LEDA를 이용한 단백질 상호작용의 분석과 가시화

윤지현° 조환규

부산대학교 컴퓨터공학과, Algorigene Lab.

{jhyoon°, adagio}@pearl.cs.pusan.ac.kr

Analysis and Visualization for Protein-Protein Interaction Using LEDA

Ji-Hyun Yoon° Hwan-Gue Cho

Algorigene Lab., Dept. of Computer Engineering, Pusan National University

요 약

PPI(Protein-Protein Interaction) 데이터는 생물체 내에서 서로 상호작용하는 단백질(protein)들에 대한 정보이다. 단백질 상호작용은 실제 생체 내에서 어떠한 작용이 일어나게 하는 원인이므로, 많은 생물학자들이 관심을 가지고 연구하고 있으며, 그 결과로 몇몇 데이터베이스가 만들어졌다. 이런 데이터베이스들은 다른 연구자들을 위해 데이터를 공개하고 있지만, 대부분의 데이터베이스가 탭으로 분리된 텍스트 형태로 제공한다. 하지만, 텍스트 형태의 데이터는 사람이 직관적으로 인지할 수 없기 때문에, PPI 데이터를 분석하기 쉬운 그래프 형태로 보여주는 프로그램이 필요하다. 그리고 아직 기능을 모르는 단백질이 많으므로 이를 예측하는 프로그램도 필요하다. 본 논문에서는 LEDA를 이용하여 PPI 데이터를 그래프 형태로 표현하며, 이 그래프에 그래프 이론을 적용하여 단백질의 기능을 예측하는 프로그램인 Proteinca에 대해 서술한다.

1. 서 론

단백질은 생물체를 구성하는 중요한 성분인 동시에, 생물이 살아가는데 필수 요소인 효소, 항체, 호르몬등으로 작용한다. 단백질들은 서로 상호작용을 하면서 이런 역할을 수행하는데, 상호작용하는 단백질들의 정보를 PPI(Protein Protein Interaction) 데이터라고 한다. 이런 정보는 DIP(Database of Interacting Proteins)[1], MIPS(Munich Information Center for Protein Sequences)[2], BIND(Biomolecular Interaction Network Database)[3]등 여러 데이터베이스에서 구할 수 있다.

PPI 데이터는 생물체가 어떤 메커니즘으로 살아가는지에 대한 정보를 담고 있으므로, 생물의 대사작용(metabolism) 연구, 신약 개발 등의 분야에서 중요하게 사용된다. 하지만, 대부분의 데이터베이스는 텍스트 형태로 정보를 제공하므로, 이 데이터를 그래프의 형태로 바꾸어주는 프로그램이 필요하다.

PPI 데이터에는 아직 기능을 모르는 단백질들이 많다. 이런 단백질의 기능을 밝히기위해서는 여러가지 생물학적 실험을 해야하는데, 시간이 오래걸리고 비용이 많이 든다. 따라서, 최소한의 실험으로 많은 단백질의 기능을 밝히기 위해, 관계가 있을 법한 단백질들을 예측하는 방법이 필요하다.

Proteinca(Protein Interaction Cabaret)는 여러 데이터베이스의 PPI 데이터를 그래프로 가시화하여 사용자가 PPI 데이터를 보다 쉽게 분석할 수 있도록 도와주며, 그래프 이론을 통해서 단백질의 기능을 예측하는 프로그램이다[4].

2장에서는 PPI 데이터와 Proteinca에서 사용한 라이브러리인 LEDA에 대해 설명하고, 3장에서는 Proteinca의 기능 및 구현 환경에 대해 설명한다. 끝으로 4장에서는 결론과 향후 과제에 대해 논한다.

2. 관련 연구

PPI 데이터는 생물체 내에서 어떤 단백질들이 상호작용하는 지에 대한 정보를 제공하므로, 생물학 분야에서 중요한 데이터이다. 하지만, 텍스트 형태로 정보를 제공하기 때문에, 직관적으로 분석 가능한 그래프로 가시화하는 프로그램이 필요하다. LEDA는 그래프를 위한 자료구조와 알고리즘, 가시화하는 기능을 가진 라이브러리로써, PPI 데이터를 다루는데 적합하다.

