단백질-단백질 상호작용 경로 분석 알고리즘의 설계 및 구현*

Design and Implementation of the Protein to Protein Interaction Pathway Analysis Algorithms

이재권, 강태호, 이영훈, 유재수
충북대학교 정보통신공학과

Lee Jae-Kwon, Kang Tae-Ho, Lee Young-Hoon, Yoo Jae-Soo
Dept. of Computer and Communication Eng., Chungbuk National University

요약

Post-genome 시대에는 유전체뿐만 아니라 단백질에 대한 연구의 필요성이 증대되고 있다. 특히 단백질-단백질 상호작용 및 단백질 네트워크에 대한 연구를 기반으로 전체 생물 시스템을 분석하는 연구가 중요한 이슈로 떠오르고 있다. 기존에 생물학자들이 실험을 통해서 증명한 사실들을 논문이나 기타 매체를 통해서 공개를 하고 있다. 하지만 공개된 정보의 양이 방대하므로 생물학자들이 정보를 효율적으로 이용하지 못하는 경우가 많다. 인터넷의 발달로 하루에도 수 없이 쏟아져 나오는 연구 성과들에 쉽게 접근이 가능해졌다. 이러한 매체로부터 생물학적 의미를 가지는 정보를 효과적으로 추출하는 일이 중요하게 대두되었다. 따라서 본 연구에서는 인터넷상에 공개된 다량의 논문 및 기타 정보 매체로부터 단백질-단백질 상호작용 정보를 추출한 데이터베이스로부터 단백질의 네트워크를 구성하고 단백질 네트워크를 통해서 생물학적 의미를 가지는 여러 가지 경로 분석 알고리즘을 설계하고 구현한다.

Abstract

In the post-genomic era, researches on proteins as well as genes have been increasingly required. Particularly, work on protein-protein interaction and protein network construction have been recently establishing. Most biologists publish their research results through papers or other media. However, biologists do not use the information effectively, since the published research results are very large. As the growth of internet, it becomes easy to access very large research results. It is significantly important to extract information with a biological meaning from varisous media. Therefore, in this research, we efficiently extract protein-protein interaction information from many open papers or other media and construct the database of the extracted information. We build a protein network from the established database and then design and implement various pathway analysis algorithms which find biological meaning from the protein network.

* 본 연구는 산업자원부의 지역혁신 인력양성사업의 연구결과로 수행되었음.

