최적 연관 속성 규칙을 이용한 비명시적 단백질 상호작용의 예측

(Prediction of Implicit Protein - Protein Interaction Using Optimal Associative Feature Rule)

엄 재 홍 † 장 병 탁 ††
(Jae-Hong Eom) (Byoung-Tak Zhang)

요 약 단백질들은 서로 다른 단백질들과 상호작용 하거나 복합물을 형성함으로써 생물학적으로 중요한 기능을 한다고 알려져 있다. 때문에 대부분의 세포작용에 있어 중요한 역할을 하는 단백질 상호작용의 분석 및 예측에 대한 연구는 여러 연구그룹으로부터 풍부한 데이타가 산출되고 있는 현(現) 게놈시대에서 또 하나의 중요한 이슈가 되고 있다. 본 논문에서는 효모(*Saccharomyces cerevisiae*)에 대해 공개되어있는 단백질 상호작용 데이타들에서 속성들 간의 연관을 통해 유추 가능한 잠재적 단백질 상호작용들을 예측하기 위한 연관속성 마이닝 방법을 제시한다. 단백질의 속성들 중 연속값을 가지는 속성값들은 최대상호의존성에 기반을 두어 이산화 하였으며, 정보이론기반 속성선택 알고리즘을 사용하여 단백질들 간의 상호작용 예측을 위해 고려되는 단백질의 속성(attribute) 수 증가에 따른 속성차원문제를 극복하도록 하였다. 속성들 간의 연관성 발견은 데이타마이닝 분야에서 사용되는 연관규칙 발견(association rule discovery) 방법을 사용하였다. 논문에서 제안한 방법은 발견된 연관규칙을 통한 단백질 상호작용 예측문제에 있어 최대 약 96.5%의 예측 정확도를 보였으며 속성필터링을 통하여 속성필터링을 하지 않는 기존의 방법에 비해 최대 약 29.4% 연관규칙 발견속도 향상을 보였다.

키워드 : 단백질 상호작용, 연관속성 마이닝, 연관규칙, 데이타마이닝, 바이오인포메틱스

Abstract Proteins are known to perform a biological function by interacting with other proteins or compounds. Since protein interaction is intrinsic to most cellular processes, prediction of protein interaction is an important issue in post-genomic biology where abundant interaction data have been produced by many research groups. In this paper, we present an associative feature mining method to predict implicit protein-protein interactions of *Saccharomyces cerevisiae* from public protein interaction data. We discretized continuous-valued features by maximal interdependence-based discretization approach. We also employed feature dimension reduction filter (FDRF) method which is based on the information theory to select optimal informative features, to boost prediction accuracy and overall mining speed, and to overcome the dimensionality problem of conventional data mining approaches. We used association rule discovery algorithm for associative feature and rule mining to predict protein interaction. Using the discovered associative feature we predicted implicit protein interactions which have not been observed in training data. According to the experimental results, the proposed method accomplished about 96.5% prediction accuracy with reduced computation time which is about 29.4% faster than conventional method with no feature filter in association rule mining.

Key words : Protein-protein interaction, Feature association mining, Association rule, Data mining, Bioinformatics

· 이 연구는 과학기술부 국가지정연구실사업(NRL)에 의하여 일부 지원되었음

† 학생회원 : 서울대학교 전기컴퓨터공학부
jheom@bi.snu.ac.kr
†† 종신회원 : 서울대학교 전기컴퓨터공학부 교수
btzhang@bi.snu.ac.kr

논문접수 : 2005년 8월 25일
심사완료 : 2006년 2월 22일

1. 서 론

1.1 서론

게놈관련 기술의 발전과 생물에 대한 게놈수준의 연구기술이 발전함에 따라, 유전자 생성물이 세포내에서 어떻게 시·공간(視·空間)적으로 행동하고 어떠한 예정

된 단계를 거쳐 최종 산물인 단백질의 유전 정보가 결정되며, 어떻게 단백질들마다 정해진 기능을 달성하고 생물체의 기관 형성을 위하여 상호작용 하는지에 대한 이해는 게놈시대의 가장 큰 도전 과제중의 하나가 되었다. 또한 단백질-단백질의 상호작용(PPI; protein-protein interaction)이 생물의 기관에서 일어나는 매우 기초적인 생화학 반응들 중의 하나로 여러 가지 생물학적 반응 과정에서 중요한 기능을 한다는 것은 이미 여러 연구자들에 의해 밝혀졌으며 지금까지도 다양한 생물학 도메인(domain)에 대하여 활발하게 연구되고 있다[1]. 이에 따라 단백질들 간의 상호작용에 대한 포괄적인 이해와 함께 보다 심층적인 분석은 여러 가지 생물학적 현상들과 문제들의 이해에 중요한 기여를 할 수 있을 것으로 전망되고 있다.

게놈시대에 활발하게 연구되고 있는 여러 생물체들 중에서 효모(yeast)는 비교적 단순한 구조, 높은 대사활성 및 빠른 성장속도와 함께 그 응용의 다양성 등으로 인해 그동안 집중적인 연구가 이루어진 대표적인 생물 중의 하나이다. 주로 자낭균류와 담자균류에서 발견되는 효모는 균류에 속하는 단세포 생물로 흙 속이나 식물의 표면에서도 발견되고, 특히 당분이 많은 꽃이나 과일의 표면에 많이 붙어 있다. 효모는 치마아제(zymase), 인베르타아제(invertase)와 같은 효소를 가지고 있어 당분을 에탄올과 이산화탄소로 분해한다. 효모에 의해 생기는 에탄올은 술의 주성분이 되며 이산화탄소는 빵을 부풀게 한다. 때문에 효모의 이 같은 당 발효를 통한 에탄올 및 이산화탄소 생산 성질은 오래전부터 맥주의 제조나 빵의 발효에 널리 이용되어왔다.

효모의 생활사는 1935년 Winge에 의하여 처음으로 세대교번이 밝혀졌으며 1960년에는 염색체지도가 완성되어 유전학의 좋은 재료로 사용되어왔다. 약 4년간의 연구 기간을 거쳐 1997년 발아효모의 DNA 염기서열 해독(sequencing)이 완료됨에 따라 이후 많은 연구자들이 약 6,300여개가 넘는 효모 단백질들에 대한 기능적 분석과 관련된 연구를 수행하였으며, 이에 따라 효모의 단백질-단백질 상호작용과 관련된 풍부한 실험적 데이타 및 분석적 데이타가 존재하게 되었고 다양한 분야의 여러 방법들이 이러한 데이타의 분석에 적용되어 긍정적인 결과를 보여 왔다[2].

1.2 관련 연구

유전자발현(gene expression) 데이타나 단백질 상호작용 데이타와 같은 여러 데이타들을 이용하여 단백질의 기능이나 기존에 관찰되지 않은 단백질들 간의 상호작용을 예측하기 위한 여러 가지 연구들이 수행되어 왔다. Eisen 등과 Pavlidis 등은 유사 기능을 하는 유전자들이 함께 발현되는 경우가 높다는 점을 이용한 유전자 발현 데이타에 대한 군집화(clustering) 분석을 적용하여 의미 있는 결과를 얻었다[3,4]. Wu 등은 효모 염기서열의 전사 클러스터(transcriptional cluster)간 중복성 분석을 통하여 효모 단백질의 기능을 분석하는 서열분석 기반의 연구를 수행하였다[5]. 또한 Park 등은 진화적으로 연관성 있는 단백질 도메인들의 구조적 그룹(structural family)들 간의 상호작용 분석법을 이용하여 단백질들 도메인들 간의 상호작용을 분석하였으며[6], Iossifov 등과 Ng 등은 기존에 밝혀진 단백질들 간의 상호작용 데이타를 이용하여 각각 확률추론(probabilistic inference)과 계산학적 추론(computational inference) 방법을 이용하여 새로운 상호작용을 예측하였다[7,8].

Fields[9]등에 의해 소개된 two-hybrid system은 효모에 응용되어 새로운 단백질 상호작용 분석방법인 결합단백질탐색법(Y2H; yeast two-hybrid)을 가능케 하였고, 이 방법은 주어진 단백질들 간의 모든 가능한 조합에서 가능한 단백질-단백질 상호작용을 검출할 수 있게 하였다. Ito 등은 Y2H 방법을 이용하여 발아효모(budding yeast) 단백질들의 기능적 상호작용에 대한 포괄적인 분석을 하였고[10,11], Uetz 등은 어레이스크리닝(array screening)과 라이브러리스크리닝(library screening)을 함께 이용하여 효모 단백질의 광범위한 단백질-단백질 상호작용을 조사하는 연구를 수행하였다[12].

단백질 상호작용 데이타가 풍부해지면서 상호작용 네트워크의 구조기반 연구들도 진행되었다. Bu 등은 기존의 효모 단백질들 간의 상호작용 데이타를 이용하여 상호작용 네트워크를 구성한 후 이 네트워크의 구조 분석을 통해서 주요 상호작용을 추출함으로써 추가적인 상호작용을 분석하였다[13]. 최근에는 Bu 등의 분석법과 같이 기존에 실험적으로 구성된 단백질들 간의 상호작용 네트워크를 분석하려고 하는 여러 가지 연구들이 제시되고 있으며 다양한 접근방법들도 제안되었다[14,15].