2.1 PPI Data

텍스트 형태의 PPI 데이터는, 한 행마다 상호작용하는 두 단백질과 단백질 혹은 상호작용에 대한 정보가 명시되어 있다. 대표적인 정보로는 단백질의 데이터베이스 ID, 상호작용을 밝혀낸 실험 방법, 종에 대한 정보 등이다.

PPI 데이터는 단백질을 노드로, 두 단백질 간의 상호작용을 에지로 가지는 그래프 데이터이다. 따라서 PPI 데이

터는 그래프로 가시화할 수 있으며, 그래프 이론을 통해서 분석 가능하다. PPI 데이터를 가시화한 대표적인 프로그램으로 InterViewer를 들 수 있다[5].

2.2 LEDA

LEDA(Library of Efficient Data Types and Algorithms)는 독일의 Algorithmic Solutions Software사에서 개발한 C++ 라이브러리로써, 기본적인 자료구조는 물론, 그래프와 기하학에 관련된 자료구조와 알고리즘, GUI 기능을 제공한다[6]. LEDA는 전세계적으로 많은 분야에서 사용되고 있으며, 대표적인 응용 분야로 네트워크이나 VLSI 디자인, 스케줄링, 생물정보학, GIS등을 들 수 있다.

Proteinca는 PPI 데이터 가시화와 단백질 기능 예측을 위해 그래프 관련 자료구조와 알고리즘, GUI 기능을 이용한다.

3. Proteinca의 기능

그림 1은 본 연구에서 구현한 Proteinca는 구조도이다.

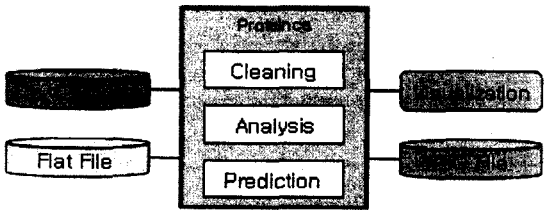

그림 1. Proteinca의 구조도. Proteinca는 DB나 파일로부터 PPI 데이터를 로드하여, 데이터를 분석하거나 예측하고, 그 결과를 가시화하거나 자체 파일 포맷으로 저장할 수 있다.

Proteinca의 궁극적인 목표는 주어진 PPI 데이터를 그래프로 가시화하는 것과, Graph Theory를 이용하여 단백질의 기능을 예측하는 것이다. PPI 데이터는 파일과 본 연구에서 구축한 데이터베이스를 통해 입력받을 수 있다. Proteinca에서 사용하는 PPI 데이터베이스는 DIP, MIPS, 그리고 BIND이다.

Proteinca에 PPI 데이터를 입력하면, 메인 화면에 그래프의 특성을 분석하여 보여준다. 그래프의 기본적인 분석이 끝나면 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.

Proteinca는 MFC를 이용하여 구현되었으며, 가시화를 비롯한 그래프 관련 기능은 LEDA를 이용하였다. Visual C++ 7.1을 이용하여 컴파일하였으며, Intel Pentium 4, Windows XP SP1a 환경에서 테스트하였다. Proteinca가 제공하는 기능은 다음과 같다.

① 그래프 가시화 기능
② 그래프 연산 기능
③ 데이터베이스 기능
④ 단백질의 기능 예측
⑤ 기타

3.1 가시화 기능

Protienca의 중요한 기능은 PPI 데이터를 가시화하는 것이다. PPI 데이터는 복잡한 그래프이므로 알아보기 쉽도록, Straight-Line Spring 알고리즘으로 노드를 배치한다. 단백질의 차수는 단백질의 중요도를 가늠하는 요소이므로, 차수에 따라 노드의 크기를 다르게 표시한다. 그리고 Proteinca는 단백질마다 단백질의 기능, 단백질이 존재하는 종, ORF, 데이터베이스 ID, OMIM 등의 정보를 저장하고 있으므로, 사용자가 원하는 정보을 선택하여 보여줄 수 있다. 이외에도 사용자가 원하는 속성을 가진 노드만 보여주는 필터링 기능과 특정 노드와 패스를 추출하여 보여주는 기능도 제공한다. 그림 2는 DIP의 PPI 데이터 중 일부를 가시화한 그래프이고, 그림 3은 문헌에서 검색한 PPI 데이터를 가시화한 그래프 중 일부를 캡쳐한 그림이다.

