Proteinca : A System for Analysis/Visualization of Protein-Protein Interaction Networks

Proteinca : 단백질-단백질 상호작용 네트워크의 분석 및 가시화 시스템

Ji-Hyun Yoon, Hee-Jeong Jin, Hwan-Gue Cho

[1] Department of Computer Engineering, Pusan National University, Busan, Korea

*To whom correspondence should be addressed. E-mail: hjjin@pearl.cs.pusan.ac.kr

Abstract

단백질-단백질 상호작용(PPI : Protein-Protein Interaction) 데이터는 생물체가 어떠한 메커니즘으로 생명을 유지하는지에 대한 정보를 담고 있다. 최근에는 생물학자들의 실험에 의해 많은 데이터가 축적되어 있으며, 데이터베이스로 구축되어 인터넷에 공개되어 있다. PPI 데이터는 단백질를 노드(node)로, 상호작용은 에지(edge)로 갖는 그래프(Graph) 구조로 표현 가능하다. 본 논문에서는 사용자가 PPI 데이터를 쉽게 가공하고 분석할 수 있도록 그래프 이론 기반에 기반하여 구현한 Proteinca(PROTEin INteraction CAbaret) 시스템에 대해 소개한다.

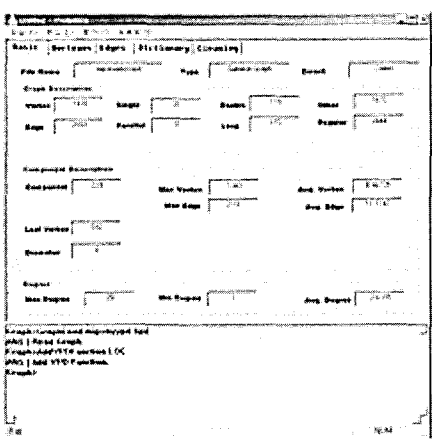
Proteinca에 대한 자세한 정보는 http://jade.cs.pusan.ac.kr/~proten 에서 볼 수 있다.

Introduction

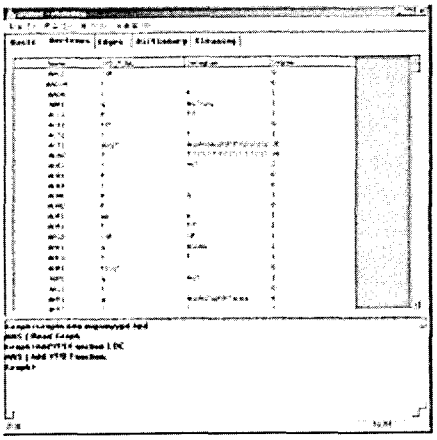
인간을 비롯한 많은 종들의 유전자 지도가 밝혀짐으로써, 인류는 Post Genome 시대를 맞이하게 되었다. Post Genome 시대에는 유전자보다는 단백질이 주요 연구 대상이 된다. 왜냐하면, 하나의 유전자는 발현되는 세포의 종류나 상황에 따라 여러 가지 단백질로 발현될 수 있으므로, 그 기능을 정확하게 파악할 수 없기 때문이다. 따라서, 최근 유전자가 발현된 형태인 단백질에 대한 연구가 늘고 있는데, 이는 단백질이 생물체 내의 물질대사에 직접 작용하므로, 단백질을 연구하면 신약개발이나 생물의 생명현상을 밝히는데 직접적인 도움을 주기 때문이다. 그 결과, 단백질의 3차원 구조, 단백질의 서열, PPI 데이터 등 단백질에 관한 많은 정보가 밝혀지고 있으며, 이 데이터들은 데이터베이스로 구축되어 인터넷을 통해 다른 연구자들에게 공개되어 있다.

단백질은 다른 단백질들과 상호작용을 통하여 생명 현상에 관여하는 기능을 수행하

게 되며, 단백질의 상호작용 정보를 단백질-단백질 상호작용(PPI : Protein-Protein Interaction) 데이터라고 한다. PPI의 데이터는 DIP(Database of Interacting Proteins)[1], MIPS(Munich Information Center for Protein Sequences)[2], BIND(Biomolecular Interaction Network Database)[3], STRING(Search Tool for the Retrieval of Interacting Genes/Proteins)[4] 등 여러 데이터베이스에서 구할 수 있다. 하지만, PPI 데이터의 방대한 양과 복잡한 구조로 인하여 사람이 직접 분석하는 것은 불가능하다. 본 논문에서는 방대한 PPI 데이터를 연구자가 다양한 방법으로 손쉽게 분석할 수 있는 워크벤치(workbench) 시스템인 Proteinca(PROTEin INteraction CAbaret)에 대하여 소개한다. Proteinca는 다양한 데이터베이스의 PPI 데이터를 그래프로 가시화하여 사용자가 직관적으로 이해할 수 있도록 도와주며, 그래프 이론에 기반한 PPI 데이터 분석 기능을 제공한다. 그림 1은 Proteinca의 실행 모습이다.