앞서 살펴본 것처럼 생물학 분야의 데이타가 점차 거대해짐에 따라 방대한 양의 데이타로부터 유용한 정보나 사실을 발굴하기 위한 '데이타마이닝(data mining)' 기법들이 관심을 끌기 시작하였으며, 지금까지 바이오인포메틱스 분야에 데이타마이닝 기법을 활용하는 여러 가지 연구들이 진행되어왔다. 가장 대표적인 데이타마이닝 방법은 Agrawal 등에 의해 개발된 '연관규칙 발견(association rule discovery)'[16] 방법으로, Satou등은 다양한 게놈 데이타에서 추출된 단백질의 서열이나 구조, 그리고 기능 등과 같은 이종(heterogeneous) 데이타들 간의 연관성을 찾기 위하여 연관규칙 발견 방법을 사용하였다[17]. 이와 같은 데이타마이닝 기법은 DNA 마이크로어레이(microarray) 데이타를 분석하는 데에도 활용되었다[18]. 이처럼 데이타마이닝기법은 생물정보학

의 여러 분야에서 성공적으로 응용되어왔다. 이러한 데이타마이닝의 성공적 응용의 예로 Fellenberg등의 연구를 들 수 있는데 그들은 데이타마이닝 기법을 이용하여 단백질의 상호작용 데이타를 분석하는 통합 시스템을 구성하였다. 이 시스템은 당시까지 밝혀지지 않은 단백질 각각의 특징을 분석하는 목적으로 사용되었다[19]. 그러나 Fellenberg등의 시스템은 단백질 각각의 속성을 밝혔을 뿐 하나의 단백질이 다른 단백질과 작용하는 보다 일반적인 규칙을 발견하지는 못하였다. 이러한 한계를 해결하기 위하여 Oyama등은 데이타마이닝 기법을 이용하여 단백질간의 상호작용 데이타에서 연관규칙을 찾는 기초 연구를 수행하여 의미 있는 결과를 얻었다[20].

앞에서 살펴본 것과 같이 데이타마이닝 방법들이 생물학 데이타를 이용한 마이닝을 위해 널리 사용되었지만 일반적으로 수천 개 이상의 속성값을 가질 수 있는 생물학 분야의 데이타를 처리하는데 있어 '속성차원(feature dimension)'의 증가에 다른 여러 가지 문제를 안고 있었다. 데이타마이닝에서 각 데이타 개체(instance)를 표현하는 속성들은 일반적으로 차원(dimension)으로 고려될 수 있다. 실제로, Oyama[20]등은 약 5,240개가 넘는 속성을 가지는 데이타를 이용하여 데이타마이닝 작업을 수행하였는데, 연관규칙 발견 방법의 경우 규칙 발견을 위해 고려해야 할 속성들이 증가함에 따라 전체 규칙발견 과정의 속도 및 발견된 규칙의 정확성이 이러한 속성차원에 의해 적지 않은 영향을 받는다.

본 논문에서는 기존에 연구된 연관규칙 발견을 통한 상호작용 예측 방법을 기반으로 이에 상호작용 단백질들의 연속값을 가지는 속성들에 대한 이산화 방법과 효율적인 속성차원축소(feature dimension reduction) 방법을 적용함으로써 향상된 단백질 상호작용 예측 방법을 제시한다. 이를 위해 본 논문에서는 Yu[21] 등에 의해 제안된 정보이론 기반의 속성 선택 방법의 기본 흐름을 사용하였으며 Oyama[20] 등에 의해 생물학 데이타에 대해 최초로 수행된 연관규칙 발견을 통한 단백질 상호작용 예측 개념을 사용하였다. 본 논문에서는 단백질-단백질 상호작용을 서로 상호작용하는 각 단백질들의 '속성-속성'의 연관(feature-feature association)으로 고려하여 새로운 상호작용의 예측을 수행한다. 상호작용하는 각각의 속성들에 대한 고찰을 통해서 얻어진 연관성 정보를 기반으로 새로운 단백질-단백질 상호작용을 예측하기 위하여 본 논문에서는 가능한 한 각 단백질들에 대하여 한 풍부한 속성값들을 사용하기 위해 MIPS (Munich Information Center for Protein Sequences의 Yeast Genome Database)[1], DIP(Database of Interacting Proteins)[2] 및 SGD(Saccharomyces Genome Database)[3]와 같은 효모에 대한 공인된 공개 데이타베이스를 보조적으로 사용하여 단백질들의 추가적인 속성을 수집하였다.

논문에서 사용하는 정보이론 기반의 '속성차원축소필터(FDRF; Feature Dimension Reduction Filter)'는 상호작용 예측 작업에 대해 가장 정보성이 높은 속성들을 선택을 통해 전반적인 규칙발견 과정의 속도 및 예측 정확성을 향상시키기 위해 사용되었다. 논문에서는, 단백질-단백질 상호작용 예측에 대해 유용한 속성들을 선택함으로써 전체 속성차원을 축소 한 후 연관규칙 발견 방법을 적용하여 속성들 간의 연관성 정보를 추출하고 이를 이용하여 훈련데이타에 제공되지 않은 테스트 데이타에 존재하는 상호작용(묵시적 상호작용; implicit interaction)을 예측함으로써 제안한 방법의 예측 정확률(accuracy rate)을 측정하였다.

1.3 논문의 구성

본 논문은 다음과 같이 구성된다. 먼저 2장에서는 연속값을 갖는 속성값을 이산화하는 방법 및 고차 속성집합에서 유의미한 정보를 제공하는 속성들을 선택하기 위한 '속성차원 축소필터'에 대한 기본 개념과 필요한 정의 및 전체 과정에 대해 기술하며, 3장에서는 논문에서 사용한 연관규칙 발견 방법을 이용한 '단백질-단백질' 상호작용의 '속성-속성' 연관 발견 방법을 통한 새로운 단백질-단백질 상호작용 예측 방법에 대해 기술한다. 4장에서는 실험에 사용한 단백질-단백질 상호작용 데이타들과 상호작용하는 각각의 단백질들에 대해 고려된 속성에 대한 설명과 실험 결과 및 분석 내용을 다룬다. 마지막으로 5장에서는 논문의 결론 및 논문에서 다룬 방법의 전반적인 성능 향상을 위한 향후 연구과제에 대해 기술한다.

2. 속성차원의 축소(Feature Dimension Reduction)

2.1 속성 이산화(Feature Discretization)

본 논문에서는 단백질-단백질 상호작용을 예측하기 위하여 상호작용이 알려진 단백질들에 대하여 MIPS나 SGD와 같은 공개 데이타베이스에서 단백질의 특징을 나타내는 여러 가지 속성들을 보조적으로 수집하여 사용하였다. 이렇게 수집된 단백질의 속성들 중에는 크게는 게놈상의 위치(locus)와 같은 속성과 함께 보다 세부적으로는 반데발스 볼륨(van der Waals volume), 등전점(isoelectric point) 등과 같이 연속적(continuous) 값을 가지는 속성들이 여럿 존재한다. 연속값을 가지는 이러한 속성들은 3.2절에서 설명하고 있는 방법과 같이 이

1) http://mips.gsf.de/genre/proj/yeast/
2) http://dip.doe-mbi.ucla.edu/
3) http://www.yeastgenome.org/

진코딩(binary coding)을 이용하여 속성들 간의 연관규칙을 추론하기 위해서는 적절하게 이산화(discretization)되어야 한다.

논문에서는 단백질-단백질 상호작용 예측을 위해 사용한 단백질들의 연속적 속성들을 이산화하기 위하여 속성을 점진적으로 분할하여 가는 Kurgan와 Cios의 하향식(top-down) 이산화 방법을 사용하였다[22]. 논문에서 사용한 하향식 이산화 방법은 우선 연속값을 가지는 각 속성변수들의 간격(interval)을 임시로 분할점을 기준으로 나누는 것으로 시작된다. 다음으로 임시 분할점에 의하여 나누어진 속성변수들에 대하여 해당 속성변수와 클래스와의 상호의존성 검정(class-attribute interdependence test)을 수행한다. 이때 계산된 상호의존성 값이 임시 분할점에 의하여 줄어들지 않는 경우 해당 분할점은 최종 분할점 목록에 추가되고 연속값을 가지는 속성변수는 이 분할점을 기준으로 나뉘어 이산화 된다. 이러한 과정은 연속값을 가지는 각 속성변수들에 대하여 반복적으로 적용되어 속성값을 분할해나간다. 이산화 과정은 궁극적으로 연속적 값을 가지는 각 속성값에 대하여 최종적으로 나누어진 이산(離散) 간격들의 수(속성차원)를 최소화함과 동시에 나누어진 간격들이 클래스 라벨에 대하여 갖는 상호의존성 값을 최대로 유지하여 이산화 과정에서 손실되는 정보량을 최소화 하도록 한다.