Proteinca는 가시화된 그래프상에서 노드나 에지를 추가 또는 제거하는 기능을 제공한다. 이 기능은 실험자가 실험한 데이터를 그래프에 첨가하거나 수정할 때 유용하게 사용될 수 있다. 뿐만 아니라 기존의 그래프의 데이터를 수정하거나 좌표 상의 위치, 색깔, 크기 등도 변경할 수 있다.

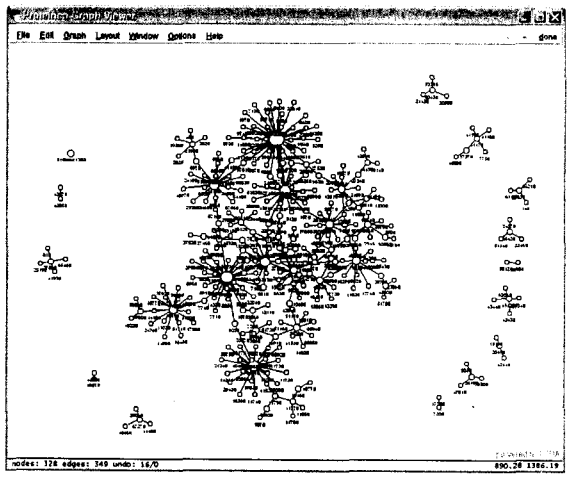

그림 2. DIP의 PPI 데이터 중 일부를 Proteinca로 가시화한 그래프. DIP은 실험으로 밝혀진 단백질 상호작용 데이터를 모아놓은 데이터베이스이다.

3.2 그래프 연산 기능

PPI 데이터를 분석할 때, 서로 다른 그래프 간의 교집합이나 합집합에 대한 정보가 필요할 수 있다. 이를 위해 Proteinca는 두 그래프의 합집합, 교집합, 차집합을 구하거나

제일 큰 컴포넌트를 추출할 수 있는 그래프 연산 기능을 제공한다. 이 기능을 이용하면 여러 PPI 데이터 베이스의 데이터를 합쳐서 보다 많은 단백질 상호작용을 볼 수 있고, 교집합하여 정확도가 높은 그래프만 선택해 볼 수 도 있다.

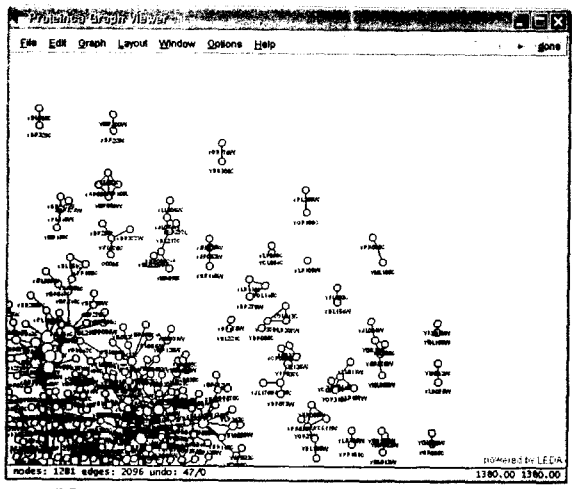

그림 3. 문헌에서 검색한 PPI 데이터를 가시화한 그래프 중 일부를 캡쳐한 그림

3.3 데이터베이스 기능

실험을 통해 밝혀진 PPI 데이터는 여러 데이터베이스에서 쉽게 구할 수 있다. 하지만 데이터베이스마다 다른 ID로 단백질을 나타내며, 포함되어 있는 상호작용도 다를 뿐만 아니라 제공하는 정보도 제각각이다. 그리고 단백질의 기능이나 질병 정보는 다른 데이터베이스에서 제공한다. 따라서 이런 데이터들을 통합할 수 있는 방법이 필요하며, 본 연구에서는 여러 데이터베이스에서 구한 자료를 이용하여 PPI 데이터와 단백질의 정보를 저장하고 있는 자체 데이터베이스를 구축하였다.