(a) Proteinca 실행 화면



(b) PPI의 단백질 리스트

그림 1. (a) Proteinca의 실행 화면, 입력된 PPI 그래프에 대한 컴포넌트의 수, 단백질의 수, 상호작용의 수와 같은 기본적인 정보를 보여준다. (b) MIPS PPI 데이터의 상세 정보를 사용자에게 리스트로 보여주는 화면, 단백질의 아이디, 기능, 상호작용하는 단백질들의 기능, 상호작용하는 단백질의 수를 살펴볼 수 있다.

Protein-Protein Interaction Data

PPI(Protein-Protein Interaction)는 서로 상호작용하는 단백질들에 대한 정보들의 집합이다. PPI 데이터는 단백질들이 서로 상호작용을 통해 특정 기능을 수행하기 때문에 생물체의 물질대사나 질병 등을 연구할 때 중요하게 사용된다. 많은 생물학자들이 질량 분석법, Yeast Two-Hybrid, Protein Microarray 등의 실험을 통해 단백질 간의 상호작용을 밝혀내고 있으며, 그 결과를 데이터베이스화하여 인터넷으로 공개하고 있다. 현재 다양한 PPI에 관한 데이터베이스들이 존재하지만, 각 데이터베이스들에 중복되어 있는 단백질 데이터는 많지 않으며, 각 데이터베이스들의 정확도도 높지 않다. 따라서 보다

정확한 PPI 정보를 얻기 위해서는 여러 데이터베이스에서 중복되는 데이터들을 사용하는 것이 좋다[5]. 공개된 PPI 데이터베이스로는 다음과 같은 것이 있다.

① DIP[1]

DIP(Database of Interacting Proteins)은 여러 단백질들의 상호작용에 관한 데이터베이스로, 전체 데이터베이스에서 Drosophila Melanogaster, Saccharomyces Cerevisiae, Caenorhabditis Elegans의 정보가 대부분을 차지한다.

② MIPS[2]

MIPS(Munich Information centre for Protein Sequences)는 독일의 Max-Planck-Institute bioinformatics 그룹에서 생성한 단백질 데이터베이스로, genome sequence의 기능 분석 및 분류에 중점을 두고 있다.

MIPS에서는 genome에 대한 총체적인 정보를 제공하기 위해서 PEDANT라는 서버를 운영 중이다.

③ BIND[3]

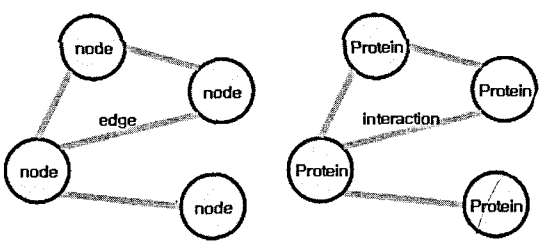
BIND에는 단백질의 상호작용 정보 외에 상호작용들의 pathway 정보를 함께 제공한다. BIND에서 제공하는 pathway 정보는 " 어떠한 질병과 관련이 있다 또는 cell cycle에 포함된다" 식의 추가 정보를 제공받는다.

④YPD[6]

YPD 데이터베이스는 워싱턴 대학교의 P.Uetz 등이 대량의 효모 유전자를 대상으로 하여, yeast two-hybrid 기법과 protein microarray를 이용한 단백질 상호작용 실험의 결과로 대량의 단백질 상호작용에 대한 정보를 획득하면서 구축되었다.

PPI 데이터는 인터넷에서 플랫 파일의 형태로 다운받을 수 있으며, 이 파일에는 한 행마다 상호작용한 두 단백질과 그 단백질의 정보가 들어 있다. PPI 데이터의 크기는 2003년12월 기준으로, MIPS는 4,336개의 단백질과 10,467개의 상호작용의 정보가 저장되어 있으며, DIP에는 4,718개의 단백질과 15,128개의 상호작용 정보가 존재한다.

