그래프 이론 기반의 단백질-단백질 상호작용 데이타 분석을 위한 시스템

(An Analysis System for Protein-Protein Interaction Data Based on Graph Theory)

진 희 정 † 윤 지 현 †† 조 환 규 †††
(Hee-Jeong Jin) (Ji-Hyun Yoon) (Hwan-Gue Cho)

요 약 단백질-단백질 상호작용(PPI : Protein-Protein Interaction) 데이타는 생물체가 어떠한 메커니즘으로 생명을 유지하는지에 대한 정보를 담고 있다. 질병 연구나 신약 연구를 위해서 PPI 데이타를 이용한 많은 연구들이 이루어지고 있다. 이러한 PPI 데이타의 크기는 Yeast-two-hybrid, Mass spectrometry과 Correlated mRNA expression과 같은 방법들로 인하여 점차 그 증가량이 커지고 있다. 따라서 단백질-단백질 상호작용 데이타의 방대한 양과 복잡한 구조로 인하여 사람이 직접 분석하는 것은 불가능하다. 다행히도 PPI 데이타는 단백질은 노드로, 상호작용은 에지로 표현함으로써 전산학의 그래프 구조로 추상화될 수 있다. 본 논문에서는 방대한 단백질-단백질 상호작용 데이타를 연구자가 다양한 방법으로 손쉽게 분석할 수 있는 워크벤치(workbench) 시스템인 Proteinca(PROTEin INteraction CAbaret)에 대하여 소개한다. Proteinca는 다양한 데이타베이스의 PPI 데이타를 그래프이론 기반의 분석 기능들을 제공하며, 그래프로 가시화하여 사용자가 직관적으로 이해할 수 있도록 도와준다. 또한, 중력 모델 기반의 간략화 방법을 제공하여 사용자에게 중요 단백질 중심의 가시화를 제공한다.

키워드 : 생물정보학, 단백질-단백질 상호작용, 그래프 이론, 가시화, 그래프 마이닝

Abstract PPI(Protein-Protein Interaction) data has information about the organism has maintained a life with some kind of mechanism. So, it is used in study about cure research back, cause of disease, and new medicine development. This PPI data has been increased by geometric progression because high throughput methods are developed such as Yeast-two-hybrid, Mass spectrometry, and Correlated mRNA expression. So, it is impossible that a person directly manage and analyze PPI data. Fortunately, PPI data is able to abstract the graph which has proteins as nodes, interactions as edges. Consequently, Graph theory plentifully researched from the computer science until now is able to be applied to PPI data successfully. In this paper, we introduce Proteinca(PROTEin INteraction CAbaret) workbench system for easily managing, analyzing and visualizing PPI data. Proteinca assists the user understand PPI data intuitively as visualizing a PPI data in graph and provide various analytical function on graph theory. And Protenica provides a simplified visualization with gravity-rule.

Key words : bioinformatics, protein-protein interaction, graph theory, visualization, graph mining

1. 서 론

최근 유전체 사업의 활성화로 인하여, 인간을 비롯한 많은 종들의 유전자 지도가 밝혀지게 되었다. 이로 인하여 생물체 내부의 생명 현상의 메카니즘을 구현하려는 포스트 게놈(Post Genome) 시대를 맞이하게 되었다. 하나의 유전자는 발현되는 세포의 종류나 상황에 따라 여러 가지 단백질로 발현되기 때문에 그 기능을 정확하게 파악할 수 없지만, 단백질은 생물체 내의 물질대사에 직접 작용하므로, 연구 결과를 신약개발이나 생물의 생명현상을 밝히는데 응용할 수 있다. 따라서 포스트 게놈 시대에는 유전자보다는 단백질이 주요 연구 대상이 되었다. 그 결과, 단백질의 3차원 구조, 단백질의 서열, PPI 데이타 등 단백질에 관한 많은 연구가 이루어지고

† 학생회원 : 부산대학교 컴퓨터공학과, 정보통신연구소
hjjin@pearl.cs.pusan.ac.kr
†† 정 회 원 : 부산대학교 컴퓨터공학과
jhyoon@pearl.cs.pusan.ac.kr
††† 정 회 원 : 부산대학교 정보컴퓨터공학부, 정보통신연구소 교수
hgcho@pusan.ac.kr
논문접수 : 2005년 1월 24일
심사완료 : 2006년 2월 1일

있으며, 이런 연구 결과들은 데이타베이스로 구축되어 인터넷을 통해 다른 연구자들에게 공개되고 있다.

1.1 단백질-단백질 상호작용의 개요

단백질에 관한 많은 데이타 중, 단백질-단백질 상호작용(PPI : Protein-Protein Interaction) 데이타는 서로 상호작용하는 단백질에 관한 데이타이다. 단백질들은 생물체 내에서 다른 단백질들과 상호작용을 통해 여러 가지 기능을 수행하므로, 단백질을 응용하는 분야에서 PPI 데이타는 중요하게 사용된다. PPI 데이타는 *Yeast two-hybrid, Mass spectrometry, Correlated mRNA expression, Genetic interactions*, 그리고 유전체 분석을 통한 *in-silico* 예측 등의 방법으로 밝혀지고 있다[1]. *Yeast two-hybrid*는 테스트하려는 두 단백질을 직접 섞어서 반응 여부를 조사하는 단순한 방법이며, *Mass spectrometry*는 발광 염료로 식별해놓은 단백질을 질량 분석법으로 정제하여 상호작용 여부를 식별하는 방법이다. *Correlated mRNA expression* 방법은 특정 상태의 세포 안에 존재하는 mRNA 수를 세어서, *Genetic interactions*는 치명적인 돌연변이를 조사함으로써 상호작용하는 단백질을 유추한다. 유전체 분석을 통한 *in-silico* 예측 방법은 생물정보학의 한 분야로써 오페론(operon), 계통도 프로파일(phylogenetic profile) 등의 이론을 바탕으로 PPI 데이타를 추측한다. 이런 방법들은 대량의 PPI 데이타를 한 번의 실험으로 밝혀낼 수 있다는 장점이 있지만, false positive와 false negative가 높다는 단점이 있다. 이러한 실험 결과들은 여러 해외 기관들에서 데이타베이스로 구축되어 공개되고 있다. 대표적인 PPI 데이타베이스로는 다음과 같은 것들이 있다.

- DIP(Database of Interacting Proteins)[2] : DIP은 여러 단백질들의 상호작용에 관한 데이타베이스로, 전체 데이타베이스에서 초파리(*Drosophila Melanogaster*), 효모(*Saccharomyces Cerevisiae*), 예쁜 꼬마 선충(*Caenorhabditis Elegans*)의 정보가 대부분을 차지한다.
- MIPS(Munich Information centre for Protein Sequences)[3] : MIPS는 독일의 Max-Planck-Institute Bioinformatics 그룹에서 생성한 단백질 데이타베이스로, 유전체 서열(genome sequence)의 기능 분석 및 분류에 중점을 두고 있다. MIPS에서는 Genome에 대한 총체적인 정보를 제공하기 위해서 PEDANT라는 서버를 운영 중이다.
- YPD(the Yeast Proteome Database)[4] : YPD는 워싱턴 대학교의 P. Uetz 등이 대량의 효모 유전자를 대상으로 하여 *yeast two-hybrid*와 *protein microarray*를 이용한 실험을 통해 얻은 대량의 PPI 데이타를 바탕으로 구축되었다.
- BIND(the Biomolecular Interaction Network Database)[5] : BIND에는 단백질의 상호작용 정보 외에 상호작용들의 경로(pathway) 정보를 함께 제공한다. BIND에서 제공하는 경로 정보는 '어떠한 질병과 관련이 있다.' 또는 'cell cycle에 포함된다.'라는 식의 추가 정보를 제공한다.
- STRING(Search Tool for the Retrieval of Interacting Genes/Proteins)[6] : STRING은 가장 많은 정보를 가지고 있는 PPI 데이타베이스 중의 하나로, 여러 종들의 PPI 데이타를 포함하고 있다. STRING에는 실험 데이타뿐만 아니라 자체적인 알고리즘으로 예측한 데이타도 포함되어 있다. 따라서 STRING는 데이타의 정확도에 따라 세 분류의 데이타 셋을 제공한다.
- SGD(Saccharomyces Genome Database)[7] : SGD는 *Saccharomyces cerevisiae*의 분자 생물학과 유전학에 관련된 과학적 데이타베이스로, 물리적 지도, 유전적 지도, 각 염색체를 연관시킨 유전체 그래픽와 포유류 유전체를 포함한다, 이 데이타베이스는 알려진 유전자 혹은 염기 서열, 특정 염색체 부위 혹은 표본 DNA, 단백질 아미노산 서열로 검색 가능하다.
- GRID(General Repository for Interaction Datasets)[8] : GRID는 유전적, 물리적, 기능적으로 상호작용하는 데이타들의 데이타베이스이다. GRID에는 논문으로 알려진 모든 상호작용 데이타들을 포함하고 있으며, 현재 BIND[5], MIPS[3]를 포함하여 13,830개의 상호작용데이타를 저장하고 있다. 또한, 이들 데이타를 효율적으로 표현해주기 위해서 "Qsprey" 가시화 프로그램을 제공하고 있다.