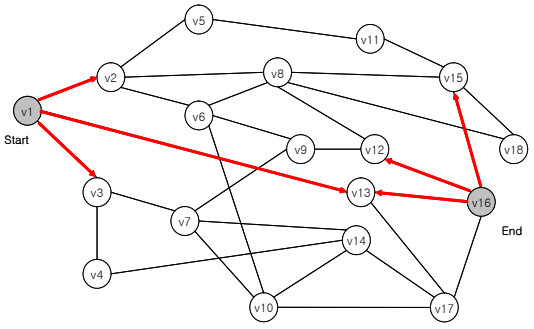
Ⅰ. 서 론

최근 초고속 인터넷의 보편화로 인해 거의 대부분의 연구 분야에서 발표하는 연구 실적들이 인터넷을 통해서 접근이 가능해졌다. 생물학자들의 연구 결과도 마찬가지로 인터넷을 통해서 접근이 가능하다. 따라서 많은 양의 연구 결과물들을 통해서 단백질-단백질 상호작용정보를 표현하는 네트워크를 구성하기가 쉬워졌다. 이러한 연구 환경의 변화로 기존의 생물학자들이 실험을 통해서만 얻을 수 있었던 사실들을 미리 가상실험을 통해서 가설을 세우고 예비 검증을 하거나 구축된 네트워크를 통해서 여러 가지 생물학적 의미를 찾을 수 있을 것이다. 또한 중요한 허브 단백질 등을 찾아내어 신약개발에도 도움이 될 수 있을 것이다. 최단 (shortest) 경로 알고리즘을 통하여 시작 단백질로부터 목표 단백질까지의 최단거리 경로를 얻어 생체에서 진행되고 있는 경로를 예측하고 실험에 적용할 수 있는 실마리를 제공할 수 있을 것이다. 기능 경로(functional pathway)를 바탕으로 하는 알고리즘을 이용하여 같은 기능을 수행하는 단백질의 군 (cluster)을 발굴할 수 있으며 나아가서 기능이 알려져 있지 않은 일련의 단백질 군의 기능을 유추할 수 도 있다. 이는 궁극적으로 시행착오를 통해서 실시되는 많은 생화학 및 생명과학 실험을 좀 더 용이하게 수행 할 수 있도록 도울 수 있을 것이다. 실제 생체 실험을 통해 얻은 가중치 (weight value)를 바탕으로 한 알고리즘을 통해 가능한 경로를 분석할 수 있으며, 이를 바탕으로 신호전이 네트워크 및 경로 분석이 가능하게 된다.

본 연구에서는 단백질-단백질 상호작용 정보가 들어 있는 데이터베이스로부터 단백질 네트워크를 구성하고, 구성된 네트워크를 이용하여 생물학적 의미를 찾을 수 있는 경로분석 알고리즘을 구현한다. 본 논문의 2장에서는 경로분석 프로그램에 사용한 여러 가지 알고리즘을 설명하고 3장에서는 경로분석 프로그램의 구조와 기능에 대한 설명을 한다.

Ⅱ. 경로분석 알고리즘

단백질 경로 분석 알고리즘은 생물학적인 의미를 지니는 경로를 찾는 것이 목적이므로 일반적인 검색 알고리즘과는 다른 접근이 필요하다. 경로 비용이 없는 경우의 최단 경로 검색을 효율적으로 수행하기 위한 단순 양방향 검색 알고리즘과 경로 비용이 있는 경우에는 가중치에 따라서 경로를 설정하는 가중치 검색 알고리즘 그리고 단백질의 생화학적 기능 경로를 따라서 검색을 하는 기능 중심 경로 검색 알고리즘으로 나누어 볼 수 있다.



▶▶ 그림 1. 양방향 최단 경로검색

1. 양방향 최단 경로검색 알고리즘

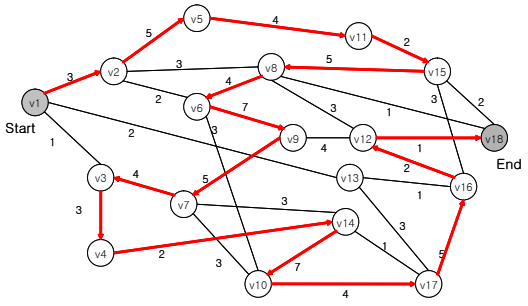
상호작용 하는 두 단백질 사이의 최단경로를 검색한다. 경로비용이 모두 같다는 전제로 이루어지는 검색으로 가장 적은 개수의 노드를 거치는 경로를 탐색하게 된다. 검색 속도 향상을 위해서 양방향으로 검색을 진행한다. [그림 1]에서 v1이 시작 노드가 된다. 그리고 v16을 종료 노드로 보게 되면 단 방향 알고리즘은 시작 노드 v1로부터 시작해서 모든 노드들을 한 번씩 방문하게 된다. 그렇게 되면 시간과 비용이 증가 하므로 양방향으로 검색을 하게 된다. 즉, v1노드에서 인접한 노드 v2, v3, v13을 평가 하고 난 후 똑같이 v16에서도 인접 노드 v12, v13, v15를 평가하게 된다. 이 과정에서 서로 상대방 측에서 접근한 노드를 먼저 발견한 경우 그 노드를 포함한 경로가 최단 경로가 되는 원리 이다. [그림 1]에서는 v13노드를 통

한 경로 v1, v13, v16이 최단 경로로 검색되어 진다.