PPI 데이터의 각 단백질은 그래프의 노드로, 단백질 간의 상호작용은 그래프의 에지로 표현 가능하므로, 전산학의 그래프 자료구조로 변환이 가능하다. 그림 2는 PPI 데이터를 그래프로 변환하는 것을 나타낸다. (a)는 일반적인 그래프 데이터이며, (b)는 PPI 데이터를 그래프로 표현한 것이다.



(a) 일반적인 그래프 (b) PPI의 그래프 표현

그림 2. (a) 일반적인 그래프, (b) PPI 데이터의 단백질은 그래프의 노드로, 상호작용은 그래프의 에지로 매핑하여 그래프 자료구조로 표현

본 장 이후부터는 PPI를 그래프구조로 표현하여, PPI의 단백질을 노드, 상호작용을 에지라고 표현한다.

Proteinca(PROTEin INteraction CAbaret)

Proteinca는 대량의 PPI 데이터를 그래프 기

반 알고리즘을 이용하여 사용자가 쉽게 분석할 수 있도록 도와주는 워크벤치 시스템이다.

Proteinca의 구조

Proteinca 시스템은 Proteinca와 Proteinca 데이터베이스로 나누어진다. Proteinca는 PPI 데이터를 읽고 분석하며, 가시화하는 등의 기능을 제공하는 프로그램이며, Proteinca 데이터베이스는 몇 가지 PPI 데이터를 저장하고 있는 데이터베이스이다. Proteinca 데이터베이스는 Proteinca의 데이터베이스 관리 모듈을 통해 접근할 수 있으며, 추가, 수정 및 삭제가 가능하다.

Proteinca는 PPI 데이터를 다른 데이터베이스의 플랫 파일로부터 읽거나 데이터베이스 관리 모듈을 통해 Proteinca 데이터베이스로부터 읽을 수 있으며, 입력된 데이터를 그래프로 변환하여 분석하거나 가시화할 수 있다. Proteinca는 그림 3과 같이 크게 여섯 개의 모듈로 나눌 수 있다.

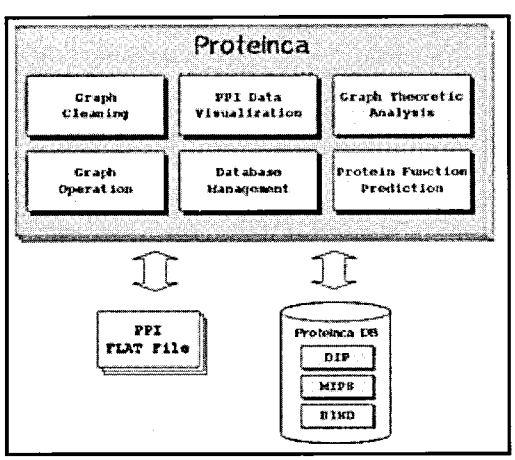


그림3. Proteinca의 구조도 : Proteinca는 기존의 다양한 PPI 데이터베이스를 입력으로 하여, 그래프 이론을 바탕으로 PPI 데이터를 분석할 수 있다.

각 모듈의 기능은 다음과 같다.

o Graph Cleaning : PPI 데이터를 전처리 과정을 수행하는 모듈이다. PPI 전체 데이터에서 Singleton이나 Loop와 같이 불필요한 데이터를 삭제할 수 있다. Singleton은 상호작용이 없는 단백질을 뜻하며, Loop는 자기자신에게 영향을 미치는 단백질을 뜻한다. 만약, 이러한 정보가 필요한 경우에는 전처리 과정을 생략할 수 있다.

o PPI Data Visualization : PPI 데이터를 가시화하는 모듈이다. LEDA[7]를 이용해 구현하였으며, 다양한 레이아웃을 제공한다.

o Graph Theoretic Analysis : PPI 데이터를 그래프 이론에 기반하여 분석하는 모듈이다. Shortest Path, Diameter등의 알고리즘을 제공하다.

o Graph Operation : PPI 데이터들을 연산하는 모듈이다. Union, Intersection, Difference등의 기능을 제공한다.

o Database Management : Proteinca 데이터베이스를 관리하는 모듈이다. 데이터베이스를 읽거나 업데이트하는 기능을 제공한다.

o Protein Function Prediction : PPI 데이터를 이용해 단백질의 기능을 예측하는 모듈이다. Majority[8,9], Chi-Square[10] 방법을 적용하여 단백질의 기능을 예측할 수 있다.