데이터베이스에는 그래프 정보와 단백질 정보가 저장되어 있다. 그래프 정보에는 DIP과 MIPS, BIND등의 PPI 데이터가 포함되어 있으며, 단백질 정보에는 GenBank ID, ORF, 단백질의 기능, EC Number, OMIM 데이터 등이 포함되어 있다. Proteinca는 이 데이터베이스에 접근하여 PPI 그래프나 단백질 정보를 사용자에게 보여줄 수 있을 뿐만 아니라, 다른 데이터베이스로부터 다운받은 데이터를 이용하여 업데이트할 수도 있고, SQL 명령을 이용하여 데이터베이스에 쿼리를 보내 그 결과를 볼 수도 있다.

3.4 단백질 기능 예측 및 부가적인 기능

Proteinca는 단백질의 기능을 예측하기 위해서 간단한 다수결법(majority rule)을 이용한다[7]. 다수결법은 상호작용하는 단백질들은 같은 기능을 가질 것이라는 가정을 기반으로 하며, 단백질의 기능을 예측하는 방법이다. 예측하는 방법은 기능이 밝혀지지 않은 단백질과 이웃하고 있는 단백질의 기능들을 통계적으로 분석하여, 상대적으로 출현 빈도가 높은 기능을 해당 단백질에게 할당한다. 다른 알고리즘으로는 이웃하고 있지않은 단백질도 거리에 따라 가중치를 달리하여 통계적 분석에 참여시키는 방법이 있다. 이런 방법들은 기능을 예측할 때 뿐만 아니라, 질병에 관련된 단백질을 찾을 때도 적용할 수 있다.

이외에도 Proteinca는 사용자가 수정한 그래프를 그대로 재현할 수 있도록 임베디드 그래프 기능을 제공한다. 그리고 두 점 사이의 최단거리나 모든 최단거리 중 가장 긴 최단거리인 지름을 구하는 기능, 컴포넌트를 둘로 나누는 분절점(Cut Vertex)을 구하는 기능 등 그래프 이론에 관련된 정보를 구하는 기능도 제공한다.

4. 결론 및 향후 연구

본 논문에서는 PPI 데이터를 그래프로 가시화하여, 사용자가 단백질의 상호작용을 분석하는데 도움을 주고, 다수결법을 이용하여 아직 기능이 밝혀지지 않은 단백질의 기능을 예측하는 Proteinca에 대해 설명하였다. Proteinca는 사용자의 선택에 따라 단백질의 정보와 그래프를 바꾸어 볼 수 있는 기능과, 다양한 단백질 정보를 저장하고 있는 데이터베이스를 제공하므로써 풍부한 정보와 편의를 제공한다.

향후 연구로는 인터넷에 산재해있는 데이터베이스의 PPI Data를 데이터베이스에 추가하고 가시화하는 방법과 단백질의 세포내 위치에 따른 가시화 기능에 대한 연구가 필요하다. 그리고, 보다 정확한 예측을 위해 그래프 이론에 기반한 기능 예측 알고리즘이 필요하다.

참고 문헌

[1] Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu

[2] Munich Information Center for Protein Sequences, http://mips.gsf.de

[3] Biomolecular Interaction Network Database, http://www.blueprint.org/bind/bind.php

[4] Proteinca, http://garnet.cs.pusan.ac.kr/~proten

[5] Kyung-Sook Han, Byong-Hyon Ju, Jong H. Park, "InterViewer : Dynamic Visualization of Protein-Protein Interactions", Lecture Notes in Computer Science, Vol. 2528, 364-365, 2002

[6] Algorithmic Solutions Software GmbH, http://www.algorithmic-solutions.com

[7] Benno Schwikowski, Peter Uetz, Stanley Fields, "A Network of Protein-Protein Interaction in Yeast", Nature Biotechnology, Vol. 18, 1257-1261, 2000