PPI 데이타는 여러 데이타베이스에서 플랫 파일의 형태로 제공된다. 플랫 파일은 여러 행으로 이루어진 단순한 텍스트 파일이며, 각 행에는 서로 상호작용하는 한 쌍의 단백질과 단백질의 기능, ID 등의 데이타로 이루어져 있다. 그러므로 플랫 파일만으로는 전체적인 단백질 상호작용 데이타를 직관적으로 이해할 수 없다. 뿐만 아니라, 최근에는 상호작용하는 단백질들을 대량으로 조사할 수 있는 high-throughput 실험 방법들이 많이 개발되어 PPI 데이타가 기하급수적으로 증가하고 있다. 이와 같은 이유로 PPI 데이타를 사람이 직접 관리하고 분석하는 것은 불가능하며, PPI 데이타를 분석하는데 도움이 되는 시스템의 개발이 필요하다. PPI 데이타의 크기는 2003년 12월 기준으로, MIPS가 4,336개의 단백질과 10,467개의 상호작용 정보를 가지며, DIP은 4,718개의 단백질과 15,128개의 상호작용 정보를 가진다. 표 1은 DIP과 MIPS의 PPI 데이타의 특징을 Proteinca (PROTEin INteraction CAbaret)를 이용하여 조사한

표 1 DIP과 MIPS의 PPI 데이타 특징. PPI 데이타는 4,500 여개의 단백질과 10,000개가 넘는 상호작용을 가지는 방대한 데이타이다. 싱글톤은 단백질 하나로만 이루어진 컴포넌트, 더블톤은 두 개의 단백질로 이루어진 컴포넌트를 의미하며, 평행한 상호작용은 연결된 두 단백질이 두개이상의 동일한 상호작용들을, 루프는 하나의 단백질이 자기 자신에게 상호작용을 하는 정보를 말한다. 컴포넌트는 분리된 서브 그래프를, 디그리는 노드에 연결된 간선의 수를 의미한다.

속성	DIP	MIPS
단백질의 수	4,718	4,336
싱글톤의 수	3	9
더블톤의 수	39	94
상호작용의 수	15,128	10,467
평행한 상호작용의 수	0	1,234
루프의 수	269	268
컴포넌트의 수	46	115
상호작용이 가장 많은 컴포넌트의 상호작용의 수	282	289
상호작용의 평균	6.41	4.83

결과이다.

PPI 데이타의 중요한 특징은 단백질을 노드로 하고, 단백질 간의 상호작용을 간선으로 하는 그래프 데이타로 추상화할 수 있다는 점이다. 따라서 이미 많은 연구가 이루어진 그래프 이론을 이용하여 PPI 데이타를 분석할 수 있고, 그 결과로 상호작용 데이타의 복잡도나 특정 단백질들 간의 상호작용 패스 등과 같은 의미 있는 정보를 얻을 수 있다.

1.2 관련 연구

PPI 데이타는 실험에 의해 밝혀진 데이타이지만, 과연 실제로 생물학적 의미를 가지는지에 대해 살펴볼 필요가 있다. 이에 관한 연구로 Hui Ge et al.[9]의 논문이 있다. 이 논문에서는 랜덤하게 생성한 그래프와 PPI 데이타를 비교함으로써 PPI 데이타가 유의미함을 증명하였다. 만들어진 랜덤 그래프는 비교할 PPI 데이타와 동일한 단백질을 가지지만 상호작용하는 단백질들을 랜덤으로 정한다. 물론, 상호작용 수(간선의 총 수)는 동일하다. Hui et al.은 랜덤 그래프가 가질 수 없는 두 가지 특징을 PPI 데이타에서 밝혀냈다. 첫 번째는 동일한 기능 분류를 가지는 단백질들의 상호작용 비율이고, 두 번째는 같은 클러스터에 포함된 단백질들 간의 상호작용 비율이다. 실제 PPI 데이타에서 이 두 가지의 비율은 랜덤으로 생성한 그래프에서 보다 월등히 높다. 그림 1은 Hui et al.의 실험 결과로, (a), (c), (e)는 기능 분류들 간의 상호작용 비율을 보여주며, (b), (d), (f) 그림은 클러스터 간과 클러스터 내의 상호작용 비율을 보여준다. 그림 1의 (a), (c), (e) 그래프의 양 축은 단백질 기능 분류를 나타내며, 따라서 대각선은 같은 기능 간의 상호작용을 의미한다. 그리고 각 셀의 색깔이 노란색일수록 상호작용이 강하다는 것을 의미이다. 그림 1의 (a), (c), (e)는 실제 PPI 데이타 그래프와 랜덤하게 생성한 그래프의 쌍으로 이루어져 있다. 이를 살펴보면, 실제 PPI 데이타는 노란색 셀이 대각선에 많이 위치하지만, 랜덤 그래프에서는 그렇지 않음을 볼 수 있다. 따라서 실제 PPI 데이타는 랜덤하게 생성한 그래프와는 달리, 같은 기능 분류를 가지는 단백질 간에 상호작용이 더 잘 일어남을 알 수 있다. 그림 1의 오른쪽 그림 (b), (d), (f)는 그래프를 클러스터로 나누어서, 클러스터 간의 상호작용과 클러스터 내에 존재하는 단백질 간의 상호작용 비율을 나타낸 것이다. 그림 1(b), (d), (f)의 초록색은 클러스터간의 상호작용을 나타내는 것이며, 보라색은 클러스터 내의 상호작용을 나타내는 것이다. 실제 PPI 데이타에서는 클러스터 내의 단백질 간의 상호작용이 클러스터 간의 상호작용에 비해 훨씬 많으나, 랜덤 그래프에서는 둘 다 비슷함을 알 수 있다. 따라서 PPI 데이타는 유사한 기능끼리 상호작용을 더 잘하고, 서로 상호작용하는 단백질들끼리 클러스터를 이룬다는 것을 알 수 있다. 이 결과는 이미 알려진 단백질의 특성과 일치하므로 PPI 데이타는 실제로 생물학적인 의미를 가진다는 것을 알 수 있다[1].

최근 생물정보학에서 많이 다루어지고 있는 주제는 단백질 기능 예측이다. 실험을 통해서 단백질의 기능을 밝혀내기 위해서는 실험해야 할 대상이 많기 때문에, 많은 시간과 비용이 소요된다. 따라서 정확한 단백질의 기능 후보들을 선별해 낼 수 있다면 실험 시간과 비용을 획기적으로 축소시킬 수 있기 때문에, 단백질의 기능을 예측하기 위해 많은 연구가 이루어지고 있다. 단백질의 기능을 예측하는 방법에는 크게 세 분류가 있다. 우선 단백질의 유전자 서열인 ORF(Open Reading Frame)의 유사도를 이용하는 방법을 들 수 있다[10]. ORF의 유사도를 이용하는 방법은 ORF의 배열이 조금만 달라져도 단백질의 3차원 구조가 달라지므로, ORF의 유사도가