이산화를 수행하려는 전체 데이타에서 훈련 데이타집합이 총 M개의 예제들을 가지고 있고 각각의 예제들은 전체 클래스 집합 S중 하나의 클래스에 속하고, 연속값을 가지는 연속속성 F에 대한 이산화 방법을 D라 하면 전체 이산화 과정은 다음과 같이 진행된다. 우선 연속 도메인 속성 F를 속성-클래스 상호의존성 계산을 통하여 n개의 간격으로 나눈다고 하면, F에 대한 이산화 방법 D는 $D = \{[d_0,d_1],(d_1,d_2], \cdots ,(d_{n-1},d_n]\}$로, 즉 연속값을 가지는 속성의 값을 나누는 최종 범위들의 집합으로 구성될 수 있다. D에서 '['와 ']'는 각각 하한 값과 상한 값을 포함하는 경계를 나타내며 '('는 하한 값을 포함하지 않는 경계를 의미한다. 이러한 이산화 방법 D가 가지는 연속값들의 최종 분할범위들은 분할기준점들의 임시 집합인 경계집합(boundary set) $B = d_0,d_1,d_2, \cdots ,d_{n-1},d_n$에서 선택되어진 것으로 $D \subseteq B$의 관계가 성립하게 된다. 속성 F의 각 값들은 앞에서와 같이 나누어진 전체 n개 구간중 하나에 속하게 되며 특정 이산화 간격에 속하는 속성 F의 값들이 갖는 소속값(membership value)은 임시 경계집합 B에서 선택되어 새로이 구성되는 이산화방법 D에 따라 변하게 된다. 이때 각 데이타의 클래스를 의미하는 클래스 변수와 경계집합 B의 원소로 표시되는 이산변량(變量)들, 즉 임시 분할경계값들은 일종의 확률변수(random variable)로 생각할 수 있으며 이 두 확률변수에 대해서는 다음의 표 1과 같이 전체 클래스들과 분할간격들에 대한 이산화 행렬을 구성할 수 있다[22].

표 1에서 C_i는 각각의 클래스를, q_{ir}은 i번째 클래스에 대해서 구분간격 $(d_{r-1},d_r]$에 속하는 연속값들의 총 개수를 의미한다. M_{i+}와 M_{+r}은 각각 i번째 클래스에 속하는 전체 객체(이산화된 값들)의 수와 이산화 간격 $(d_{r-1},d_r]$에 속하는 속성 F의 연속값들의 수를 의미하며 색인변수 $i=1,2,...,S$로 전체 클래스의 개수를, $r=1,2,...,n$로 연속속성이 분할된 전체 구간의 개수를 의미한다.

이산화 과정에서 연속값의 분할 위치는 이산화를 통하여 생성되는 이산화간격들과 클래스 라벨과의 상호의존성이 최대가 되는 지점을 선택하게 되며, 이러한 상호의존성은 식 (1)과 같이 속성값과 클래스 변수 사이의 최대상호의존성 척도 θ(maximal interdependency)에 의해 평가된다.

$$\theta(C,D|F) = \frac{\sum_{r=1}^{n} \frac{\max_r^2}{M_{+r}}}{n}. \quad (1)$$

식 (1)에서 n은 연속값을 나누는 전체 간격의 수를 의미하며 $\max_r$은 표 1에서 $i=1,2,...,S$에 대하여 q_{ir}의 최댓값을 의미하여 n은 θ를 전체 이산화 간격의 수로 정규화하기 위해 사용되었다. 표 2는 식 (1)의 최대상호

표 1 클래스 변수와 경계집합에 대한 이산화 행렬

클래스 (class)	변수의 이산화 간격들(variable intervals) $[d_0,d_1] \cdots (d_{r-1},d_r] \cdots (d_{n-1},d_n]$	전체 클래스들 (total classes)
C_1	$q_{11} \cdots q_{1r} \cdots q_{1n}$	M_{1+}
$\vdots$	$\vdots \ddots \vdots \ddots \vdots$	$\vdots$
C_i	$q_{i1} \cdots q_{ir} \cdots q_{in}$	M_{i+}
$\vdots$	$\vdots \ddots \vdots \ddots \vdots$	$\vdots$
C_S	$q_{S1} \cdots q_{Sr} \cdots q_{Sn}$	M_{S+}
전체 구간들 (total intervals)	$M_{+1} \cdots M_{+r} \cdots M_{+n}$	M

표 2 최대상호의존성 기반 연속값 속성의 이산화 과정

입력: ($\boldsymbol{M}$, $\boldsymbol{S}$, $\boldsymbol{F_i}$)

$\boldsymbol{M}$: 각각 클래스 레이블을 가지는 전체 예제들의 집합.

$\boldsymbol{S}$: 각각의 예제들이 속하는 클래스들의 집합.

$\boldsymbol{F_i}$: 연속값을 가지는 연속속성들 중 이산화 할 i번째 속성.

출력: ($\boldsymbol{D}$)

$\boldsymbol{D}$: 연속속성 F_i를 이산화하기 위한 최종 이산화 경계점들의 집합

이산화 알고리즘

1. 초기화 단계:

1-1. $d_0 \leftarrow \min(F_i)$, $d_n \leftarrow \max(F_i)$

1-2. $make\_distinct(F_i)$, $sort_{ascending}(F_i)$

1-3. 임시 경계집합 $B \leftarrow \left(d_0, \frac{d_i + d_{i+1}}{2}, d_n\right)$, $i = 0,...,n-1$

1-4. $c \leftarrow |B|$, $D \leftarrow \{[d_0, d_n]\}$, $\theta \leftarrow 0$, $\theta^t \leftarrow 0$

2. 분할점탐색 단계:

2-1. $k \leftarrow 1$

2-2. 각각의 $d_j \notin D$ ($j = 1,...,c$)에 대하여 $D \leftarrow D \cup B(d_j)$ 및
식 (1)을 이용하여 각각의 θ_j^t값 계산

2-3. 단계 (2-2)에서 계산된 θ_j^t값들 중에서 최댓값을 가지는 θ_j^t의 d_j 선택

2-4. (a) $\theta_j^t > \theta$이고 $k < S$ 인 경우 아래의 갱신작업 후 (2-2)부터 반복

$D \leftarrow D \cup d_j$ (이산화 방법에 연속값 분할경계 추가)

$\theta \leftarrow \theta_j^t$ (현재까지 계산된 최대 상호의존성 값 갱신)

$k \leftarrow k+1$ (전체 진행단계 카운트변수 갱신)

(b) $\theta_j^t \leq \theta$이거나 $k \geq S$인 경우 분할점 탐색단계 종료, D 반환

의존성 값 θ를 이용하여 연속값을 이산화하는 알고리즘을 나타낸다. 표 2에서 $make_distinct(F_i)$는 속성 F_i의 값들 중에서 중복된 값들을 제거하여 유일한 속성값들의 집합으로 만들어주는 함수를, $sort_{ascending}(F_i)$는 속성 F_i의 값들을 속성값 기준으로 오름차순으로 정렬해 주는 함수를 의미한다.

2.2 속성차원의 축소

게놈프로젝트와 같은 생물학 관련 여러 응용데이타의 크기는 열과 행, 즉 데이타의 수와 데이타가 가지는 속성의 면에서 모두 점차 방대해지고 있다. 때문에 데이타의 차원을 의미하는 속성이 다수 존재하는 고차원(high dimensional) 데이타를 기계학습 방법을 이용하여 보다 효율적으로 분석하기 위해서는 '속성 선택(feature selection)' 과정이 요구된다[21]. 또한, 각각의 객체가 가지는 다수의 속성들은 앞서 2.1절에서 다룬 '속성 이산화' 단계를 통하여 각각 이산화되기 때문에 전체적으로 고려해야 하는 속성 차원은 더욱 증가한다.

속성 선택은 속성 집합에서 특정 평가 척도에 따라서 부분 속성 집합을 선택함으로써 본래의 속성 공간(feature space)을 줄이는 것을 의미한다. 일반적으로, 속성은 클래스 개념에 관한 명확한 긍정·부정적 정보를 제공하면서 다른 속성들과 중복되지 않는 경우 해당 클래스 개념을 다루는 문제에 있어서 좋은 속성으로 간주된다. 속성의 선택은 속성 자체와 속성이 속하는 클래스를 각각 확률변수로 고려하여 처리할 수 있다.