Proteinca의 기능

Proteinca에서는 사용자가 PPI 데이터들을 쉽게 가공하고 분석할 수 있는 다양한 기능을 제공한다.

PPI 데이터베이스 (PPI Database)

Proteinca 시스템은 기존에 존재하는 DIP,

MIPS와 같은 PPI 데이터베이스의 플랫 파일에 기반하여 생성한 Proteinca 데이터베이스라는 자체 데이터베이스를 제공한다. 사용자가 DIP이나 MIPS 등의 PPI 데이터를 사용하기 위해서는 매번 여러 플랫 파일을 다운받아 입력하여야 한다. 하지만 PPI 데이터베이스를 이용하면 플랫 파일 입력 단계를 생략함으로써 보다 편리하게 PPI 데이터를 사용할 수 있다. 또한, 사용자가 분석한 결과를 데이터베이스에 저장할 수 있으므로 연구의 연속성을 보장한다.

PPI 데이터의 입출력 (PPI Data I/O)

PPI 데이터 I/O 기능은 PPI 데이터를 읽고 쓰는 기능이다. Proteinca는 두 가지 방법으로 PPI 데이터를 읽을 수 있다. 하나는 공개된 PPI 데이터베이스에서 제공하는 플랫 파일과 단백질의 기능과 같은 부가적인 플랫 파일을 읽는 방법이다. 두 번째는 Proteinca 데이터베이스를 이용하는 방법이다. Proteinca에서는 DIP, MIPS, BIND의 플랫 파일을 이용하여 제작한 Proteinca 데이터베이스를 제공하는데, 이 데이터베이스를 이용함으로써 PPI 데이터를 안전하고 편리하게 읽을 수 있다.

Proteinca에서 사용할 수 있는 플랫 파일은 DIP과 MIPS, BIND의 플랫 파일이다. 하지만, 일반적인 PPI 데이터도 지원하기 위해 단순히 상호작용하는 두 단백질 정보만으로 이루어진 데이터를 읽는 기능도 제공한다. 이외에도, Proteinca의 플랫 파일 포맷인 PIG 포맷으로 PPI 데이터를 읽고 저장할 수 있다. PIG 포맷은 크게 세 부분으로 구성되며, 첫 번째 부분은 노드와 에지 수 등 그래프의 기본적인 특징에 관한 부분이며, 두 번째는 노드, 세 번째 부분은 에지의 속성에 관하여 설명된 부분이다. 이러한 PIG 포맷은 노드의 위치와 색깔, 에지의 색깔에 관한 정보도 함께 가지고 있는 Embeded 버전과 이러한 정보가 없는 Normal 버전이 있다.

입력 데이터 필터링(Input Data Filtering)

데이터베이스나 플랫 파일로부터 입력 받은 PPI 데이터에는 그래프를 분석하는데 불필요한 정보도 포함되어 있다. 따라서, 분석에 앞서 불필요한 요소를 제거하는 전처리 과정이 필요하며, Proteinca 시스템에서는 이를 위해 Graph Cleaning 기능을 제공한다. 그래프를 분석하는데 불필요한 데이터로는 Loop, Singleton, Doubleton, Parallel Edge등이 있다. 시작 노드와 끝 노드가 같은 노드인 에지를 Loop, 시작 노드와 끝 노드가 동일한 두 개 이상의 에지들을 Parallel Edge라고 한다. Singleton은 다른 노드와의 연결이 없는 하나의 노드, Doubleton은 다른 노드와의 연결은 없지만, 서로 연결된 두 노드를 뜻하며, Proteinca는 이들을 삭제하는 기능을 제공한다.

분석 기능(Analysis Property)

PPI 데이터는 그래프 데이터로 추상화할 수 있으므로, 그래프 이론을 통해 분석할 수 있다. Proteinca는 컴포넌트의 수나 Degree등의 간단한 그래프 정보에서부터, Shortest Path, Diameter, Cut Vertex등의 그래프 알고리즘을 이용하여 분석하는 기능도 제공한다. Shortest Path는 주어진 노드 사이의 최단 거리를 구하는 알고리즘이며, PPI 데이터의 모든 에지는 가중치가 '1'이라고 가정한다. Diameter는 모든 Shortest Path 중에 가장 긴 것을 의미하며, Cut Vertex는 그 노드를 제거하면 하나의 컴포넌트가 두 개의 컴포

넌트로 나누어지는 노드를 말한다. Proteinca에서 지원하는 그래프 알고리즘은 LEDA 라이브러리를 이용하여 구현하였다.

