Relating Whole-genome Expression Data with Protein-Protein Interaction". Genome Research Vol.12, pp.37-46, 2002.

[13] N. Bhardwaj and H. Lu, "Correlation between Gene Expression Profiles and Protein-Protein Interactions within and across Genomes". Bioinformatics vol.21, pp.2730-2738, 2005.

[14] L. R. Mattews, P. Vaglio, J. Reboul, H. Ge, et al. "Identification of Potential Interaction Networks using Sequence-Based Searches for Conserved Protein-Protein Iinteractions or "Interologs"". Genome Research, Vol.11, pp.2120-2126, 2001.

[15] T. Sato, Y. Yamanishi, M. Kanehisa, and H. Toh, "The Inference of Protein-Protein Interactions by Co-evolutionary Analysis is Improved by Excluding the Information about the Phylogenetic Relationships". Bioinformatics Vol.21, pp.3482-3489, 2005.

[16] R. Jansen, H. Yu, D. Greenbaum et al. "A Bayesian Network Approach for Predicting Protein-Protein Interactions from Genomic Data", Science 203, 449-153, 2003.

[17] L. J. Lu, A. Paccanaro, H. Yu, "Assessing the Limits of Genomic Data Integration for Predicting Protein Networks", Genome Research 15, 9455-953, 2005.

[18] M. Deng, F, Sun, T. Chen, "Assessment of the Reliability of Protein-Protein Interactions and Protein Function Prediction", Symp. Biocomputing, 140-151, 2003

[19] A. Patil and H. Nakamura, "Filtering High-throughput Protein-Protein Interaction Data using a Combination of Genomic Features", BMC Bioinformatics, 6:100-112, 2005

[20] I. J. Witten and E. Frank, Data Mining: Practical Machine Learning Tools with Java Implementations. Morgan Kaufmann, San Francisco, CA. 2000.

[21] P. N. Tan, M. Stenbach, V. Kumar, Introduction to Data Mining, Addison Wesley, 2005.

[22] I. Guyon and A. Elisseff, "An introduction to variable and feature selection", Journal of machine learning research, 3, 1157-1182, 2003.

[23] R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA. 1993.

[24] D. Aha and D. Kibler, "Instance-based Learning Algorithms". Machine Learning vol.6, pp.37-66, 1991.
[25] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers". Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence.pp.338-345, Morgan Kaufmann, San Mateo. 1995.
[26] J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization". Advances in kernel methods - support vector learning, Schoelkopf, B., Burges, C. and Smola, A. eds., MIT Press. 1998.
[27] H. W. Mewes, D. Fishman, K. F. X. Mayer, et al, "MIPS: Analysis and Annotation of Proteins from Whole Genomes in 2005", Nucleic Acids Research 34, D169-D172, 2005
[28] U. Guldener, M. Munsterkotter, M. Oesterheld, et al. "MPact: the MIPS Protein Interaction Resource on Yeast", Nucleic Acids Research 34, D436-D441, 2006
[29] The Gene Ontology Consortium, "Gene Ontology: Tool for the unification of biology", Nature Genetics 25, 25-29, 2000
[30] A., Ruepp, A. Zollner, D. Maier, K. Albermann, et al.: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 32, 5539-5545. 2004

**이 민 수**

e-mail : ssue@ewhain.net
2001년 이화여자대학교 수학과(학사)
2003년 이화여자대학교 컴퓨터학과 (공학석사)
2003년~현재 이화여자대학교 컴퓨터학과 박사과정
관심분야 : 바이오인포매틱스, 지식표현, 데이터마이닝, 시멘틱 웹 등

**박 승 수**

e-mail : sspark@ewha.ac.kr
1974년 서울대학교 수학과(학사)
1976년 한국과학기술원 전산학 (공학석사)
1988년 미국 텍사스 대학 전산학 (공학박사)
1988년~1991년 미국 캔사스대학 컴퓨터학과 조교수
1991년~현재 이화여자대학교 컴퓨터학과 정교수
관심분야 : 인공지능, 데이터마이닝, 바이오인포매틱스 등

**이 상 호**

e-mail : shlee@ewha.ac.kr
1979년 서울대학교 계산통계학과 학사
1981년 한국과학기술원 전산학과 석사
1987년 한국과학기술원 전산학과 박사
1983년~현재 이화여자대학교 컴퓨터학과 정교수
관심분야 : 정보보호, 암호프로토콜, 알고리즘 설계, 계산기하, 그래프 드로잉, 데이터 마이닝, 바이오인포매틱스

**용 환 승**

E-mail : hsyong@ewha.ac.kr
1983년 서울대학교 컴퓨터공학과 학사
1985년 서울대 대학원 컴퓨터공학과 공학석사
1985년~1989년 한국전자통신연구소 연구원
1994년 서울대 대학원 컴퓨터공학과 공학박사
2002년 8월~2003년 2월 IBM T.J. Watson 연구소 객원연구원
1995년~현재 이화여자대학교 컴퓨터학과 부교수
관심분야 : 데이타베이스 시스템, 데이터 마이닝, 유비쿼터스 컴퓨팅

**강 성 희**

e-mail : kangsh@mju.ac.kr
1991년 이화여자대학교 전자계산학과 학사
1995년 이화여자대학교 대학원 전자계산학과 공학석사
2001년 이화여자대학교 과학기술대학원 컴퓨터학과 공학박사
2001년~현재 명지대학교 방목기초교육대학 부교수
관심분야 : 인공지능, 에이전트, 데이터마이닝, 바이오인포매틱스