두 확률변수 사이의 상관관계는 크게 두 가지 방법으로 측정할 수 있는데 하나는 전통적인 선형상관관계(linear correlation)에 기반을 둔 방법이고 다른 하나는 엔트로피를 이용한 정보이론에 기반을 둔 방법이다. 선형상관계수(linear correlation coefficient)나 최소자승에러(least square regression error), 또는 최대정보압축색인(maximal information compression index)과 같은 선형상관관계 기반의 확률변수들 간의 상관관계 측정 방법들은 몇 가지 이점을 가지고 있다. 우선, 이러한 접근 방법들은 속성들 중에서 고려하는 클래스 개념에 대한 상관관계가 거의 '0'인 것들을 대부분 제거할 수 있으며 최종적으로 선택된 속성들 간의 중복성도 제거할

수 있다. 그렇지만 선형상관관계 기반의 확률변수들 간의 상관관계 측정 방법들은 본질적으로 선형적 속성을 가지지 않는 상관관계들을 제대로 찾아내기가 힘들며 상관관계의 계산을 위해서는 모든 속성들이 수치값(numerical value)을 가져야 한다는 제약이 따른다. 따라서 본 논문에서는 정보이론 기반의 상관관계 분석을 이용하여 선형상관관계 기반 확률변수들 간의 상관관계 측정 방법의 이러한 약점들을 극복하도록 시도하였다.

먼저, 데이타의 각 속성은 확률변수로 볼 수 있는데, 어떤 확률변수의 불확실성은 엔트로피(entropy)로 측정이 가능하며 어떤 확률변수 X의 엔트로피는 다음과 같이 정의된다.

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)), \quad (2)$$

두 확률변수 X와 Y를 고려할 때, 확률변수 Y를 관찰한 이후의 변수 X의 엔트로피는 다음과 같이 정의된다.

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)), \quad (3)$$

식 (3)에서 $P(y_j)$는 변수 Y의 인스턴스인 변수 y_j의 사전확률(prior probability) 값이며 $P(x_i|y_j)$는 주어진 변수 Y에 대해 변수 X가 x_i의 값을 가질 사후확률(posterior probability) 값을 의미한다. 변수 X에 대해서 또 다른 변수 Y가 주어짐으로써 줄어드는 변수 X에 대한 정보의 분량을 '정보이득(IG: information gain)'[23]이라 부르며, 정보이득은 다음과 같이 계산할 수 있다.

$$IG(X|Y) = H(X) - H(X|Y), \quad (4)$$

일반적으로 세 확률변수 X, Y 및 Z에 대해서 정보이득의 정도가 $IG(X|Y) > IG(Z|Y)$와 같은 경우 Y가 Z보다 X에 보다 큰 상관성을 갖는다고 정의한다. 속성들 간의 상관관계와 정보이득을 측정하기 위해서는 사용된 평가척도가 '대칭성(symmetry)'을 가져야 하지만 앞서 살펴본 정보이득의 정의는 보다 많은 값을 가지는 속성들에 대하여 그 값이 커지는 일종의 편향적 특성을 갖기 때문에 모든 속성들이 동일한 영향력을 갖기 위해서는 각 속성들의 값이 모두 정규화 되어야 할 필요가 있다. 때문에 본 논문에서는 이러한 편향적 특성을 가지지 않는 '대칭적불확실성(symmetrical uncertainty)'을 속성들 간의 상관관계를 측정하는데 사용하기로 하며 이러한 대칭적불확실성은 다음과 같이 정의 된다. 식 (5)에서 $SU(X, Y)$는 $0 \le SU(X, Y) \le 1$의 값을 갖는다[24].

$$SU(X, Y) = 2\left[\frac{IG(X|Y)}{H(X) + H(Y)}\right]. \quad (5)$$

표 3은 식 (5)의 SU값을 이용하여 확률변수들 간의 연관성 측정을 하는 '속성차원축소필터'의 전체 과정을 나타낸다. 표 3에서 $sort_{descending}(S'_{list})$는 S'_{list}의 속성값들을 $SU_{i,c}$값을 기준으로 내림차순으로 정렬하는 것을 의미하며 $first_element(S'_{list})$는 정렬된 속성목록 S'_{list}의 맨 앞 속성의 선택을, $following_element(S'_{list}, f_p)$는 속성목록 S'_{list}에서 f_p다음 값을 변환하는 함수를 의미한다.

표 3의 속성차원축소필터는 전체 속성집합에서 문제에서 고려하는 클래스 개념에 적합한 주요 속성부분집합(feature subset)을 크게 두 단계를 거치며 찾는다. 첫 번째 단계는 속성선택 단계로, 속성차원축소필터는 각각의 속성들에 대해서 대칭적불확실성 값 SU을 계산한 후 이 값을 이용하여 문제에서 고려하는 클래스 개념과 관련된 유의미한 속성을 사용자가 미리 정한 임계값 δ를 기준으로 선택하여 SU 값이 큰 순서대로 속성목록집합인 S_{list}에 1차로 유지한다. 두 번째 단계는 중복성 제거 단계로 앞 단계에서 생성한 속성집합 S_{list}의 각 속성들과 클래스와의 상관관계를 고려하여 중복된 속성들을 S_{list}에서 제거하여 앞서 1차적으로 선택되었던 속성들 중에서 중복성 없는 주요 속성들만을 선택하여 새로운 주요 속성부분집합을 생성한다. 이렇게 선택된 최종 속성들은 속성부분집합 $S'_{list}(=S_{best})$에 유지된다.

3. 연관 속성 마이닝

3.1 연관성 추출

속성들의 연관을 통한 단백질 상호작용을 예측하기 위해서 본 논문에서는 Agrawal 등에 의해 제안된 연관규칙 발견 알고리즘인 Apriori 알고리즘을 활용하였다[16]. 연관규칙은 본래 대용량의 데이타를 담고 있는 데이타베이스에서 각 데이타(트랜잭션; transaction; 논문에서는 각각의 단백질 상호작용을 의미)들의 연관성을 추출하기 위하여 고안되었다. 데이타를 구성하는 항목들의 집합 $J = \{i_1, i_2, ..., i_m\}$, 고려하는 문제에 대한 데이타 집합을 D, 고유의 번호(TID; transaction id)를 가지는 각각의 트랜잭션들을 T, 각각의 데이타들이 항목들 집합의 특정 요소들로 구성되어 $T \subseteq J$의 관계를 만족한다고 하면 항목집합 A와 $B(A, B \subset J)$에 대한 연관규칙은 $A \Rightarrow B$와 같은 내포(implicit) 형태가 된다. 이때 $A \cap B = \phi$가 성립하는 규칙만을 고려하여 규칙을 구성하는 중복 항목들에 의한 비효율성을 제거한다.

일반적으로 연관규칙 R은 규칙의 조건부 집합 A와 조건이 충족 되었을 때의 결과부분을 나타내는 집합 B로 이루어지며 $R(A \Rightarrow B)$과 같은 형태로 표현할 수 있다. 이와 같은 연관규칙은 연관규칙의 특징을 표현하는 지지도(SP; support ratio)와 신뢰도(CF; confidence ratio)의 두 가지 값을 갖는다. 응용에 따라서 연관규칙

표 3 정보이득 기반 속성차원축소필터(FDRF)

입력: (S, C, δ)

S : N개의 속성들과 해당 속성집합이 속하는 클래스 변수로 이루어진 훈련 예제 $S=(f_1,...,f_N,C)$.

C : 각각의 예제들이 속하는 클래스들의 집합으로 연관규칙의 조건 부분과 결과 부분에 대한 클래스를 모두 포함하는 집합 $C=C_C\cup C_R$.

δ : 사전 정의된 SU값의 임계값

출력: (S_{best})

S_{best}: N개 속성들 중에서 선택된 최적속성집합 F'.
($F=\{f_1,...,f_N\}$, $F'=\{f_1,...,f_n\}$, $n\le N$)

속성선택 알고리즘

1. 초기화 단계:
   1-1. $S'_{list}=\phi$, $f_p=\phi$, $f_q=\phi$, $f'_q=\phi$.
2. 임시 후보속성 선택 단계
   2-1. 식 (5)를 이용하여 $i=1,...,N$에 대하여 각 f_i의 $SU_{i,c}$ 값 계산
   2-2. 각 $SU_{i,c}$값들 중에서 $SU_{i,c}\ge\delta$인 경우 $S'_{list}\leftarrow S'_{list}\cup f_i$ ($i=1,...,N$)
   2-3. $sort_{descending}(S'_{list})$
   2-4. $f_p\leftarrow first\_element(S'_{list})$
3. 최적 후보속성 선택 단계: (중복속성 삭제)
   3-1. $f_p\neq NULL$인 경우 아래의 (3-2)~(3-4)를 반복
   3-2. $f_q\leftarrow following\_element(S'_{list},f_p)$ (f_q: S'_{list}의 f_p다음 속성)
   3-3. $f_q\neq NULL$인 경우 아래의 (3-3-1)~(3-3-3)을 반복
      3-3-1. $f_q'\leftarrow f_q$
      3-3-2. (a) $SU_{p,q}\ge SU_{q,c}$인 경우
         $S'_{list}\leftarrow S'_{list}-f_q$,
         $f_q\leftarrow following\_element(S'_{list},f'_q)$
      3-3-3. (b) $SU_{p,q}<SU_{q,c}$인 경우
         $f_q\leftarrow following\_element(S'_{list},f_q)$
   3-4. $f_p\leftarrow following\_element(S'_{list},f_p)$
4. 종료 단계:
   4-1. $S_{best}\leftarrow S'_{list}$
   4-2. 전체 속성의 부분집합인 최적속성집합 S_{best}를 반환하고 알고리즘 종료.