2. 가중치 중심 경로검색 알고리즘

두 단백질 사이의 상호작용 하는 정도에 따라 부여된 가중치가 높은 노드들을 순회하며 검색한다. 본 연구에서는 가중치의 기준으로 연구 논문에 실험 결과로 자주 등장 하게 되는 빈도수에 따라서 적절한 가중치를 부여 했다. 즉, 많은 연구 논문의 실험 결과로 등장하는 상호작용은 그만큼 신뢰도가 있다는 반증이 되므로 신뢰성에 기반을 둔 경로 검색이 필요할 경우 사용하게 되는 검색 알고리즘이다.

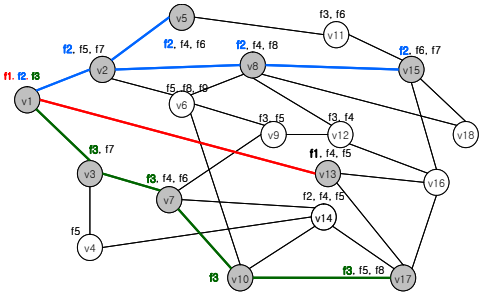
[그림 2]에서와 같이 시작 노드 v1에서는 인접한 노드 v2, v3, v13까지의 간선의 가중치를 모두 판단해본 후 가중치가 큰 v2로의 경로를 설정하게 된다. 같은 방식으로 진행을 하다가 순회가 발생하는 경우에는 이전 단계로 돌아가 다음으로 높은 가중치를 가지고 있는 경로를 설정하고 진행하게 된다.



▶▶ 그림 2. 가중치 중심 경로검색

3. 기능 중심 경로검색 알고리즘

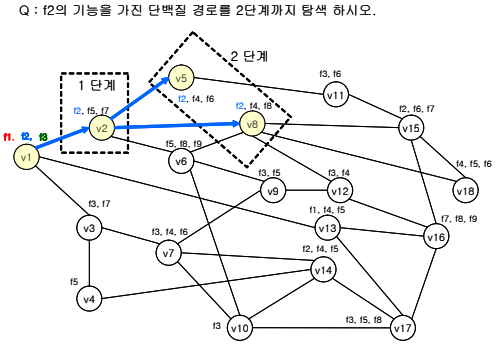
시작점으로 주어진 단백질과 연결된 단백질 중 시작점에 해당하는 단백질의 기능과 일치하는 노드들을 따라서 경로를 검색한다. [그림 3]에서 v1이 시작 노드인 경우에 해당한다. v1은 f1, f2, f3의 기능을 가지고 있으므로 먼저 순서상 검색을 원하는 기능을 선택을 하게 되면 선택된 기능에 대한 경로를 탐색하게 된다. [그림 3]에서와 같이 각각 f1기능을 가진 경로와 f2, f3의 기능 경로를 모두 검색 할 수 있게 된다.



▶▶ 그림 3. 기능 중심 경로검색

4. 단방향 기능 검색 알고리즘

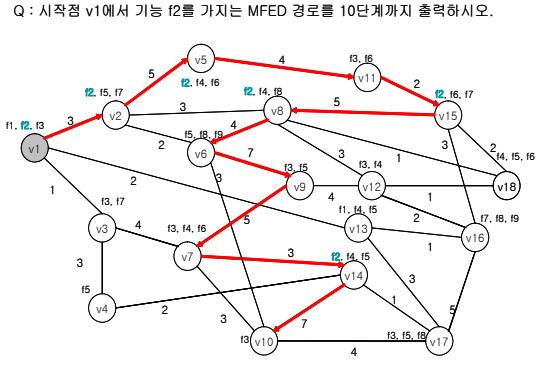
기능 경로 검색과 단 방향 검색 알고리즘을 병행한 알고리즘으로 일정 단계까지의 기능 경로를 볼 수 있는 알고리즘이다. 검색 시에 검색을 원하는 기능과 단계를 명시하게 되면 [그림 4]와 같이 f2의 기능을 가진 단백질 경로를 2단계까지 탐색하는 기능을 지원하게 된다. 먼저 노드 v1에서는 인접 노드 v2, v3, v13이 가지고 있는 기능을 검사한 후에 질의에서 요청한 f2의 기능을 가지고 있는 노드를 경로에 포함시키게 된다. 그리고 노드 v1으로 부터 같은 거리에 존재하는 모든 노드 즉 v2, v3, v13이 모두 평가가 되고 나면 단계 2로 넘어 가게 된다. 단계 2의 모든 노드에서도 단계 1에서의 평가 작업을 반복 하게 된다.