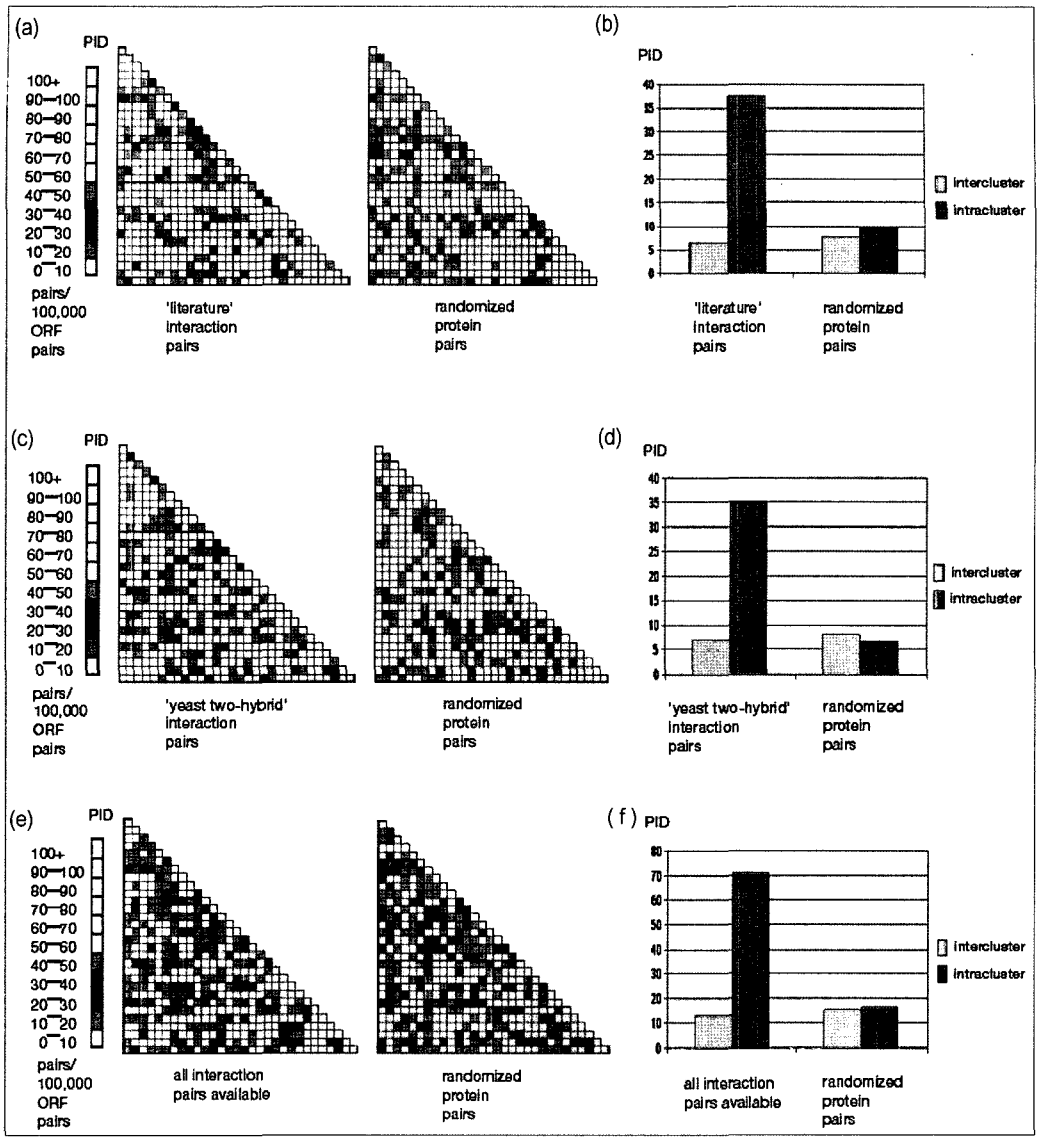


그림 1 PPI 데이타와 랜덤 데이타의 비교. 실제 PPI 데이타는 랜덤하게 생성한 그래프와는 달리, 유사한 기능끼리 상호작용을 더 잘하며 서로 상호작용하는 단백질들끼리 클러스터를 이룬다[9]. (b), (d), (f)의 연한 색으로 표현된 그래프는 클러스터간의 상호작용을 나타내는 것이며, 진한색은 클러스터 내의 상호작용을 나타내는 것이다.

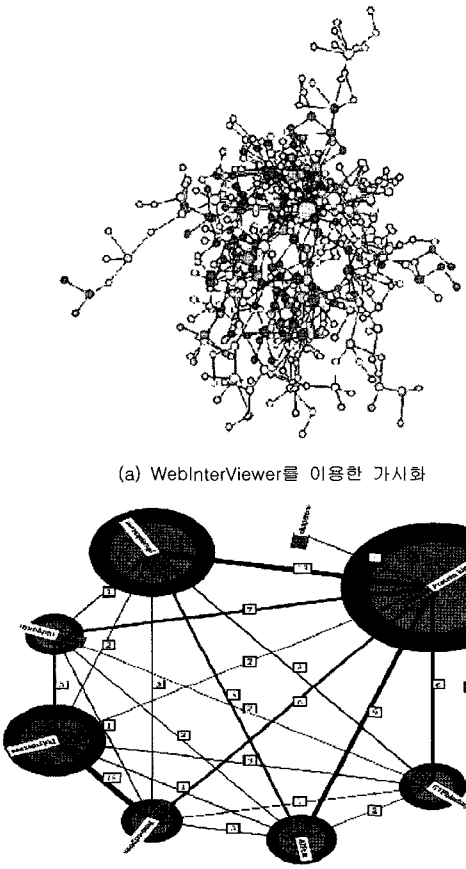
높더라도 같은 기능을 할 확률이 낮다는 문제점이 있다. 통계적인 기법을 이용하는 Markov Random Field 방법은 각 기능의 전체적인 발생 빈도를 조사하고, 이를 기반으로 단백질의 기능을 예측하는 방법이다[11,12]. 마지막 방법으로는 PPI 데이타의 그래프 특성을 이용하는 방법이다. 이를 이용한 대표적인 방법으로 Majority Rule를 들 수 있다. Majority Rule은 기능을 모르는 단백질과 상호작용하는 단백질들이 가지는 기능들의 빈도수를 세어서, 그 빈도수가 가장 높은 기능을 할당하는 방법이다[13,15]. 비슷한 방법으로 χ^2 식을 이용한 방법이 있는데, 이 방법은 기능을 할당할 때 상호작용하는 단백질의 기능 빈도수뿐만 아니라 전체 빈도수도 함께 고려한다는 차이점이 있다[15]. 또한 노드 간의 연결이 많은 HCS(Highly Connected Subgraph) 구조는 같은 기능을 가질 확률이 높다는 점을 근거로 quasi-clique을 계산하여 기능을 할당하는 방법도 있다[16].

PPI 데이타는 서로 상호작용하는 단백질들의 데이타이므로, 보다 직관적인 이해를 돕기 위해 가시화할 필요가 있다. 하지만 PPI 데이타에는 많은 노드와 간선이 포함되어 있으므로, 가시화한 결과 그래프의 간선이 교차되거나 노드가 겹치게 된다. 이는 사용자가 단백질 상호작용 데이타를 이해하는데 방해 요인이 되므로 그래프를 보다 이해하기 쉽도록 가시화할 수 있는 방법이 필요하다. WebInterViewer[17]는 PPI 데이타를 웹상에서 가시화해주는 프로그램이다. 그림 2의 (a)는 WebInterViewer로 Yeast의 PPI 데이타를 가시화한 그래프이다. 하지만 아무리 좋은 레이아웃을 개발하더라도 가시화하려는 PPI 데이타가 방대하다면 소용없다. 이러한 문제점을 해결하기 위하여, PPI 전체 그래프를 간략화하여 가시화하는 방법에 대한 연구가 진행되고 있다. CNplot[18]은 PPI 데이타에 클러스터링 정보를 적용하여 간략화하는 방법으로, 간략화한 그래프의 노드 크기는 클러스터에 포함된 단백질의 수를 나타내며 노드 경계선의 굵기는 클러스터에 포함된 상호작용 수를 의미한다. 그리고 간선의 굵기는 연결된 두 클러스터 간의 상호작용 수를 의미한다. 그림 2의 (b)는 Benno et al.[14]의 PPI 데이타를 CNplot으로 간략화한 예이다.

본 논문에서는 PPI 데이타를 다양한 방법으로 손쉽게 분석할 수 있는 워크벤치(workbench) 시스템인 Proteinca (PROTEin INteraction CAbaret)에 대하여 소개한다. Proteinca는 다양한 데이타베이스의 PPI 데이타를 그래프이론 기반의 분석 기능들을 제공하며, 그래프로 가시화하여 사용자가 직관적으로 이해할 수 있도록 도와준다. 또한, 중력 모델 기반의 간략화 방법을 제공하여 사용자에게 중요 단백질 중심의 가시화를 제공한다. Proteinca는 전 세계적으로 많이 사용되는 LEDA(Library of Efficient Data Types and Algorithms) 라이브러리[19]를 이용하여 개발되어 알고리즘의 신뢰성을 보장한다.

2. 단백질-단백질 상호작용 데이타분석을 위한 Proteinca 시스템 개발

Proteinca의 궁극적인 목표는 주어진 PPI 데이타를 그래프로 가시화하는 것과, Graph Theory를 이용하여 단백질의 기능을 예측하는 것이다. PPI 데이타는 파일과 PPI 데이타베이스를 통해 입력받을 수 있다. Proteinca에서 사용할 수 있는 PPI 데이타베이스는 DIP, MIPS, 그리고 BIND등이다. Proteinca에 PPI 데이타를 입력하면, 입력 데이타를 기반으로 유닉스의 쉘과 같은 방식으로 명령을 입력하여 불필요한 정보를 제거하거나 그래프의 특성 조사, 단백질의 기능 예측 등을 할 수 있다. 그리고 분석된 결과를 그래프로 가시화하거나 Proteinca 파일 포맷인 PIG(Protein Interaction Graph) 파일로 저장할 수 있다.



(a) WebInterViewer를 이용한 가시화

(b) CNPlot을 이용한 가시화

그림 2 PPI 데이타의 가시화. (a) WebInterViewer를 이용한 가시화[15]. (b) CNPlot을 이용한 가시화[16]. 대용량의 PPI 데이타를 사용자가 인지하기 쉽도록 가시화하기 위해서는 그 특징을 잘 나타낼 수 있는 간략화 방법이 요구된다.

Proteinca는 MFC를 이용하여 구현되었고, 가시화를 비롯한 그래프 관련 기능은 LEDA를 이용하였다. Intel Pentium 4, Windows XP SP1a 환경에서 테스트하였다. Proteinca의 기능은 다음과 같다.