PPI 데이터는 방대한 정보를 담고 있으므로, 전체 그래프를 가지고 작업하면 많은 시간이 걸리고, 가시화한 결과를 이해하기 힘들다. 따라서, 사용자가 관심이 있는 일부 단백질 데이터만을 추출하는 기능이 필요하다. Proteinca에서는 다양한 사용자의 요구를 만족할 수 있는 검색 및 추출 기능을 제공한다. 이 기능을 이용하면, 가장 큰 컴포넌트를 추출하거나 특정 단백질(노드)과 Shortest Path로부터 일정한 거리까지 상호작용하는 단백질들의 그래프, 특정 기능을 가진 단백질들의 그래프와 같은 서브 그래프를 추출할 수 있다.

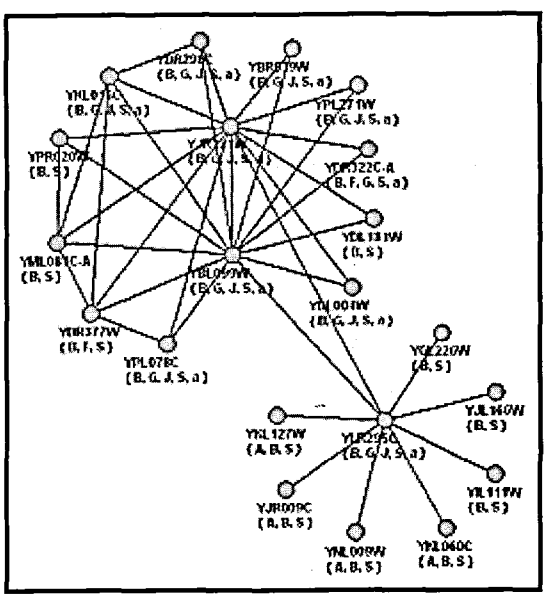
데이터 통합 기능(Data Integrating Property)

연구자는 PPI 데이터를 분석할 때는 DIP, MIPS, BIND와 같은 PPI 데이터베이스나 연구자의 개인 데이터와 같이 다양한 데이터를 사용할 수 있다. 또한, DIP 데이터와 BIND 데이터를 함께 사용할 수도 있으며, 특정 자료의 일부만을 분석하고자 할 수 있다. 따라서 다양한 사용자의 요구를 만족하기 위하여, Proteinca 시스템에서는 그래프 연산기능을 제공한다.

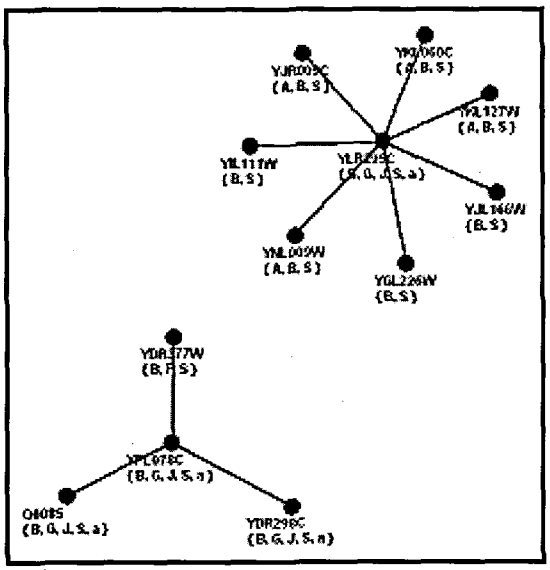
Proteinca에서 제공하는 그래프 연산 기능은 union과 intersection, difference등이 있다. 이러한 그래프 연산 기능을 적용하기 위해서는 사용되는 PPI 데이터들 간의 단백질 ID가 동일해야 한다는 제약이 있으며, 일반적인 경우 ORF ID를 사용하여 연산한다.

그림 4는 Proteinca를 이용하여 두 PPI 데이터를 union 한 결과이다. (a)는 DIP의 데이터 중 Energy에 관련된 기능을 하는 21개의 단백질로 구성된 하나의 컴포넌트이며, (b)는 MIPS의 데이터 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. 따라서 여러 개의 PPI 데이터를 함께 사용하면, 개개의 PPI 데이터에서는 나타나지 않던 정보가 추가되거나 존재했던 정보들이 사라질 수 있다.