▶▶ 그림 4. 단방향 기능 검색

5. 가중치 기능 고려 검색 알고리즘

단백질 노드의 기능 경로와 가중치 검색 알고리즘이 복합된 형태로 된 알고리즘이다. [그림 5]에서 언급된 질의가 들어오게 되면 먼저 시작노드 v1의 인접 노드 중에 같은 기능을 하는 노드를 선택하게 된다. 그리고 인접 노드 중에 같은 기능을 하는 노드가 여러 개 인 경우 노드와의 간선의 가중치를 비교한 후 가중치가 높은 쪽 경로를 선택하게 된다. 같은 방식으로 경로를 설정하다가 만약 원하는 기능의 노드가 존재 하지 않는 경우에는 간선의 가중치가 높은 경로를 선택하고 진행한다. 종료 조건은 질의에서 명시한 단계에 이르게 되거나 모든 노드를 방문한 경우에 종료하게 된다.

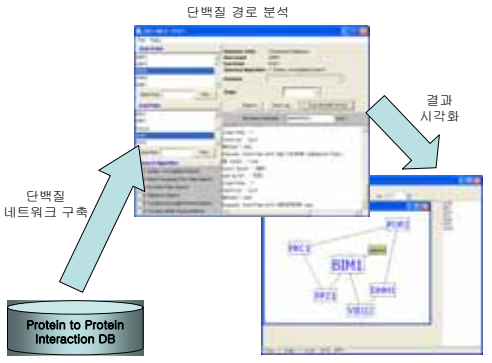


▶▶ 그림 5. 가중치 기능 고려 경로검색

Ⅲ. 단백질 경로분석 프로그램

1. 프로그램 구성

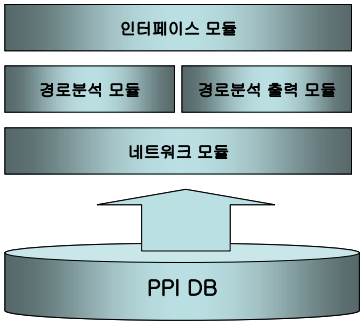
시스템의 구성은 [그림 6]과 같이 PPI(Protein-to-Protein) 데이터베이스, 단백질 경로 분석 프로그램, 결과 출력 프로그램으로 구성이 된다. 텍스트 마이닝 프로그램으로부터 수집한 PPI 데이터베이스에 접속해서 PPI 정보를 네트워크로 구성을 한다. 구성된 네트워크에 다양한 형태의 알고리즘을 적용하여 생물학적 의미를 가지는 경로 혹은 단백질을 검색한다.