- 그래프 가시화 기능
- 그래프 연산 기능
- 데이타베이스 기능

그림 3은 Proteinca의 구조도를 나타낸다. Proteinca는 그림 3과 같이 크게 6가지의 모듈로 나누어 볼 수 있다.

2.1 PPI 데이타의 입출력과 전처리

PPI 데이타 입출력 기능은 PPI 데이타를 읽고 쓰는

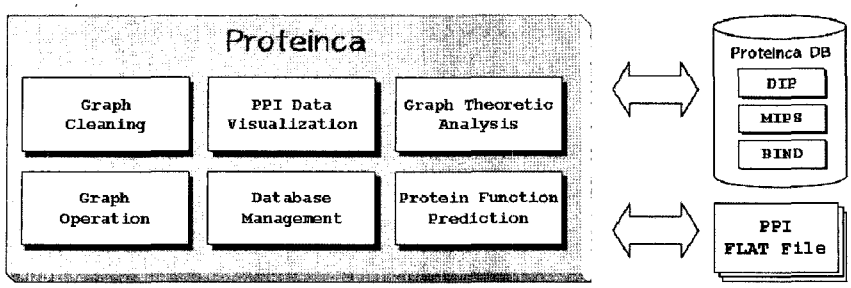


그림 3 Proteinca의 구조도 : PPI 데이타는 파일과 PPI 데이타베이스를 통해 입력받을 수 있다. Proteinca는 입력된 PPI 데이타를 이용하여 다양한 그래프이론 기반의 분석 기능과 가시화 기능을 제공한다.

표 2 Proteinca의 6가지 모듈. PPI 데이타를 다양한 방법으로 손쉽게 분석할 수 있다.

모 듈	설 명
Graph Cleaning	PPI 데이타를 전처리 하는 모듈이다. 그래프 분석 목적에 따라, 싱글톤이나 더블톤, 루프등의 데이타를 제거할 수 있다.
PPI Data Visualization	PPI 데이타를 가시화하는 모듈이다. LEDA를 이용하여 구현되었으며, 다양한 레이아웃을 제공한다.
Graph Theoretic Analysis	PPI 데이타를 그래프 이론을 바탕으로 분석하는 모듈이다. Shortest path, Difference 등의 연산을 제공한다.
Graph Operation	PPI 데이타들을 연산하는 모듈이다. Union, Intersection, Difference 등의 연산을 제공한다.
Database Management	Proteinca DB를 관리하는 모듈이다. 데이타를 읽거나 수정, 삭제하는 기능을 제공한다.
Protein Function Prediction	PPI 데이타를 이용하여 단백질의 기능을 예측하는 모듈이다. Majority Rule[11,12]과 χ^2[13] 방법을 적용하여 예측한다.

기능이다. Proteinca는 두 가지 방법으로 PPI 데이타를 읽을 수 있다. 하나는 공개된 PPI 데이타베이스에서 제공하는 플랫 파일을 읽는 방법이며, 두 번째는 Proteinca DB를 이용하는 방법이다. Proteinca 시스템에서는 DIP, MIPS, BIND의 플랫 파일을 이용하여 제작한 Proteinca DB를 제공하는데, 이 데이타베이스를 이용함으로써 PPI 데이타를 안전하고 편리하게 읽을 수 있다. 또한 데이타베이스에 접근할 수 있도록 Database Management 모듈도 제공한다. Database Management 모듈에서는 데이타베이스를 검색하거나 데이타를 추가 혹은 삭제할 수 있으며, 새로운 플랫 파일을 이용하여 데이타베이스를 새롭게 업데이트할 수 있다. 또한, 사용자가 분석한 결과를 데이타베이스에 저장함으로써 연구의 연속성을 보장한다.

Proteinca에서 사용할 수 있는 플랫 파일은 DIP과 MIPS, BIND 등의 플랫 파일이다. 하지만 일반적인 PPI 데이타도 지원하기 위해, 단순히 상호작용하는 두 단백질만으로 이루어진 데이타를 읽는 기능도 제공한다. 이외에도, Proteinca의 플랫 파일 포맷인 PIG(Protein Interaction Graph) 포맷으로 PPI 데이타를 읽고 저장할 수 있다. PIG 포맷은 크게 세 부분으로 구성되며, 첫 번째 부분은 노드와 간선 수 등 그래프의 기본적인 특징에 관한 부분이며, 두 번째는 노드, 세 번째 부분은 간선의 속성에 관하여 설명된 부분이다. 이러한 PIG 포맷은 노드의 위치와 색깔, 간선의 색깔에 관한 정보도 함께 가지고 있는 Embedded 버전과 이러한 정보가 없는 Normal 버전 등 두 종류가 있다. 데이타베이스나 플랫 파일로부터 입력 받은 PPI 데이타에는 분석하고자 하는 목적에 따라 그래프를 분석하는데 불필요한 정보도 포함되어 있다. 따라서 분석에 앞서 불필요한 요소를 제거하는 전처리 과정이 필요하며, Proteinca에서는 이를 위해 Graph Cleaning 기능을 제공한다. Protenica에서 제공하는 데이타 필터링 기능은 루프(loop), 싱글톤, 더블톤, 평행 간선(parallel edge) 제거 등이 있다.

2.2 그래프이론 기반의 분석

PPI 데이타는 그래프 데이타로 추상화할 수 있으므로, 이를 바탕으로 분석할 수 있다. Proteinca에는 컴포넌트의 수나 degree 등의 간단한 그래프 정보뿐만 아니라, 최단 패스(shortest path), 지름(diameter), 절단점(cut vertex) 등의 그래프 알고리즘을 이용하여 분석하는 기능도 제공한다. Shortest path는 주어진 노드 사이의 최단 거리를 구하는 알고리즘으로, PPI 데이타의 모든 간선을 가중치가 1이라고 가정한다. 지름은 모든 최단 패스 중에 가장 긴 패스를 의미하며, cut vertex는 그 노드를 지우면 하나의 컴포넌트가 두 개 이상의 컴포넌트로 나누어지는 노드를 말한다. 그림 4는 Proteinca 시스템에서 1333N 단백질과 1877N 단백질 사이의 최단 패스를 나타낸 것이다.

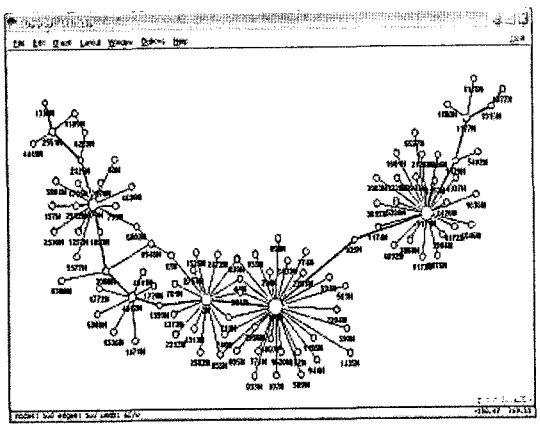


그림 4 Proteinca에서 제공하는 최단 패스의 결과 화면 : Proteinca 시스템에서는 전체 PPI 데이타에서 특정 단백질쌍의 최단 패스를 알려준다. 1333N과 1877N 단백질사이의 최단 거리는 14이다.

2004년 발표된 Natasa Przulj et al.[20]의 연구에서는 PPI 데이타를 그래프 이론을 사용하여 PPI 데이타에서 돌연변이에 의해 치명적인 작용을 하는 단백질들의 특징을 밝혔다. Natasa Przulj et al.에서는 그래프 이론에서 사용하는 HCS(Highly Connected Subgraph), 병목점, 절단점 등의 구조를 PPI 데이타에서 찾아내며, 이를 바탕으로 치명적인 단백질이 HCS 구조에 포함될 뿐만 아니라 병목점이라는 점도 밝혀내었다.

최근 PPI 데이타를 이용한 단백질 기능예측 알고리즘 개발에도 많은 연구가 진행되고 있다. 현재까지 PPI를 이용한 단백질 기능 예측 방법론에는 Majority rule[21]이나, Random Markov Model[22], PPI 구조 분석[19] 등이 있다. Majority rule은 그래프 기반의 기능 예측 방법 중 가장 먼저 연구된 방법이다. 이 방법에서는 기능을 알고자하는 단백질의 기능을 알기위해, 그 단백질과 상호작용 하는 단백질들의 기능을 살펴보는 것이다. 상호작용 하는 이웃 단백질들의 기능들을 살펴보고, 그 중에서 가장 빈도가 높은 기능을 할당하는 것이다. 만약 빈도가 가장 높은 기능이 여러 가지일 경우에는 모두 할당한다. 그림 5는 Majority rule의 예를 보여준다.