에 대해 보다 다양한 값을 고려함으로써 보다 제약되고 정확한 규칙을 고려하기도 하지만 일반적으로 지지도와 신뢰도 두 값으로도 충분히 좋은 결과를 얻을 수 있기 때문에 본 논문에서 다루는 문제인 단백질 상호작용 예측을 위한 응용에서는 지지도와 신뢰도 두 가지 값만을 고려하도록 한다.

연관규칙의 지지도는 모든 속성들이 규칙에 함께 나타나는 빈도를 의미하며, 신뢰도는 규칙의 정확성을 나타내는 것으로 지지도 값을 규칙의 조건부에 속성들이 나타나는 빈도로 나누어 계산한다. 연관규칙의 지지도와 신뢰도는 0.0~1.0 사이의 값으로 표현 할 수 있으나 일반적으로 0%~100%의 값으로 표현하기도 하며 본 논문에서는 후자의 %표기법을 사용하였다. 지지도와 신뢰도는 확률적으로 표현으로 다음과 같이 표현할 수 있다.

$$SP(A\Rightarrow B)=P(A\cup B), \quad (6)$$

$$CF(A\Rightarrow B)=P(B|A). \quad (7)$$

여기서 '$A\Rightarrow B$'는 두 항목 A, B의 연관규칙을 나타내며 규칙의 항목 A와 B는 단일항목(단일속성)이거나 항

목집합(속성들의 집합)일 수 있으며 단일항목 보다는 집합을 표현하는 경우가 일반적이다. 연관규칙은 사용자가 미리 정한 임계값보다 큰 값을 가지는 규칙들을 모두 고려함으로써 발견 가능하며 사용자가 정하는 임계값으로는 지지도의 최소값($SF_{\min}$)과 신뢰도의 최소값($CF_{\min}$) 두 가지가 있다. 두 최소지지도 값과 최소신뢰도 값을 모두 만족하는 규칙을 '강한규칙(strong rule)'이라 하며 단백질 상호작용 예측을 위한 본 응용에서는 이러한 강한규칙들만을 고려하기로 한다.

3.2 속성연관을 이용한 단백질-단백질 상호작용의 표현

본 논문에서 다루는 각각의 단백질-단백질 상호작용들은 서로 상호작용하는 두 단백질의 쌍으로 표현된다. 단백질들 간의 상호작용을 각 단백질 속성들의 관련성에 기초하여 예측하기 위해서 본 논문에서는 각각의 단백질-단백질 상호작용을 3.1절에서 설명한 데이타베이스의 트랜잭션(T)으로 고려하고 단백질 속성들의 집합을 항목들의 집합 J로 고려하였다. 각각의 단백질 상호작용 트랜잭션들은 항목집합들이 해당 트랜잭션에 존재하는지를 0과 1로 나타내는 방식의 이진벡터(binary vector) 형태로 표현된다. 그림 1은 단백질 상호작용 트랜잭션들에 대한 이진벡터 형식의 표현을 나타낸다. 예를 들어, 단백질의 속성들이 5개 존재한다고 할 때 상호작용하는 두 단백질 A, B가 각각 1, 3, 5번째 속성과 2, 4, 5번째 속성들을 갖는 경우 두 단백질들 간의 상호작용은 $A(10101) \rightarrow B(01011)$과 같이 표현된다. 이와 같은 연관규칙 발견을 통하여 본 논문에서는 특정 상호작용들을 일반화 하는 속성들 간의 연관규칙을 찾아낸다.

각각의 단백질-단백질 상호작용들은 속성값들의 이진벡터로 해당 속성이 상호작용에 존재하는지와 그들 간의 묵시적 연관성을 나타내는 방식으로 표현된다. 사용자가 미리 지정한 임계값보다 적은 정보성(SU 값)을 가지는 속성들은 표 3의 FDRF에 의해 제거된다. FDRF에 의해 제거된 속성들은 'don't care'로 설정되어 연관규칙 발견과정에서 배제되며 정해진 값보다 SU 값이 큰 속성들만이 연관규칙 발견에서 고려된다(지지도와 신뢰도 계산에서 제외된다). 그림 1의 FDRF 우측에는 이러한 'don't care' 속성들이 표현되어있지 않다.

4. 실험 및 결과

4.1 실험 설정

4.1.1 실험 데이타

실험에서는 주요 단백질-단백질 상호작용 데이타로 Oyama[20] 등이 논문에서 사용한 상호작용 데이타를 사용하였다. Oyama가 논문에서 사용한 데이타는 MIPS와 YPD, 그리고 Ito[11]와 Uetz [13]의 Y2H 실험 결과를 포함한다. 본 논문에서는 상호작용하는 각 단백질들의 속성을 보다 풍부하게 인코딩하기 위하여 추가적으로 SGD도 함께 사용하였다. 표 4는 각각의 데이타들이 가지는 단백질 상호작용 개수를 나타낸다. 표 4의 오른쪽은 이 같은 전체 상호작용들을 이진벡터로 표현하기 위해 사용한 속성들의 개수와 표 3의 FDRF로 정보성이 적은 속성들을 제거한 후의 속성의 개수를 나타낸다.

그림 2는 훈련데이타로 주어진 단백질 상호작용 데이

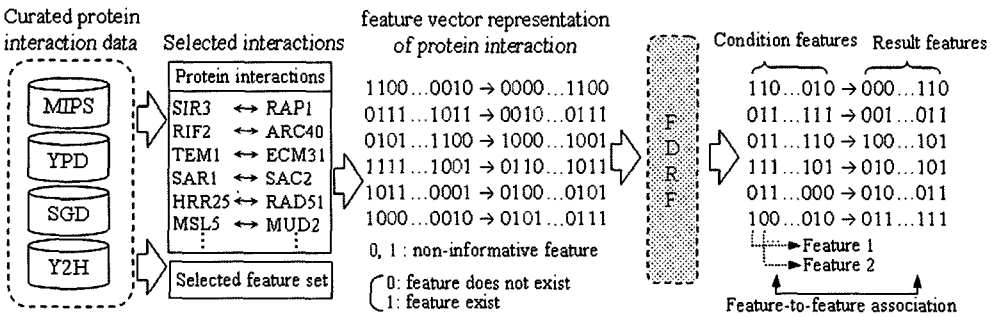

그림 1 속성 벡터들을 이용한 단백질 상호작용의 표현과 속성필터링

표 4 실험에 사용한 상호작용 데이타들의 수와 속성의 개수

상호작용 데이타	상호작용 개수	초기 속성의 개수	필터링 후 속성의 개수
MISP	10,641	6,232	1,293
YPD [25]	2,952		
SGD	1,482		
Y2H (Ito *et al.* [11])	957		
Y2H (Uetz *et al.* [13])	5,086		