▶▶ 그림 6. 경로 분석 프로그램의 구성

2. 프로그램 모듈의 기능

경로 분석 프로그램의 기능은 크게 사용자 인터페이스 모듈, 데이터베이스 모듈, 경로 분석 모듈, 경로 분석 출력 모듈로 구성이 된다. [그림 7]과 같이 PPI DB로부터 네트워크 모듈이 데이터를 읽어와 네트워크를 구축한다.[1] 경로분석 모듈은 네트워크 모듈에 의해서 논리적인 네트워크가 구축이 되면 인터페이스 모듈을 통해서 사용자의 요청을 받게 된다. 사용자의 다양한 요청이 들어오게 되면 인터페이스 모듈은 경로 분석 모듈을 호출하여 네트워크를 적절한 알고리즘을 이용하여 순회하게 된다. 실행결과로 나오는 경로는 네트워크상에 기록이 되고 경로 출력 모듈을 호출하게 되면 적절하게 기록된 경로를 InterViewer[6]와 같은 프로그램을 통해서 보여 줄 수 있는 형태의 파일구조로 출력을 한다.



▶▶ 그림 7. 경로 분석 프로그램 모듈

Ⅳ. 결 론

본 논문에서는 단백질-단백질 네트워크를 구축하고 구축한 단백질 네트워크를 통해서 단백질 경로 분석을 할 수 있는 프로그램을 설계하고 구현하였다. 단백질 경로 분석 프로그램에서는 두 단백질 사이의 최단 경로검색, 상호관계 특성에 따라 부여된 가중치가 높은 경로를 분석, 단백질의 생물학적인 기능에 따른 경로분석 등의 기본적인 분석 알고리즘을 제시하였고, 기본 알고리즘을 복합적으로 적용하여 단계적으로 검색할 수 있도록 하는 복합 알고리즘을 개발하여 제시하였다. 분석된 결과는 사용자가 직관적으로 알 수 있도록 별도의 InterViewer[6]와 같은 GUI 툴과 연동 할 수 있도록 구현했다. 전체 상호작용 관계를 네트워크로 구축함으로 인해 이를 통해 HUB 단백질을 중심으로 클러스터가 형성되는 것을 확인하는 게 가능하며, 다양한 알고리즘을 통한 분석 방법을 제공함으로서 신호전이 경로를 예측할 수 있게 하였다.

향후 연구방향으로는 구축된 네트워크를 통해서 생물학적 의미를 찾을 수 있는 다양한 알고리즘을 더 개발하고 사용자가 직관적으로 파악하고 다양한 질의를 처리할 수 있는 경로 출력 모듈을 개발 하고자 한다.

▌참고문헌▐

[1] Adam Drozdek, Data Structures and Algorithms in Java, BROOKS/COLE, USA, 2001.
[2] Baker, P.G. and Brass, A., Recent development in biological sequence databases, Current Opinion in Biotechnology, 9:54-58, 1998.
[3] Benton, D., Bioinformatics-principles and potential of a new multidisciplinary tool, Trends in Biotechnology, 14:261-272, 1996.
[4] Q. Jacobson, E. Rotenberg, and J. Smith, "Path-based Next Trace Prediction", Proc. 30th. Annual IEEE/ACM Intl. Symp. on Microarchitecture, pp. 14-23, December, 1997.
[5] Sairam Subramanian, Parallel and Dynamic Shortest-Path Algorithms for Sparse Graphs, Department of Computer Science Brown University Providence, Rhode Island 02912, May 1995.
[6] InterViewer, 인하대학교 Web Intelligence 연구실
[7] N. Collier, "Progress on human-computer interaction in the GENIA project on Internet," In Proc. of Natural Language Pacific Rim Symposium, 1999.
[8] K. Fukuda, A. Tamura, T. Tsunoda and T. Takagi, "Toward IE: Identifying protein names from biological papers," In Proc. of the Pacific Symposium on Biocomputing, pp.707~718, 1998.
[9] U. Hahn, M. Romacker, S. Schulz, "Creating knowledge repositories from Biomedical reports: The MEDSYNDIKATE text mining system," International Journal of Medical Informatics, pp. 1~28, 1999.
[10] JBuilder, "http://www.borland.com/jbuilder", Borland Software Corporation.
[11] Human Protein Reference Database, http://hprd.org/, Johns Hopkins University and the Institude of Bioinformatics, 2002-2003.