이러한 연구들은 "유사한 기능을 하는 단백질들이 상호작용한다."는 개념[9]을 바탕으로 한 것이다. 하지만, 특정 단백질들은 그 단백질과 상호작용하는 단백질들이 가지고 있는 기능들과는 전혀 다른 기능을 하는 경우가 있다. 그림 6은 이러한 경우의 예를 나타낸다. 그림 6은 YPD의 데이타로, cellular role 기능을 가진 단백질들이다. 그림 6의 단백질 RPS28A와 RPS28B 단백질의 기능을 예측하기 위해 Majority rule을 사용할 경우, RPS28A와 RPS28B 단백질들의 이웃하는 단백질의 기

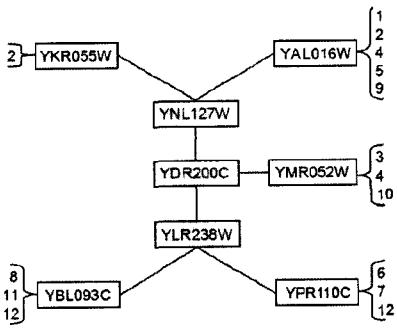


그림 5 Majority rule의 예 : 사각형은 각 단백질을 뜻하며, 에지로 연결된 단백질들은 서로 상호작용하는 단백질을 뜻한다. 단백질 옆에 적혀있는 숫자는 해당 단백질의 기능을 간단하게 나타내는 것이다. 여기서 숫자가 없는 회색 사각형은 기능을 알고자 하는 단백질이다. 단백질 YNL127W의 기능을 알기 위해 Majority rule을 이용하면, 상호작용하는 단백질인 YAR055W, YAL016W와 YDR200C의 기능을 살펴본다. 하지만, 이중에서 YDR200C는 기능을 알지 못하는 단백질이므로 이를 제외하고 살펴보면, YNL127W와 상호작용하는 단백질들의 기능에는 2가 2번, 기능 1,4,5,9가 있다. 따라서 YNL127W의 기능은 2로 할당한다.

능에 RPS28A와 RPS28B의 실제 기능이 포함되어 있지 않기 때문에 올바른 예측을 할 수 없다. 그림 6의 그림은 Proteinca 시스템에서 RPS28A와 RPS28B 단백질을 중심으로 각각 깊이 1만큼씩 전체 YPD 데이타에서 추출한 것이다.

Proteinca에서는 그림 6과 같이 한 단백질이 그 이웃하는 단백질들과 같은 기능을 포함하고 있지 않은 경우, D-protein(Disjoint protein), 몇몇 기능만이 포함되는 경우 O-protein(Overlap protein), 모든 기능이 이웃하는 단백질들의 기능에 포함될 경우 S-protein(Subset protein)으로 정의하고, 입력받은 PPI 데이타에서 이들을 찾아주는 기능을 제공하고 있다. 표 3은 MIPS에서 D-protein, O-protein, S-protein의 수를 Proteinca를 사용하여 조사한 것이다. 따라서 MIPS 데이타에서 Majority rule을 사용하면, 11.05%의 D-protein의 기능은 전혀 알 수 없고, 31.50%의 O-protein 기능은 일부만을 예측할 수 있게 된다.

2.3 PPI 데이타 가시화

Proteinca는 입력받은 PPI 데이타를 사용자가 분석하기 쉽도록 그래프로 가시화할 수 있으며, 그래프의 특성에 따라 이해하기 쉬운 레이아웃을 선택할 수 있도록 다양한 레이아웃을 제공한다. 뿐만 아니라 색깔과 굵기, 레이블을 달리하여 시각적 효과를 줄 수도 있고, 줌 기

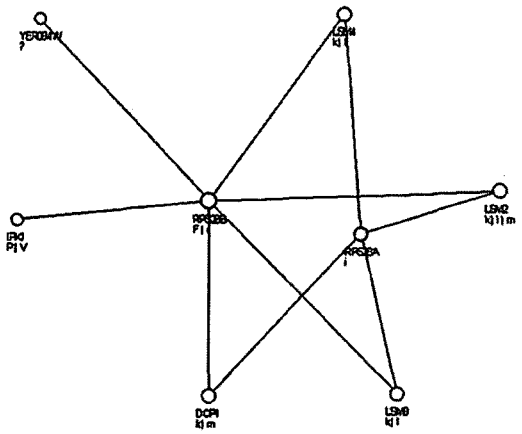


그림 6 하나의 단백질과 상호작용하는 단백질의 기능이 중복되지 않는 예 : RPS28A 단백질은 i(Protein translocation)기능을 가지지만, RPS28A와 상호작용하는 단백질인 LSM2와 DCP1 단백질의 기능에는 i(Protein translocation)기능이 없다. RPS28B 단백질도 F(Cell polarity) 기능을 가지지만, RPS28B와 상호작용하는 단백질인 LSM4, YBR094W, IPK1에는 F(Cell polarity) 기능이 없다.

표 3 MIPS 데이타에서의 D-protein, O-protein, S-protein 조사

전체 단백질 수	D-protein	O-protein	S-protein
3,387	374(11.05%)	1,067(31.50%)	1,946(57.45%)

능도 제공한다. 그리고 가시화한 화면에서 노드와 간선을 추가하거나 삭제할 수 있는 Visual Editing 기능을 제공한다.

가시화 기능을 이용하면, 사용자가 실험한 데이타를 기존의 PPI 데이타에 쉽게 추가할 수 있으며, 잘못된 데이타를 삭제할 수도 있다. 그림 7의 (a)는 Proteinca가 제공하는 다양한 레이아웃으로 PPI 데이타를 가시화한 것으로, Proteinca에서는 Random, Circular, Spring Embedder, Straight Line 등과 같은 다양한 레이아웃을 제공한다. 그림 7의 (b)는 MIPS에서 추출한 HCS 구조를 가지는 단백질들을 가시화한 그래프로써, 모두 미토콘드리아 전사에 위치한 단백질이다.

PPI 데이타는 아주 방대한 양의 단백질 상호작용 정보를 가지고 있다. 따라서 전체 PPI 데이타를 가시화하는데 많은 시간이 걸릴 뿐만 아니라, 가시화하더라도 그래프의 많은 간선이 서로 교차되고, 노드와 레이블들이 겹치는 등의 문제로 인해 그래프의 구조를 직관적으로 파악하기 힘들다. 그러므로 Proteinca에서는 그래프를 간략화하여 사용자가 이해하기 쉽도록 가시화하는 기능을 제공한다. 간략화 된 그래프는 허브 노드 간의 연결로 이루어진 그래프이므로, 사용자는 이 그래프를 살펴봄으로써 PPI 데이타의 대략적인 구조를 파악할 수 있다. Proteinca에서는 대용량의 PPI 데이타를 간략화하여 사용자에게 가시화해주는 기능을 제공하고 있다. Proteinca에서 제공하는 간략화 기능의 알고리즘은 표 4와 같다.

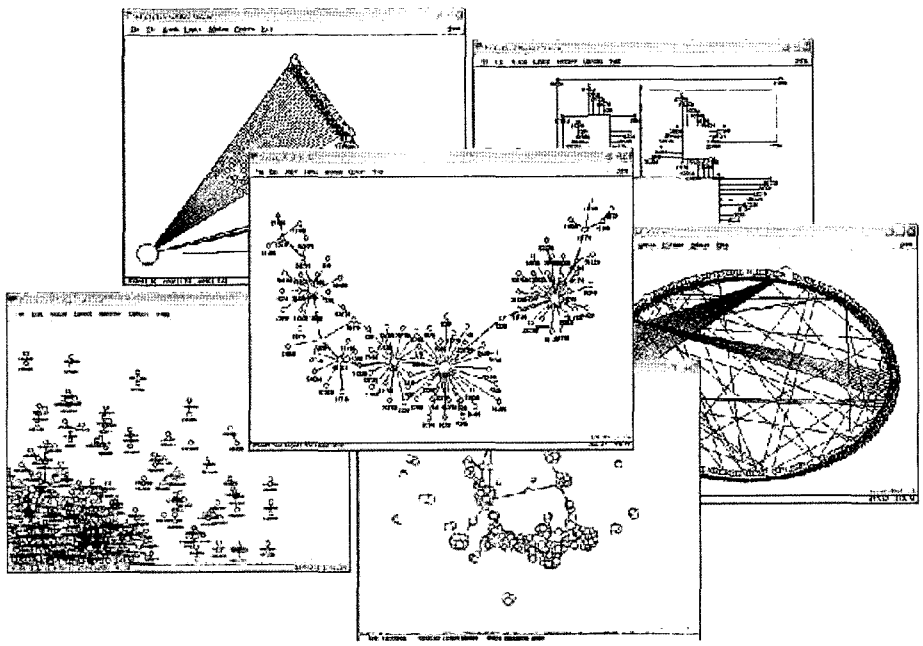
그림 8은 DIP 데이타의 일부 그래프를 간략화 되는 과정을 보여주는 그림이다. 그림 5의 (a)는 간략화하려는 원본 그래프이고, (f)는 간략화 된 결과 그래프이다. 간략화 된 그래프, 즉 (f)의 노드는 간략화 된 컴포넌트를 의미하며 간선은 각 컴포넌트 간의 상호작용을 의미한다. 따라서 노드의 크기가 클수록 그 속에 포함된 단백질들의 수가 많다는 것을 의미하며, 에지가 굵을수록 에지로 연결된 두 노드에 포함된 단백질 간의 상호작용이 많다는 것을 의미한다. 즉, 노드와 간선의 굵기는 컴포넌트에 포함된 노드의 상대적인 크기와 컴포넌트 간의 상호작용의 상대적인 빈도를 의미한다.