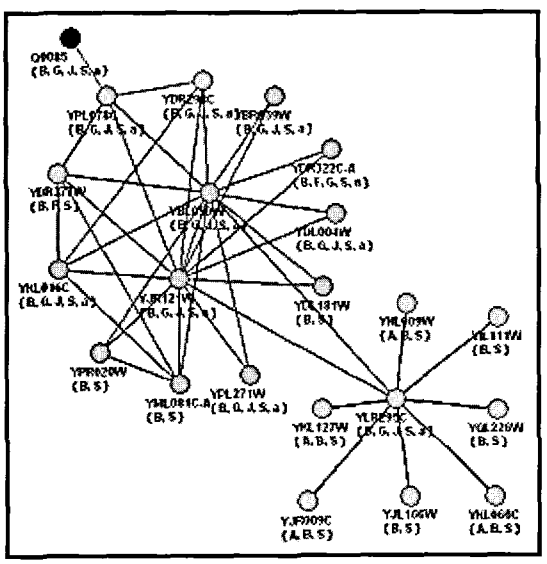
Von Mering C et al.[5]와 같이 보다 신뢰성 있는 PPI 데이터를 분석하고자 한다면, Proteinca의 Intersection 기능을 사용하여 여러 PPI 데이터에 공통적으로 존재하는 데이터들을 추출하여 생성한 PPI 데이터를 분석하면 된다. 반대로, 보다 많은 정보를 원한다면 Proteinca의 union 기능을 사용하여 생성한 PPI 데이터를 분석하면 된다. 이외에도 union 기능을 이용하면 개개의 실험실에서 실험한 PPI 데이터를 기존의 PPI 데이터에 추가하여 분석할 수 있다.



(a) DIP의 일부 PPI 데이터



(b) MIPS의 일부 PPI 데이터



(c) (a)의 DIP 데이터와 (b)의 MIPS 데이터의 union 결과

그림 4. Proteinca에서의 union 기능 : (a) DIP 데이터 중 Energy에 관련된 기능을 하는 21개의 단백질로 구성된 컴포넌트, (b) MIPS 데이터 중 Energy에 관련된 기능을 하는 12개의 단백질로 구성된 두 개의 컴포넌트, (c) (a)와 (b)의 합집합 : 하늘색 노드는 DIP과 MIPS의 공통적으로 존재하는 단백질을 뜻하며, 주황색은 DIP, 갈색은 MIPS에만 존재하는 단백질을 뜻한다. 마찬가지로, 파란색 에지는 DIP과 MIPS에 공통적으로 존재하는 상호작용을 뜻하며, 붉은 색은 DIP, 검은색은 MIPS에만 존재하는 상호작용을 뜻한다.

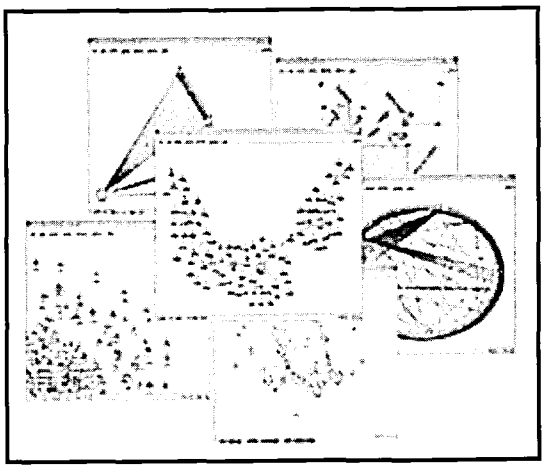
Proteinca 시스템에서 제공하는 그래프 연산은 다음과 같다.

- o Extract Largest Component : 그래프에서 가장 큰 컴포넌트를 추출하여, 새로운 그래프를 생성한다.
- o Union : 두 그래프를 합하여 새로운 그래프를 생성한다.
- o Union To Embedded Graph : 합하는 두 그래프의 공통 노드, 개별 노드들의 색깔을 달리하여, 두 그래프를 union한다.
- o Intersection : 두 그래프의 교집합으로 새로운 그래프를 생성한다.
- o Difference : 두 그래프의 차집합으로 새로운 그래프를 생성한다.