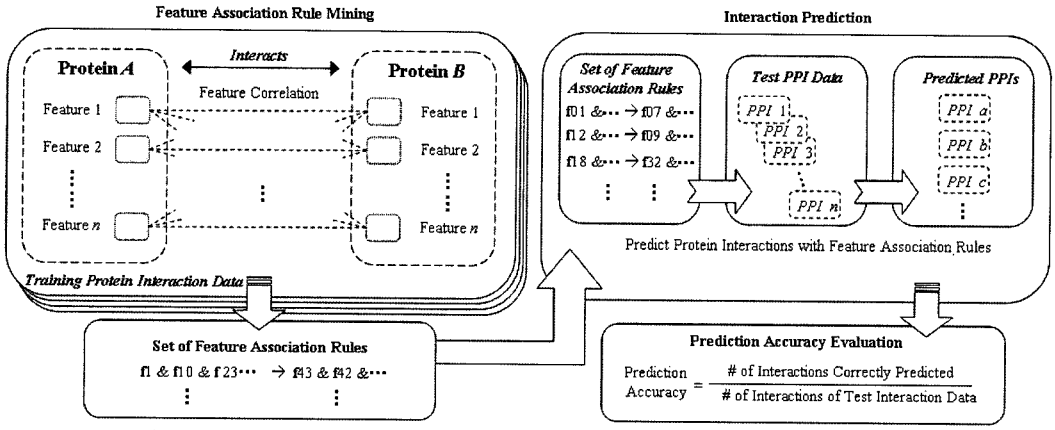

그림 2 PPI 데이타에서 연관규칙의 학습 및 연관규칙을 이용한 단백질 상호작용 예측

표 5 속성연관을 통한 단백질 상호작용 예측에 사용된 단백질 속성들과 개수

속성	설명	사용된 속성의 개수
YPD 범주	- 세포역할이나 생화학적 기능 등을 이용한 분류	275
EC 번호	- SWISS-PROT, PIR의 기능에 따라 4단 계층 범주로 나뉘어 할당된 번호(번호가 '1.2.3.4'과 같을 경우, 1, 1.2, 1.2.3, 1.2.3.4와 같이 모든 단계를 풀어서 각 단계를 표현할 수 있도록 하여 사용)	616
키워드 집합	- 단백질의 기능적 성질·분류 등을 의미하는 단백질마다 할당되어져 있는 키워드 집합(SWISS-PROT, PIR의 키워드 사용)	491
PROSITE 모티프	- PROSITE 데이타베이스에서 제공하는 각 단백질별 DNA서열 정보	940
밀집 아미노산 집합	- 단백질의 아미노산 서열들 중에서 특정 아미노산들이 밀집되는 특정을 정해진 크기의 서열 길이에 존재하는 각 아미노산들의 집중도를 0~100%로 5% 단위로 표기한 아미노산 밀집특성	20
아미노산 속성	- 알라닌(A; Ala; Alanine)부터 티로신(Y; Tyr; Tyrosine)까지 전체 20개의 아미노산들에 대해서 밀집된 아미노산들의 곁사슬성질(친수성/소수성/산성/염기성) 및 기타 특성들(positive/negative, polarity, charged, small/tiny, aromatic, aliphatic, van der Waals volume, isoelectric point 등)	152
상동서열집단 (homologue sequence cluster)	- 모든 효모 단백질들에 대해서 BLASTP를 이용한 아미노산 서열정보에 대해 상동서열집단을 구하고 동일 아미노산이 집중적으로 군집화된 클러스터를 제외한 클러스터 패턴의 단백질별 존재여부	3,724
도메인 연관 서열패턴	- 특정 도메인에 바인딩 한다고 알려져 있는 아미노산 메타 패턴집합의 존재여부 및 그 수 (예: SH3 도메인과 아미노산 패턴 RxxPxxP, PxxPxR)	14

타에서 속성들 간의 연관규칙을 추출한 후 이를 이용하여 테스트집합의 단백질 상호작용을 예측하는 과정 전반을 나타낸다. 2.1의 '속성 이산화' 과정은 그림 2의 연관규칙 발견과정 이전에 연속값을 가지는 속성들에 대하여 적용되며 2.2의 '속성차원의 축소' 과정은 2.1의 방법을 이용하여 속성값의 이산화 후 이산화된 속성들에 적용되어 정보성이 없는 속성들을 걸러내는 역할을 한다.

속성집합을 이용한 단백질 상호작용의 확장은 표 5와 같은 단백질들의 속성을 이용하여 이루어졌다. 표 5는 다양한 데이타베이스에서 수집한 속성들로 단백질 상호작용 데이타를 분석하기 위하여 각각의 단백질을 그림 2와 같이 '단백질-단백질' 상호작용에서 '속성집합-속성집합'의 상호작용으로 확장하기 위하여 사용되었다. 속성들은 각 단백질들의 YPD 범주, EC 번호, SWISS-PROT(SWISS-PROT database)[4]과 PIR(Protein Information Resource)[5]에서 각 단백질에 할당된 키워드집합, PROSITE(Dictionary of protein sites and patterns)[6]의 모티프 등의 정보들이 사용되었다.

4.1.2 속성 필터 임계값(δ)의 설정

표 3의 정보이득 기반 속성차원축소필터 알고리즘에서 속성값의 필터링을 위한 SU값의 임계값은 UCI Machine Learning Repository[7]와 UCI KDD Archive[8]에서 수집한 7개의 데이타 집합(Lung-cancer, Pro-

4) http://www.psc.edu/general/software/packages/swiss/swiss.html
5) http://pir.georgetown.edu/
6) http://www.psc.edu/general/software/packages/prosite/prosite.html
7) http://www.ics.uci.edu/~mlearn/MLRepository.html
8) http://kdd.ics.uci.edu/

표 6 속성 필터링 임계값(δ) 설정을 위한 데이타 집합들의 통계정보

데이타 집합	사용한 예제 개수	전체 속성의 개수	예제 분류 클래스 개수
LUNG-CANCER	32	57	3
PROMOTERS	106	59	2
SPLICE	3,190	62	3
US-CENSUS	9,338	68	3
CoIL2000	5,822	86	2
CHEMICAL	936	151	3
MUSK2	6,598	169	2

moters, Splice, US-Census90, CoIL2000, Chemical, Musk2)을 사용한 반복 실험에서 각 실험별로 최적의 결과는 보이는 δ값을 선정하여 평균값을 계산하여 δ = 0.73으로 정하였다. 표 6은 δ설정을 위해 사용된 데이타 집합들의 간단한 통계 정보를 나타나며 '예제 분류 클래스 개수'는 데이타들이 분류될 실제 클래스의 수를 나타낸다.

4.2 실험 결과 및 분석

실험에서는 정보성이 높은 속성들을 표 3의 필터링 과정을 이용하여 선택하였다. 속성 필터링을 통해 전체 6,232개 속성에서 선택된 1,293개의 속성으로 구성된 단백질 상호작용 데이타를 이용하여 연관규칙 발견과정을 수행하였다. 연관규칙 발견을 위한 지지도와 최소신뢰도 값은 실험 성능의 비교를 위하여 Oyama 등이 실험에서 사용한 설정과 동일하게 각각 9와 75%로 설정하였다. 다음으로 앞의 과정을 통하여 얻은 속성간의 연관규칙을 이용하여 속성연관 학습에 사용되지 않은 검증 데이타집합의 단백질 상호작용에 대하여 예측 성능을 측정하였다. 표 7의 예측 정확도는 각각 4,628개의 훈련 집합을 이용하여 연관규칙을 찾고 훈련 집합의 약 10%정도인 테스트집합의 상호작용을 얼마나 예측 하는지 예측된 상호작용의 비율을 10-집단 교차검증(10-fold cross validation)을 이용하여 측정하였다.

표 7은 연속속성 이산화와 비정보성 속성필터링(FDRF)을 이용한 단백질 상호작용 예측 정확률과 소요시간 비교 결과를 나타낸다. 표에서 예측정확도의 최대 성능향상폭(§)과 연관규칙 발견 소요시간의 최대 성능향상폭(♯)은 '연속값을 가지는 속성들에 대한 균등분할 이산화 적용 후 속성들의 연관규칙 추론을 수행하여 발견된 연관규칙으로 검증집합의 단백질 상호작용을 예측하는 방법(ⓐ)'과 '최대상호의존성기반 이산화방법으로 연속값을 가지는 속성들을 이산화한 후 정보이론 기반의 속성필터로 정보성이 적은 속성들을 제거한 속성집합에 대하여 연관규칙 발견 및 발견된 연관규칙을 기반으로 검증집합의 단백질 상호작용을 예측하는 방법(ⓓ)'의 성능 차이를 구한 것이다. 이 외의 방법 ⓑ와 ⓒ는 각각 '연속값을 가지는 속성들에 대한 균등분할 이산화 적용 후 정보이론 기반의 속성필터로 정보성이 적은 속성들을 걸러낸 후 연관규칙 발견 및 발견된 연관규칙을 기반으로 검증집합의 단백질 상호작용을 예측'하는 방법과 '최대상호의존성기반 이산화방법으로 연속값을 가지는 속성들을 이산화한 후 연관규칙 추론을 수행하여 발견된 연관규칙으로 검증집합의 단백질 상호작용을 예측'하는 조합을 의미한다.

표 7은 정보성이 없는 속성을 제거함으로써 연관규칙 발견의 속도 및 정확도의 측면에서 성능향상을 얻을 수 있음을 나타낸다. 연속값을 갖는 속성들을 균등하게 분할하여 이산화한 후 속성필터(FDRF)를 적용함으로써

표 7 연속속성 이산화와 비정보성 속성필터링(FDRF)을 이용한 단백질 상호작용 예측 정확률과 연관규칙 발견 소요시간 비교(※ 소요시간은 속성집합을 이용하여 연관규칙을 생성하는데 소요되는 시간이며 펜티엄4 3.2GHz, 2GB RAM 윈도우 시스템에서 측정함. 예측방법의 약자들은 각각 n.Dt.: 균등분할 연속속성 이산화방법, i.Dt.: 최대상호의존성기반 연속속성 이산화방법, FDRF: 속성필터기반 비정보성 속성필터링 방법, Asc.: 연관규칙 발견방법을 의미함)

예측방법	상호작용의 개수			예측 정확도 (|P|/|T|)	소요시간 (연관규칙 발견)
	훈련 집합	테스트집합 (T)	예측된 상호작용 (P)		
ⓐ: n.Dt. + Asc.	4,628	463	423	91.4 %	212.34 sec
ⓑ: n.Dt. + FDRF + Asc.	4,628	463	439	94.8 %	163.35 sec
ⓒ: i.Dt. + Asc.	4,628	463	431	93.0 %	197.98 sec
ⓓ: i.Dt. + FDRF + Asc.	4,628	463	447	**96.5 %**	**149.91 sec**
성능향상	-	-	-	5.1 % (§)	29.4 % (♯)

전체 속성들에서 선택된 속성들 간의 연관규칙 발견을 통한 단백질 상호작용 예측의 정확도는 약 3.4% 향상되었으며, 연관규칙 발견의 속도 또한 FDRF를 적용하지 않은 데이타집합을 이용할 때보다 최대 약 23.1% 정도 향상되었다. 연속값 속성의 이산화를 최대상호의존성 기반으로 보다 정교하게 한 경우 필터를 적용하지 않는 방법에서 약 1.6%의 향상이 있었으며 필터를 적용한 경우에도 약 1.7%의 성능 향상을 보였다. 또한, 방법 ⓐ와 ⓓ의 최대 성능차이는 약 5.1%로 상대적으로 보다 정교한 속성값의 이산화 및 속성에 대한 필터링 방법을 통해 적지 않은 예측 성능의 향상을 얻을 수 있음을 나타낸다.