그림 9와 그림 10은 그림 8의 (a) 원본 그래프에 서로 다른 매개변수를 적용하여 간략화한 결과 그래프이다. 간략화 알고리즘에서 조절 가능한 매개변수로는, 각 노드의 가중치를 계산할 때 고려하는 깊이(depth)의 크기와 간략화시 통합될 노드 쌍을 결정하는 임계값이다. 깊이를 넓게 보게 되면, 보다 멀리 떨어져있는 단백질들을 고려하게 된다. 많은 각 단백 그림 9의 그래프들은 모두 깊이가 2이고, 임계값은 (a)부터 차례로 20, 40, 50, 60인 경우이며, 그림 10의 그래프들은 깊이가 3이고, 임계값은 (a)부터 차례로 40, 60, 80, 100인 경우의 간략화 그래프이다. 그림 8의 (a)에서 색이 진하게 칠해진 노드는 상호작용이 많은 허브 노드로써, PPI 데이타에서 중요한 역할을 하는 단백질이다. 이 노드들은 그림 9과 그림 10의 그래프들에서 잘 보존되어 있음을 알 수 있다. 따라서 Proteinca에서 제공하는 간략화 알고리즘은 원본 그래프의 중요한 특징을 유지하면서 간략화 함을 알 수 있다.

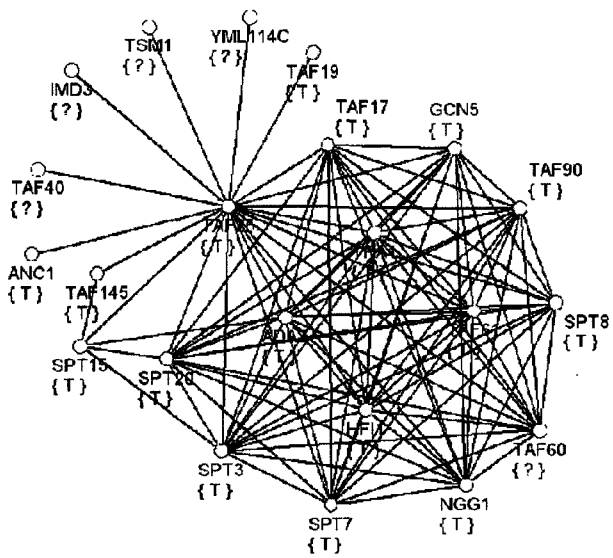
2.4 PPI 데이타 통합 및 분리 기능

연구자들은 DIP, MIPS, BIND 등과 같은 PPI 데이타베이스나 연구자의 개인 데이타 등 다양한 데이타를 이용하여 PPI 데이타를 분석한다. 또한, DIP 데이타와 BIND 데이타를 함께 사용할 수도 있으며, 특정 자료의 일부만을 사용하여 분석하고자 할 수도 있다. 따라서 다양한 사용자의 요구를 만족시키기 위하여, Proteinca 시스템에서는 그래프 연산 기능을 제공한다.

Proteinca에서 제공하는 그래프 연산 기능은 Union과 Intersection, Difference 등이 있다. 이러한 그래프 연



(a) 다양한 layout 제공



(b) HCS 추출의 예, T : Mitochondrial Transcription에 위치

그림 7 Proteinca의 가시화. (a) Proteinca가 제공하는 다양한 레이아웃. (b) MIPS에서 추출한 HCS의 가시화, 추출된 단백질들이 대부분이 Mitochondrial Transcription에 위치한 단백질들이다.

산 기능을 적용하기 위해서는, 사용되는 PPI 데이타들 간의 단백질 ID가 동일해야 한다는 제약이 있으며 일반적인 경우 ORF(Open Reading Frame)를 사용하여 연산한다. 표 5는 Proteinca에서 제공하는 그래프 연산 기능을 설명한 것이다.

그림 11은 Proteinca를 이용하여 두 PPI 데이타를 Union한 결과이다. (a)는 DIP PPI 데이타 중 Energy에 관련된 기능을 하는 21개의 단백질로 구성된 하나의

표 4 PPI 데이타의 간략화 : 복잡한 PPI 데이타를 입력으로 간략화 과정을 거치면, 상호작용이 많은 노드들, 즉 허브들 간의 연결 구조로 간략화 된 그래프를 얻을 수 있다.

간략화 알고리즘

Step 1. PPI 데이타의 모든 단백질에 대해 다음 3가지의 Majority Rule의 식 중 하나를 적용하여 가중치를 계산한다.

$$w_1 = 1$$

$$w_2 = \frac{1}{k},\ k = depth$$

$$w_3 = (\frac{1}{2})^{k-1}$$

Step 2. 모든 단백질 쌍, P_i와 P_j에 대해 gravity G_{ij}를 구한다.

$$G_{ij} = \frac{w_i / w_j}{d_{ij}^2}$$

P_i와 P_j는 상호작용하는 한 쌍의 단백질을 의미하며, w_i는 단백질 P_i의 가중치를 의미한다. d_{ij}는 단백질 P_i와 P_j의 shortest path를 의미하며, 이 값은 원본 그래프에서 계산한 값을 사용한다.

Step 3. 전체 PPI 데이타에서 일정 크기 이상의 gravity를 가지는 단백질 쌍을 하나의 단백질로 합한다. gravity 값은 두 단백질의 가중치 비에 비례하고 거리의 제곱 값에 반비례하므로, 가중치의 차가 크고 거리가 가까운 단백질일수록 쉽게 합쳐진다.

Step 4. 간략화 되는 단백질이 없을 때까지 2, 3 과정을 반복한다.

Step 5. 간략화 과정이 끝나면, 간략화 된 노드들이 모인 각 컴포넌트들의 크기는 컴포넌트에 속한 단백질의 수에 비례하여 결정하며, 컴포넌트간의 간선의 굵기는 컴포넌트들 간의 상호작용의 수에 비례하여 결정한다.

컴포넌트이며, (b)는 MIPS PPI 데이타 중 Energy에 관련된 기능을 하는 12개의 단백질로 구성된 두 개의 컴포넌트이다. (c)는 (a)와 (b)의 합집합으로 (b)에 나타나는 MIPS의 두 개의 컴포넌트가 (a)에 나타나는 DIP의 컴포넌트에 의해 하나로 연결됨을 볼 수 있다. (c)에서 가로줄 무늬 노드는 DIP과 MIPS에 공통적으로 존재하는 단백질이며, 연한색은 DIP, 진한 색은 MIPS에만 존재하는 단백질을 뜻한다. 따라서 여러 개의 PPI 데이타를 함께 사용하면, 개개의 PPI 데이타에서는 나타나지 않던 정보가 추가되거나 존재했던 정보들이 사라질 수 있다.

von Mering C. et al.[1]과 같이 보다 신뢰성 있는 PPI 데이타를 대상으로 분석하고자 한다면, Proteinca의 Intersection 기능을 사용하여 여러 PPI 데이타에 공통적으로 존재하는 데이타들을 추출하여 생성한 PPI 데이타를 사용하면 된다. 반면, 보다 많은 정보를 대상으로 실험하고자 한다면 Proteinca의 Union 기능을 사용하여 생성한 PPI 데이타를 사용하면 된다. 이외에도 Union 기능을 이용하면 개개의 실험실에서 실험한 PPI 데이타를 기존의 PPI 데이타에 추가하여 분석할 수도 있다.

3. 결 론

PPI 데이타는 상호작용하는 단백질들에 대한 정보를 담고 있으므로, 신약 개발, 질병 연구 등 다양한 분야에서 사용된다. 하지만, PPI 데이타는 방대한 양의 데이타를 저장하고 있으며 텍스트 기반의 데이타이므로, 사람이 직접 분석하기는 힘들다. 그러므로 PPI 데이타를 분석하는데 도움을 줄 수 있는 프로그램이 필요하다. 따라서 본 논문에서는 PPI 데이타를 가시화하고 분석하기위해 개발된 Proteinca 시스템에 관하여 서술하였다.

Proteinca는 PPI 데이타를 분석하는데 유용한 많은 기능을 제공한다. Proteinca를 이용하면 PPI 데이타를 한결 손쉽게 분석할 수 있으며, 많은 PPI 데이타로부터 의미 있는 정보를 얻어낼 수도 있다. Protienca는 다음과 같은 기능을 제공한다.

- 복잡한 PPI 데이타를 그래프의 허브 노드를 중심으로 간략화해주는 알고리즘을 제공함으로써, PPI 데이타의 구조를 쉽게 파악할 수 있도록 도와준다.
- 다양한 PPI 데이타를 포함한 데이타베이스를 구축함으로써, 사용자가 쉽고 간편하게 데이타를 읽을 수 있도록 도와준다.
- 그래프 연산 기능을 통해, 기존의 PPI 데이타에서 사용자가 원하는 부분만을 추출하여 사용할 수 있는 기능을 제공한다.
- 그래프 이론을 바탕으로 한 다양한 PPI 데이타 분석 기능을 제공한다.
- Proteinca는 2D 및 3D Spring Embedder, Circular, Random, Straight Line, Orthogonal 등의 다양한 레이아웃으로 PPI 데이타를 가시화할 수 있다.