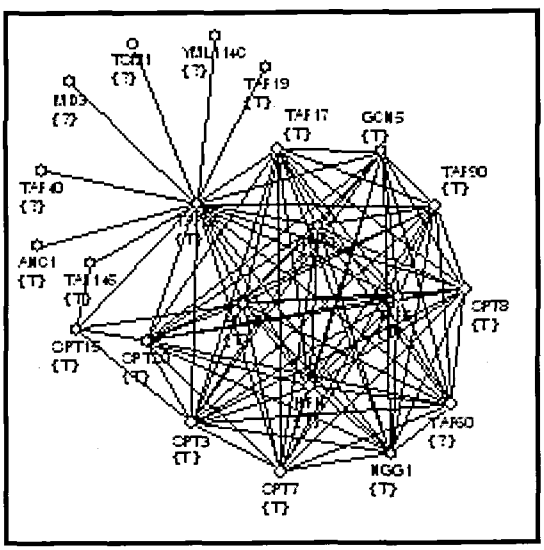
가시화 기능(Visualization Property)

Proteinca의 중요한 기능 중 하나는 PPI 데이터를 가시화하는 것이다. Proteinca는 입력된 PPI 데이터를 사용자가 인지하기 쉽도록 그래프의 특성에 맞는 다양한 레이아웃 방법을 이용하여 가시화할 수 있다. 뿐만 아니라 색깔과 굵기, 레이블을 달리하여 시각적 효과를 줄 수도 있고, 줌 기능도 제공한다. 그리고, 가시화한 화면에서 직접 노드와 에지를 추가하거나 삭제할 수 있는 Visual Editing 기능도 제공한다. 이 기능을 이용하면, 사용자가 실험한 데이터를 기존의 PPI 데이터에 쉽게 추가할 수 있으며, 잘못된 데이터를 삭제할 수도 있다. 그림5는 MIPS

에서 제공하는 데이터 중 문헌에서 발췌한 데이터를 가시화한 예이다.



(a) 다양한 레이아웃



(b) 특정 컴포넌트의 가시화

그림5. Proteinca 가시화 예, (a) 다양한 레이아웃(layout) 지원: Proteinca는 사용자가 인지하기 쉽도록, 다양한 레이아웃을 지원한다. 지원하는 레이아웃은 2D and 3D spring embedder, circular, random, straight-line, orthogonal등이 있다. (b) 컴포넌트 추출 : 파란색으로 표현된 단백질들은 MIPS에 존재하는 단백질들로, 모두 Mitochondrial Transcription에 위치하며, HCS(Highly Connected Subgraph)를 구조를 이루고 있다.

PPI 데이터는 일반적인 그래프와는 달리 가시화하기 어려운 몇 가지 특징이 있다.

o PPI 데이터에 포함된 데이터의 양은 아주 방대하므로, 이를 가시화할 때 많은 시간이 소요된다. 또한 표현해야 할 단백질(노드)의 수가 많아, 이들의 레이블을 표현하기가 어렵다.

o PPI 데이터는 에지 교차가 많은 그래프이다. 하나의 단백질은 여러 단백질들과 상호작용을 하므로 이를 그래프로 표현할 경우, 에지의 교차가 많이 발생할 수 밖에 없다.

이러한 특징으로 인하여, PPI 데이터를 가시화하면 사용자가 인지하기 어려운 복잡한 그래프가 된다. 따라서, 그래프를 간소화하여 사용자가 이해하기 쉽도록 가시화하는 방법이 필요하다. Nizar et al.은 그래프의 클러스터링 정보를 바탕으로 그래프를 간소화하는 방법을 제안하였다[11]. 하지만, 그래프를 실시간으로 클러스터링하는 것은 불가능하므로, 미리 계산해놓은 데이터를 이용해야 한다는 문제점이 있다. 따라서, 그래프의 내부 정보만으로 그래프를 간소화시키는 방법이 필요하다. 본 시스템에서는 그래프의 내부적인 특징을 이용하여 그래프 간소화 기능을 제공한다. Proteinca에서 제공하는 간소화 방법은 다음과 같다.

1. 그래프의 모든 단백질(노드)에 대해 상호작용하는 단백질 수에 비례하는

weight를 계산하고, 모든 단백질 쌍에 대해 weight에 비례하고 거리에 반비례하는 공식을 적용하여 gravity를 구한다.

2. 전체 PPI 데이터에서 일정 크기 이상의 gravity를 가지는 노드들을 하나의 노드로 합한다.
3. 더 이상 간소화되지 않을 때까지 1, 2 과정을 반복한다. 이때, 두 단백질 사이의 거리는 원본 그래프의 거리로 계산한다.
4. 최종적으로 얻어지는 그래프는 상호작용이 많은 노드들, 즉 허브들간의 연결 구조로 간소화된 그래프이다.

사용자는 이 그래프를 살펴봄으로써 그래프의 대략적인 구조를 파악할 수 있다.