속성값 이산화와 속성필터링을 통한 성능향상은 연관규칙 발견에 소요되는 시간에 대한 고찰에서도 관찰되었다. 기본적으로 속성필터를 적용한 경우(방법 ⓑ와 ⓓ)가 그렇지 않은 경우(방법 ⓐ와 ⓒ)보다 시간이 적게 걸려, 사용된 속성들 중에는 정보성이 없는 속성들이 여럿 존재했음을 확인할 수 있었다. 다음으로, 속성필터를 적용한 두 방법(ⓑ와 ⓓ)에서는 이산화 방법에 따른 소요시간의 차이를 발견할 수 있었다. 연속값을 가지는 속성값에 대한 최대상호의존성 기반 이산화방법을 적용한 경우(ⓓ) 균등분할 이산화를 적용한 예측방법(ⓑ)보다 약 8%의 속도 향상을 보여, 속성 이산화의 정교한 처리가 결과적으로 생성되는 속성의 수를 최적화 하여 전체적으로 속성차원을 낮추는 효과와 이로 인한 연관규칙 발견속도의 향상을 가져왔음을 확인할 수 있었다.

표 8은 단백질 상호작용에 사용된 서로 다른 방법들 간의 예측 정확도를 비교 결과를 나타낸다. 표 8의 방법 ①은 단백질 상호작용의 간접 및 직접 상호작용 정보를 이용하여 구성한 유사도 행렬(random forest similarity matrix)을 이용한 상호작용 단백질 분류 방법을 타나낸다[26]. 방법 ②는 단백질 상호작용에서 중요한 역할을 단백질의 구조 및 서열 보존정보를 이용하여 새로운 단백질을 예측하는 방법을 나타내며[27], 방법 ③은 도메인 정보와 염기서열 복합(amino acid composition)정보, 그리고 세포부분위치추정(subcellular localization) 정보를 복합적으로 사용하여 커널 기반 분류 방법인 SVM (Support Vector Machine) 기반으로 단백질 상호작용을 예측하는 방법을 나타낸다[28]. 다음으로 방법 ④는 확률모델 기반의 이종데이타 통합 방법에 MCMC(Markov Chain Monte Carlo) 파라미터 추정을 적용하여 단백질 상호작용을 예측하는 방법을 나타내고[29], 방법 ⑤와 ⑥은 각각 표 7의 ⓐ와 ⓓ를 나타낸다. 표 8의 예측 정확도 ⓐ는 해당 방법을 제시한 논문에서 보고된 예측 정확도를 나타내며, 예측 정확도 ⓑ는 표 7의 실험을 위해 사용한 상호작용 데이타를 공통으로 이용하여 측정한 결과를 나타낸다. 예측 정확도 ⓑ를 측정하기 위하여 사용한 데이타는 오류상호작용(false positive interaction)이 상대적으로 적은 집합을 사용하여 각 방법들이 논문에서 보고한 성능보다 높은 결과가 측정되었다.

표 8에서 상호작용들 자체의 간접 및 직접 정보만을 이용하여 상호작용 단백질의 분류를 통한 예측(①)은 가장 낮은 예측 정확도를 보였는데 이러한 결과는 상호작용 예측에 사용한 정보가 기존에 알려진 상호작용들을 이용하여 구성한 정보밖에 없기 때문이라 생각된다. 이는 상호작용 예측에 구조정보 및 서열정보와 함께 단순 상화작용 정보 이외에 보다 다양한 속성들을 함께 고려한 방법들(방법 ②, ③, ⑤, ⑥)이 상대적으로 높은 예측 성능을 갖는 것으로 추론 할 수 있다. 그러나 상호작용 정보들만을 이용하여 새로운 상호작용을 예측하는 방법을 사용하더라도 기존 상호작용이 가지는 상호 의존성이나 기타 확률적 속성들을 보다 자세히 추정함으로써 좀 더 정확한 예측이 가능하였다(방법 ④). 다양한 상호작용 데이타를 이용한 방법 ④는 다수의 속성과 함께 연속값을 가지는 속성의 이산화 및 속성 필터링을 함께 사용하여 본 논문에서 예측한 결과(⑥)와 비교하여 약 2% 미만의 근소한 차이의 성능을 보이고 있는 것을 확인 할 수 있었다. 방법 ④의 결과가 본 논문의 결과와 큰 차이를 보이지는 않지만 논문에서 사용한 방법은 예측을 위한 규칙을 확인 할 수도 있기 때문에 세부적 확률적 인자 추정을 통한 상화작용 예측보다 예측 결과를 해석하는데 유용하게 사용될 수 있다. 다만, 방법 ④의 접근방법과 속성 연관규칙을 이용한 예측의 장점을 고려한 새로운 예측 모델을 구성한다면 보다 낮은 비용으로 상대적으로 높은 예측 성능을 얻을 수 있을 것으로 예상된다.

표 8 다른 상호작용 예측 접근방법들과의 예측성능 비교

상호작용 예측 방법	예측 정확도 (ⓐ)	예측 정확도 (ⓑ)
①: k−NN classification (with random forest similarity)	70.45 %	86.34 %
②: Structure & sequence conservation based prediction	73.10 %	88.46 %
③: SVM	79.00 %	**93.28 %**
④: Generative stochastic model + MCMC estimation	86.90 %	**94.82 %**
⑤: Asc. + n.Dt.	-	**91.40 %**
⑥: Asc. + FDRF + i.Dt.	-	**96.50 %**

표 9 미확인 단백질과 연관된 새로운 단백질 상호작용

단백질	상호작용 예상 단백질	설명
YK61(*)	KC21	· 카세인키나아제(Casein kinase) 연관 단백질
	YMT9	· 리보솜리보핵산(rRNA) 처리 단백질
	NOP2	· 핵소체(Nucleolar) 단백질
YNJ2(*)	YKA2	· 엔도소말(Endosomal) 단백질
	RHO1	· 로(Rho) 단백질
	SC17	· 소포융합(Vesicular fusion) 단백질

연관속성발견을 통해 새로운 상호작용들을 발견할 수 있었는데, 표 9는 속성연관규칙 발굴을 통하여 추가적으로 예측한 단백질 상호작용들 중의 일부를 나타낸다. 표 좌측의 YK61와 YNJ2는 그 기능이 아직 알려지지 않은 단백질로 실험에서는 이 두 단백질과 연관된 상호작용 데이타는 사용되지 않았다. 표 9의 '상호작용 예상 단백질'들은 기능이 알려지지 않은 이 두 각각의 단백질들이 상호작용을 할 것으로 예측되는 대상 단백질들을 나타낸다. 즉, 단백질들의 속성연관에 기반을 두어 볼 때 이 두 단백질들이 다른 표에 제시된 단백질들과 상호작용할 수 있음을 의미하며, 이 단백질들의 유의미한 실제 상호작용 여부는 생물학적 추가 실험을 통하여 확인할 수 있을 것이다.

5. 결 론

논문에서는 최대의존성기반 연속속성 이산화 방법과 정보이론기반 속성차원 감소 필터의 사용 및 속성연관규칙의 발견을 통해 단백질 상호작용을 예측하는 방법을 제시하였다. 제안된 방법은 예측 정확도와 연관규칙 생성 시간에 있어서 연속값을 가지는 속성의 이산화와 속성필터링 세부적으로 사용하지 않는 방법에 비해 향상된 성능을 보였다. 그림 1과 표 7은 단백질-단백질 상호작용을 보다 작은 단위로 즉, 단순히 단백질 개체와 단백질 개체간의 상호작용이 아닌 속성집합들과 속성집합들 간의 상호작용으로 살펴볼 수 있으며 이처럼 보다 세부적인 상호작용 고찰을 통하여 단백질 상호작용 예측의 성능 향상이 가능함을 보여주고 있다. 즉, 논문에서 제안한 이산화 및 정보성 기반 속성선택과 연관규칙 발견을 이용한 단백질 상호작용 예측 방법은 다수의 속성집합으로 구성 가능한 단백질 상호작용 데이타에서 연속값 속성을 효율적으로 이산화하고, 정보성이 없는 속성들을 제거함으로써 속성들 간의 연관규칙 발견과정에서 잘못된 속성연관규칙의 발견 가능성을 줄여 결과적으로 연관규칙 발견의 계산속도 향상 및 예측 성능의 향상을 보인다고 할 수 있다. 또한 논문에서 사용한 방법 및 전체 과정은 실험자가 직접 고찰 가능한 속성들 간의 연관규칙을 찾을 수 있어 아직 발견되지 않은 잠재적인 단백질 상호작용도 예측할 수 있는 가능성을 확인하였다. 즉, 단백질 상호작용에 대하여 생물학적 실험을 통한 *in vitro*나 *in vivo* 분석 이전에 데이타 기반의 *in silico* 분석을 수행하여 가능한 상호작용 후보를 먼저 생성해보기 위한 가능한 한 가지 방법으로 활용될 수 있을 것이다.