본 논문에서 서술한 Proteinca는 PPI 데이타를 분석

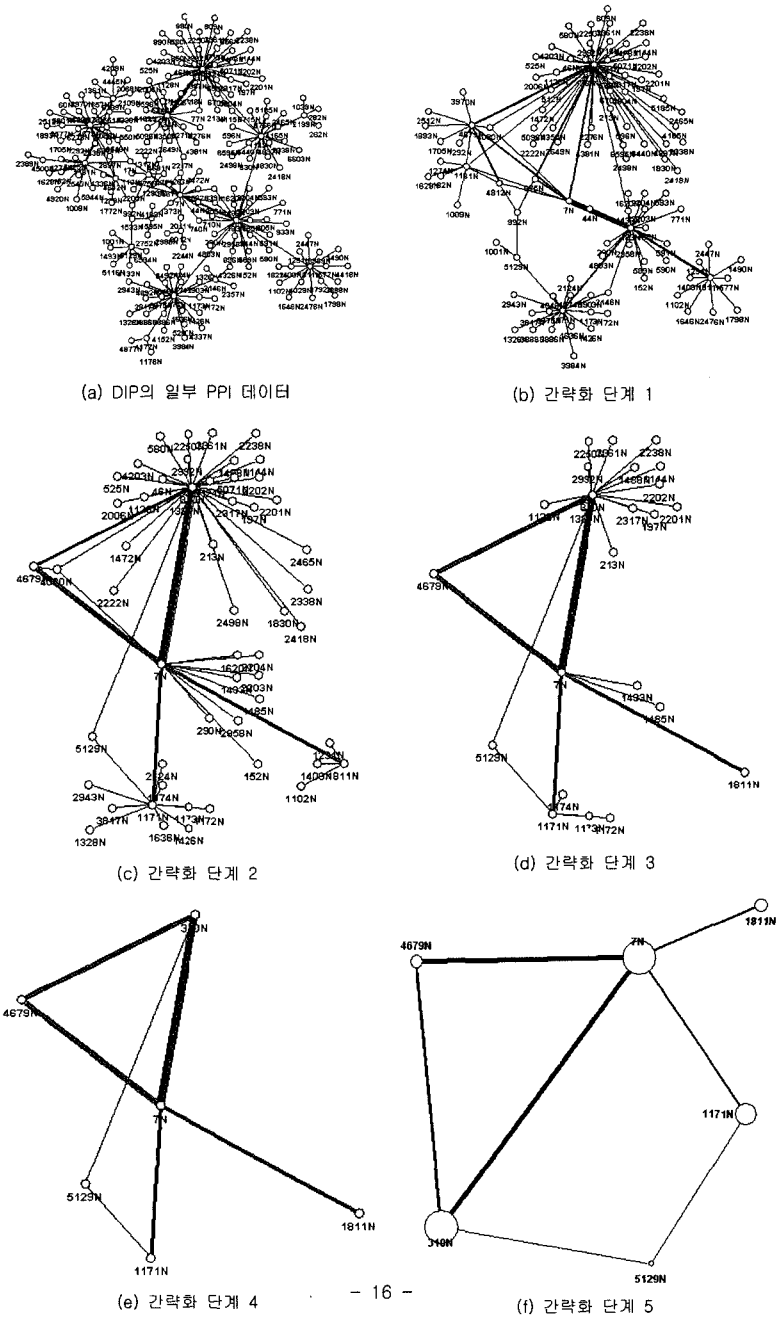


그림 8 PPI 데이타 간략화 과정. (a)는 간략화하려는 DIP의 PPI 데이타 일부를 가시화한 것이다. (b)~(f)까지는 간략화 되는 과정을 나타낸 것이다. (f)는 간략화된 결과 그래프로 노드는 간략화된 컴포넌트를 의미하며, 간선은 각 컴포넌트 간의 상호작용을 의미한다. 각 노드의 크기는 컴포넌트에 속한 단백질의 수에 비례하며, 간선의 굵기는 컴포넌트간의 상호작용의 수에 비례한다. 노드의 ID는 컴포넌트의 중심 단백질을 뜻한다. 실제 데이타인 (a) 그래프에서 진하게 색이 칠해져있는 노드는 상호작용이 많은 허브 노드이며, (f)의 간략화 결과 그래프에서의 각 컴포넌트 중심 노드들이 (a)에서 나타난 허브 노드들임을 알 수 있다.

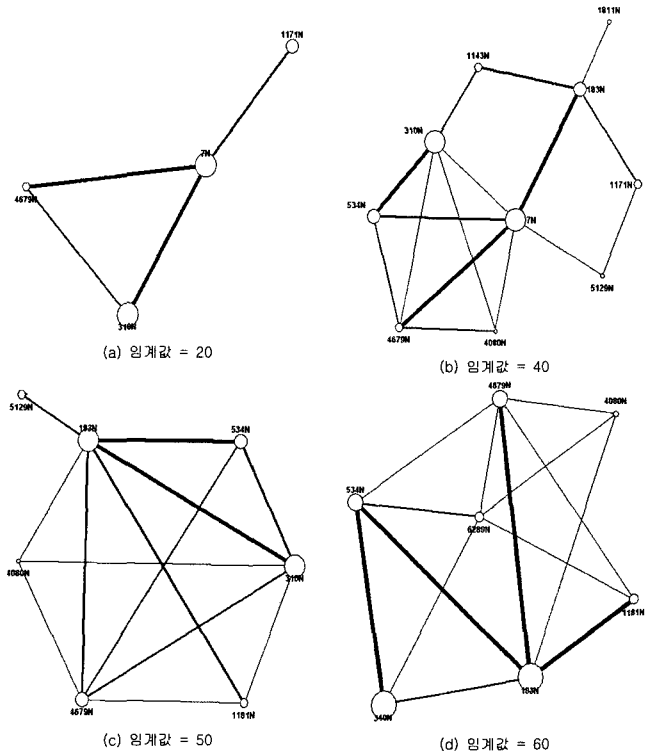


그림 9 그림 5의 (a) 그래프를 간략화 한 결과. 깊이가 2이고, 임계값은 (a)부터 차례로 20, 40, 50, 60인 경우

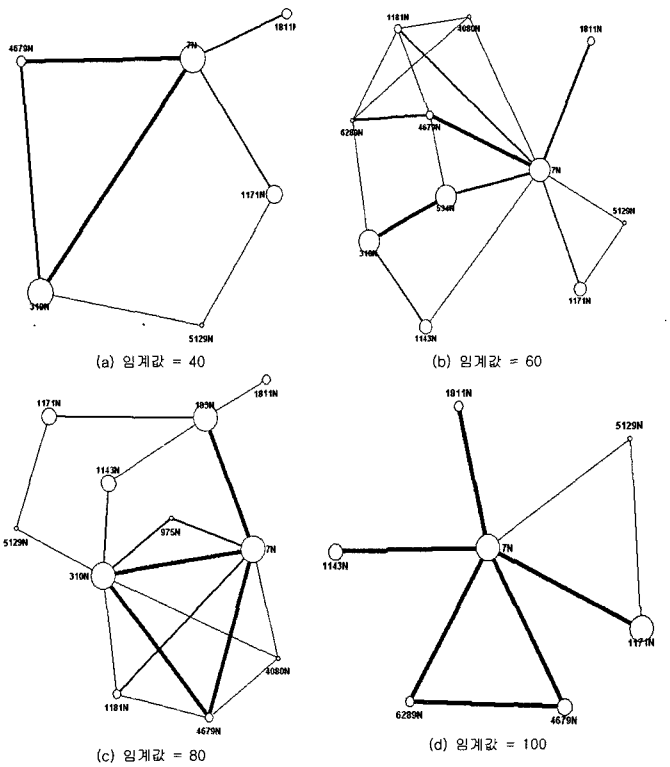
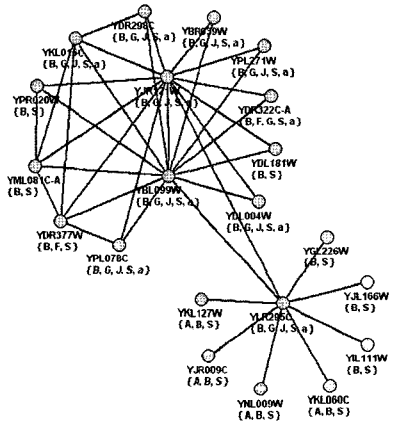


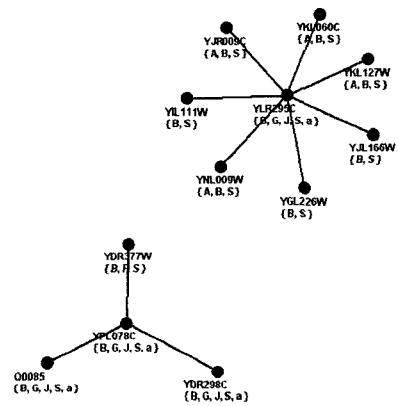
그림 10 그림 5의 (a) 그래프를 간략화 한 결과. 깊이가 3이고, 임계값은 (a)부터 차례로 40, 60, 80, 100인 경우

표 5 Proteinca에서 제공하는 그래프 연산 기능. : Proteinca에서 제공하는 그래프 연산 기능은 Union과 Intersection, Difference가 있다. 그래프 연산 기능을 이용하여 여러 PPI 데이타들을 함께 사용할 수 있다.