단백질 기능 예측(Protein Function Prediction)

최근 PPI 데이터를 이용한 연구 중 각광받고 있는 분야는 미지의 단백질의 기능을 예측하는 분야이다. 단백질을 응용하기 위해서는 단백질의 기능 정보가 필요하다. 하지만, 단백질의 기능을 알기 위한 실험에는 많은 시간과 비용이 소요된다. 따라서 가능성이 높은 기능을 예측해주는 방법이 필요하다. Proteinca는 기존에 연구된 방법 중 Majority Rule[8,9]과 Chi-Square[10] 방법을 구현하여 단백질 예측 기능을 제공한다.

Majority Rule는 PPI 데이터를 이용한 단백질 기능 예측 방법 중 가장 먼저 연구된 방법으로, 규칙이 간단하고 구현하기도 쉽다. 이 방법은 서로 상호작용하는 단백질들은 같은 기능을 할 가능성이 높다는 가정을 기반으로 한다. 따라서 Majority Rule를 이용하여 미지의 단백질의 기능을 할당하기 위해서는, 그 단백질과 상호작용하는 단백질들의 기능들에서 출현 빈도를 계산하고. 출현 빈도 값에 미리 정의해 놓은 가중치를 부여하여 각 기능에 대한 점수를 부여한다. 해당 단백질의 기능은 각 기능들 중에서 점수가 가장 높은 몇 가지의 기능으로 할당된다.

Chi-Square 방법은 Haretsugu H et al.[10]이 제안한 방법으로 Majority Rule과 유사하지만, 인접한 기능의 빈도수뿐만 아니라 전체적인 빈도수도 같이 고려한다는 차이점이 있다. 이 방법 역시 계산된 점수가 가장 높은 기능을 할당한다.

Conclusion

단백질-단백질 상호작용(PPI : Protein-Protein Interaction) 데이터를 분석함으로써, 미지의 단백질의 기능을 예측하고, 중요한 단백질을 찾을 수 있다. 본 연구에서는 현재 국내외에 개발되어 있는 다양한 단백질 상호작용 데이터를 종합적으로 통합하여, 이들 데이터로부터 단백질에 대한 새로운 지식을 얻고자 하였다. 이를 위해서 사용자가 다양한 데이터베이스의 데이터들을 쉽게 가공하고 분석할 수 있도록 Proteinca 시스템을 구현하였다.

Proteinca는 다음과 같은 기능을 제공한다.

- o 다양한 PPI 데이터베이스를 구축하였다.
- o 개개의 실험실에서 생성한 단백질 상호작용 데이터를 Proteinca의 PPI 데이터베이스에 추가할 수 있다.
- o 그래프 이론을 바탕으로 한 다양한 PPI 데이터 분석 기능을 제공한다.
- o Proteinca에서는 2D and 3D spring embedder, circular, random, straight-line, orthogonal등 다

양한 레이아웃을 이용한 가시화 방법을 제공한다.

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

References

[1] Lukasz Salwinski et al., The Database of Interacting Proteins: 2004 update, Nucleic Acids Research, Vol. 32, Database issue, 449-451, 2004

[2] H. W. Mewes et al., MIPS : a database for genome and protein sequences, Nucleic Acids Research, Vol. 28, No. 1, 37-40, 2000

[3] Gary D. Bader et al. BIND: the Biomolecular Interaction Network Database, Nucleic Acids Research, Vol. 31, No. 1, 248-250, 2003

[4] Christian von Mering et al., STRING: a database of predicted functional associations between proteins, Nucleic Acids Research, Vol. 31, No. 1, 258-261, 2003

[5] Von Mering C et al., Comparative assessment of large-scale data sets of protein-protein interactions, Nature, Vol. 417, 2002

[6] Hodges PE et al., The Yeast Proteome Database(YPD) : a model for the organization and presentation of genome-wide functional data, Vol. 27, No. 1, 1999

[7] Kurt Mehlhorn et al., LEDA : a platform for combinatorial and geometric computing, Vol. 38, No. 1, 1995

[8] Schwikowski et al., A network of protein-protein interactions in yeast, Nature Biotechnology, Vol. 18, No. 12, 2000

[9] Vazquez A et al., Global protein function prediction from protein-protein interaction networks, Nature Biotechnology, Vol.21, 2003

[10] Hishigaki H et al., Assessment of prediction accuracy of protein function from protein-protein interaction data, Yeast, Vol. 18, 2001

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