결과적으로 본 논문에서는 제안한 방법이 실제 생물학 실험을 통하여 생성된, 다수의 속성을 가지는 단백질들 중에서 상호작용을 하는 단백질들의 효율적인 예측에 활용 될 수 있는 가능성을 보였다. 그렇지만, 이러한 접근 방법은 기존의 MIPS와 SGD등의 데이타를 실험의 중요 기준(golden standard)으로 삼고 있기 때문에 실제 생물학 실험을 통해 생성된 이 같은 데이타가 적당한 수준의 오측률(false positive ratio)을 가지고 있는 경우에 적당하다. Y2H와 같은 고효율방법들로 생성된 현재의 상호작용 데이타들은 어느 정도의 오류상호작용들을 가지고 있다고 생각되고 있으며 다양한 신규 실험방법을 통한 재검사를 통해서 이러한 오류상호작용들이 지속적으로 수정되고 있다. 때문에 오류상호작용에 보다 덜 민감한 상호작용 예측 방법의 개발이 필요하며, 속성 인코딩에 따른 결과의 차이 극복과 속성값을 이산화하여 사용하는 방법에 따른 성능상의 변화에 대한 고찰 및 다수의 속성에서 보다 효율적으로 단백질의 분류 및 예측에 적합한 속성을 선택하기 위한 방법에 대한 추가적인 연구가 필요하다. 마지막으로 연관규칙의 한계를 극복할 수 있는 확률그래프모델과 같은 확률기반의 보다 정교한 연관성 예측모델에 대한 연구가 필요할 것이다.

참 고 문 헌

[1] Deng, M., Mehta, S., Sun, F., and Chen, T., "Inferring domain-domain interactions from protein-protein interactions," *Genome Res.* Vol.12, No.10, pp. 1540-1548, 2002.

[2] Goffeau, A. and Barrell, B. G. *et al.*, "Life with 6000 genes," *Science*, Vol.274, pp. 563-567, 1996.

[3] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D., "Cluster analysis and display of genomewide expression patterns," Proc. Natl. Acad. Sci., Vol.95, pp. 14863-14868, 1998.

[4] Pavlidis, P. and Weston, J., "Gene functional classification from heterogeneous data," In Proc. 5th Int. Conf. Comput. Mol. Biol. (RECOMB-2001), pp. 249-55, 2001.

[5] Wu, L. F. and Hughes, T. R. et al., "Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters," Nature Genetics, Vol.31, pp. 255-265, 2002.
[6] Park, J., Lappe, M., and Teichmann, S. A., "Mapping protein family interactions: intra-molecular and intermolecular protein family interaction repertoires in the PDB and yeast," J. Mol. Biol. Vol.307, pp. 929-39, 2001.
[7] Iossifov, I. and Krauthammer, M. et al., "Probabilistic inference of molecular networks from noisy data sources," Bioinformatics, Vol.20, No.8, pp. 1205-12013, 2004.
[8] Ng, S. K., Zhang, Z., and Tan, S. H., "Integrative approach for computationally inferring protein domain interactions," Bioinformatics, Vol.19, No.8, pp. 923-29, 2003.
[9] Fields, S. and Sternglanz, R., "The two-hybrid system: an assay for protein-protein interactions," Trends in Genetics, Vol.10, pp. 286-92, 1994.
[10] Ito, T. and Chiba, T. et al., "A comprehensive two-hybrid analysis to explore the yeast protein interactome," Proc. Natl Acad. Sci., Vol.98, pp. 4569-4574, 2001.
[11] Ito, T., Matsui, Y., Ago, T., Ota, K., and Sumimoto, H., "Novel modular domain PB1 recognizes PC motif to mediate functional protein-protein interactions," EMBO J., Vol.20, pp. 3938-3946, 2001.
[12] Uetz, P. and Giot, L. et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," Nature, Vol.403, No.6770, pp. 623-627, 2000.
[13] Bu, D. and Zhao, Y. et al., "Topological structure analysis of the protein-protein interaction network in budding yeast," Nucl. Acids. Res., Vol.31, No.9, pp. 2443-2450, 2003.
[14] Tong A. H. and Lesage G. et al., "Global mapping of the yeast genetic interaction network," Science, Vol.303, No.5659, pp. 808-813, 2004.
[15] Hartwell L., "Robust Interactions," Science, Vol.303, No.5659, pp. 774-775, 2004.
[16] Agrawal, R., Imielinski, T., and Swami, A., "Mining association rules between sets of items in large data-bases," In Proc. ACM SIGMOD-93, pp. 207-216, 1993.
[17] Satou, K. and Shibayama, G. et al., "Finding association rules on heterogeneous genome data," In Proc. Pac. Symp. Biocomput., pp. 397-408, 1997.
[18] Creighton, C. and Hanash, S., "Mining gene expression databases for association rules," Bioinformatics, Vol.19, No.1, pp. 79-86, 2003.
[19] Fellenberg, M., Albermann, K., Zollner, A., Mewes, H. W., and Hani, J., "Integrative analysis of protein interaction data," In Proc. Int. Conf. Intell. Syst. Mol. Biol., Vol.8, pp. 152-161, 2000.
[20] Oyama, T., Kitano, K., Satou, K., and Ito, T., "Extraction of knowledge on protein-protein interaction by association rule discovery," Bioinformatics, Vol.18, No.5, pp. 705-714, 2002.
[21] Yu, L. and Liu, H., "Feature selection for high dimensional data: a fast correlation-based filter solution," In Proceedings of the 20th International Conference on Machine Leaning (ICML-03), pp. 856-863, 2003.
[22] Kurgan, L. A. and Cios, K. J., "CAIM Discretization Algorithm," IEEE Trans. Knowledge and Data Eng., Vol.16, No.2, pp. 145-153, 2004.
[23] Quinlan, J. R., C4.5: Programs for machine learning, Morgan Kaufmann Publishers, San Francisco, 1993.
[24] Press, W. H. and Flannery, B. P. et al., "Numerical recipes in C: The Art of Scientific Computing," 2nd Ed., pp. 633-634, Cambridge University Press, Cambridge, 1992.
[25] Csank C. and Costanzo M. C. et al., "Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD)," Methods Enzymol., Vol.350, pp. 347-373, 2002.
[26] Qi, Y., Klein-Seetharaman, J., and Bar-Joseph, Z., "Random forest similarity for protein-protein interaction prediction from multiple sources," In Proc. Pac. Symp. Biocomput., pp. 531-542, 2005.
[27] Aytuna, A. S., Gursoy, A., and Keskin, O., "Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces," Bioinformatics, Vol.21, No.12, pp. 2850-2855, 2005.
[28] Dohkan, S., Koike, A., and Takagi, T., "Prediction of protein-protein interactions using support vector machines," In Proc. 4th IEEE Symp. Bioinfo. Bioeng. (BIBE'04), pp. 576-586, 2004.
[29] Chen, S.-C. and Bahar, I., "Mining frequent patterns in protein structures: a study of protease families," Bioinformatics, Vol.20, Suppl.1, pp. i77-i85, 2004.

엄 재 홍

1999년 2월 강원대학교 컴퓨터공학과 학사. 2001년 2월 서울대학교 전기·컴퓨터공학부 석사. 2001년~현재 서울대학교 전기·컴퓨터공학부 박사과정. 관심분야는 텍스트마이닝, 정보추출, 정보검색, 생물정보학, 기계학습

장 병 탁

1986년 서울대학교 컴퓨터공학 학사 1988년 서울대학교 컴퓨터공학 석사 1992년 독일 Bonn대학교 컴퓨터공학 박사. 1992년~1995년 독일국립정보기술연구소(GMD) 연구원. 1995년~1997년 건국대학교 컴퓨터공학과 조교수. 1997년~현재 서울대학교 컴퓨터공학부 교수, 인지과학, 뇌과학, 생물정보학 협동과정 겸임. 관심분야는 Biointelligence, Probabilistic Models of Learning and Evolution, Molecular/DNA Computation