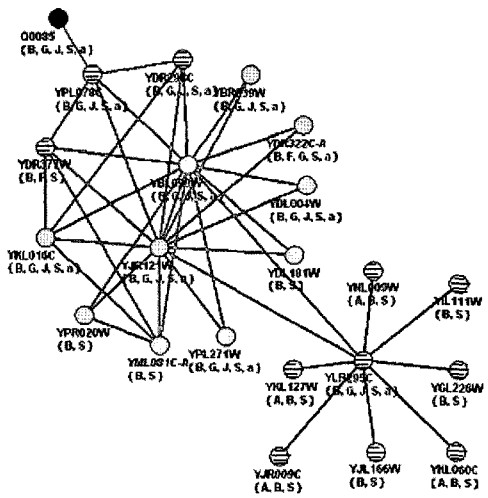
기 능	설 명
Extract Largest Component	그래프에서 가장 큰 컴포넌트를 추출하여, 새로운 그래프를 생성한다.
Union	두 그래프를 합하여 새로운 그래프를 생성한다. 합해질 때에는, 두 그래프에서 같은 노드를 찾아서, 그 노드들을 중심으로 합하게 된다.
Union To Embedded Graph	합하는 두 그래프의 공통 노드, 개별 노드들의 색깔을 달리하여, 두 그래프를 Union한다. 합해질 때에는, 두 그래프에서 같은 노드를 찾아서, 그 노드들을 중심으로 합하게 된다.
Intersection	두 그래프의 교집합으로 새로운 그래프를 생성한다. 두 그래프에서 공통되는 노드들만을 추출하고, 추출된 단백질들간에 원 그래프들 모두에서 상호작용이 있는 경우만을 상호작용을 표현해 준다.
Difference	두 그래프의 차집합으로 새로운 그래프를 생성한다. 두 그래프에서 서로 다른 부분들만을 그래프로 생성해준다.



(a) DIP의 Energy 관련 단백질



(b) MIPS의 Energy 관련 단백질



(c) DIP과 MIPS의 Energy 관련 단백질들의 union

그림 11 Proteinca의 Union 기능. (a) DIP PPI 데이타 중 Energy에 관련된 기능을 하는 21개의 단백질로 구성된 하나의 컴포넌트, (b) MIPS PPI 데이타 중 Energy에 관련된 기능을 하는 12개의 단백질로 구성된 두 개의 컴포넌트, (c) (a)와 (b)의 합집합으로 (b)에 나타나는 MIPS의 두 개의 컴포넌트가 (a)에 나타나는 DIP의 컴포넌트에 의해 하나로 연결됨을 볼 수 있다.

하는데 도움을 주는 많은 기능을 제공한다. 현재 Proteinca에서 제공하는 기능들 중 데이타베이스와 가시화 기능, 단백질 기능 예측 부분은 다음과 같이 향상시킬 수 있다.

- Proteinca DB의 강화 : Proteinca는 PPI 데이타를 손쉽게 읽을 수 있도록 Proteinca DB를 제공한다. 하지만, Proteinca DB에서 제공하고 있는 PPI 데이타는 DIP, MIPS, BIND이다. 따라서 앞으로는 YPD나 STRING과 같은 PPI 데이타도 Proteinca DB에 포함시켜야할 필요가 있다.
- 가시화 기능 강화 : Proteinca는 LEDA 라이브러리를 이용하여 PPI 데이타를 가시화한다. 하지만, LEDA에서 제공해주는 가시화 기능은, 데이타의 양이 방대한 PPI 데이타 적합하지 않다. 따라서 다른 그래프 가시화 라이브러리를 사용하여 복잡한 그래프도 사용자가 쉽게 이해할 수 있도록 가시화 기능을 강화해야 한다. 그리고 단순한 허브 구조 대신 보다 의미 있는 그래프로 간략화하기 위해 그래프 간략화 알고리즘을 향상시켜야하며, 사용자 인터페이스나 매개 변수 조절 기능 등을 추가하여 사용자가 쉽게 사용할 수 있도록 개선해야 한다.
- 단백질 기능 예측 알고리즘 개발 : Proteinca에서는 Majority Rule과 χ^2방법을 사용하여 단백질의 기능을 예측해준다. 하지만, 이 두 방법은 다른 연구자들에 의해 개발된 방법이며, 실험 데이타에 따른 성능의 편차도 심하다. 따라서 보다 안정적이고 정확하게 단백질의 기능을 예측하는 새로운 알고리즘의 개발이 필요하다.

Proteinca에 대한 자세한 정보는 http://jade.cs.pusan.ac.kr/~proten에서 얻을 수 있다.

참 고 문 헌

[1] von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P., "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, 2002.

[2] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU and Eisenberg D., "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Research*, 2004.

[3] Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S and Weil B., "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research*, 2000.

[4] Hodges PE, McKee AH, Davis BP, Payne WE and Garrels JI., "The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data," *Nucleic Acids Research*, 1999.

[5] Bader GD, Betel D and Hogue CW., "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Research*, 2003.

[6] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P and Snel B., "STRING: a database of predicted functional associations between proteins," *Nucleic Acids Research*, 2003.

[7] Karen R. Christie, Shuai Weng, Rama Balakrishnan et. al, "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms," *Nucleic Acids Research*, 2004.

[8] Bobby-Joe Breitkreutz, Chris Stark and Mike Tyer, "The GRID: The General Repository for Interaction Datasets," *Genome Biology*, 2002.

[9] Hui Ge, Zhihua Liu, George M. Church and Marc Vidal, "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae," *Nature*, 2001.

[10] Baker D and Sali A., "Protein structure prediction and structural genomics," *Science*, 2001.

[11] Minghua Deng, Kui Zhang, Shipra Mehta and Ting Chen, "Prediction of protein function using protein-protein interaction data," Journal of Computational Biology, 2003.

[12] Stanley Letovsky and Simon Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach," Bioinformatics, 2003.

[13] Vazquez A, Flammini A, Maritan A and Vespignani A., "Global protein function prediction from protein-protein interaction networks," *Nature*, 2003.

[14] Schwikowski B, Uetz P and Fields S. "A network of protein-protein interactions in yeast," *Nature Biotechnology*, 2000.

[15] Haretsugu Hishigaki, Kenta Nakai, Toshihide Ono, Akira Tanigami and Toshihisa Takagi, "Assessment of prediction accuracy of protein function from protein-protein interaction data," *Yeast*, 2001.

[16] Dongbo Bu, Yi Zhao, Lun Cai, Hong Xue, Xiaopeng Zhu, Hongchao Lu, Jingfen Zhang, Shiwei Sun, Lunjiang Ling, Nan Zhang, Guojie Li, Runsheng Chen, "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Research*, 2003.

[17] Kyungsook Han, Byong-Hyon Ju and Haemoon Jung, "WebInterViewer: visualizing and analyzing molecular interaction networks," *Nucleic Acids Research*, 2003.

[18] Nizar N. Batada, "CNplot: simple method to visualize pre-clustered networks," *Bioinformatics*, 2004.

[19] Kurt Mehlhorn and Stefan Näher, "LEDA: a

platform for combinatorial and geometric computing," *Communications of the ACM archive*, 1995.
[20] Przulj N, Wigle DA and Jurisica I., "Functional topology in a network of protein interactions," *Bioinformatics*, 2004.
[21] Peter Uetz, Benno Schwikowski and Stanlu Fields, "A network of protein-protein interactions in yeast," *Nature Biotechnoloty*, 2002.
[22] Shipara Metha, Ting Chen, Minghua Deng, Kui Zhang and Fengzhu Sun, "Prediction of protein function using protein-protein interaction data," *Journal of Moleculer Biology*, 2003.

진 희 정
2000년 부산대학교 전자계산학과 졸업(학사). 2002년 부산대학교 전자계산학과 졸업(이학석사). 2002년~2003년 국립보건원 유전체연구부 역학정보실 선임연구원. 2003년~현재 부산대학교 전자계산학과 박사과정. 관심분야는 생물정보학, 알고리즘 이론

윤 지 현
2003년 부산대학교 정보컴퓨터공학부 졸업(학사). 2005년 부산대학교 컴퓨터공학부 졸업(공학석사). 2005년~현재 (주)디오텍 선임연구원. 관심분야는 생물정보학, 임베디드시스템

조 환 규
정보과학회논문지 : 시스템 및 이론
제 33 권 제 1 호 참조