단백질 경로 분석 시스템의 설계 및 구현

Design and Implementation of Protein Pathway Analysis System

이재권*, 강태호*, 이영훈*, 유재수**
충북대학교 정보통신공학과*, 충북대학교 전기전자컴퓨터공학부**

Jae-Kwon Lee(jklee@netdb.cbnu.ac.kr)*, Tae-Ho Kang(thkang@netdb.cbnu.ac.kr)*,
Young-Hoon Lee(hopekyo@netdb.cbnu.ac.kr)*, Jae-Soo Yoo(yjs@chungbuk.ac.kr)**

요약

포스트 게놈 시대에는 단백질에 대한 연구의 필요성이 증대되고 있다. 특히 단백질-단백질 상호작용 및 단백질 네트워크에 대한 연구를 기반으로 전체 생물 체계를 분석하는 연구가 중요하게 대두되고 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. 다행히도 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 정보를 추출한 데이터베이스로부터 단백질 네트워크를 구성하고, 단백질 네트워크를 통해서 생물학적 의미를 가지는 여러 가지 경로 분석 알고리즘을 설계하고 구현한다.

■ 중심어 : | 단백질 경로 분석 시스템 | 단백질 네트워크 | 생물 정보학 |

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, because the published research results are very large. As the growth of internet field, it becomes easy to access these research results. It is important to extract information with a biological meaning from various media. Therefore, In this paper, we efficiently extract the protein 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.

■ Keyword : | Protein Pathway Analysis System | Protein Network | BioInformatics |

* 이 논문은 2005년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었습니다.

접수번호 : #050511-001
접수일자 : 2005년 05월 11일
심사완료일 : 2005년 09월 15일
교신저자 : 유재수, e-mail : yjs@chungbuk.ac.kr

I. 서 론

포스트 게놈 시대에 생명과학의 발달로 생물학적 정보의 양은 급속도로 증가하게 되었다. 이렇게 급속도로 생산되고 있는 생물학적 정보들을 효율적으로 수집, 관리, 분석 할 수 있도록 하기 위해 전산학적 접근이 요구되고 있다. 또한 DNA-DNA간, DNA-단백질간, 단백질-단백질간의 관계 및 상호작용에 대한 세포 전체의 네트워크를 구성하는 문제는 기존의 수학적 통계학적, 물리학적 접근뿐만 아니라 전산학적 접근도 필요하다.

다행히도 최근 초고속 인터넷의 보편화로 인해 거의 대부분의 연구 분야에서 발표하는 연구 실적들이 인터넷을 통해서 접근이 가능해졌다. 생물학자들의 연구 결과도 마찬가지로 인터넷을 통해서 접근이 가능하다. 대표적인 예로 생화학자들이 발표하는 논문을 통해 인간 단백질의 정보를 데이터베이스화하고 있는 HPRD (Human Protein Reference Database)[1] 그리고 지구상의 생물들의 유전자 서열 정보를 데이터베이스로 구축하는 GOLD(Genomes OnLine Database)[2]와 같은 대량의 데이터베이스가 등장했다. 이러한 많은 양의 연구 결과물들로부터 웹 로봇 그리고 텍스트 마이닝 도구를 이용해서 단백질-단백질 상호작용정보를 표현하는 네트워크를 구성하기가 쉬워졌다. 연구 환경의 변화로 기존의 생물학자들이 실험을 통해서만 얻을 수 있었던 사실들을 거시적 관점에서 미리 가상실험을 통해 가설을 세우고 예비 검증을 하거나 구축된 네트워크를 통해 여러 가지 생물학적 의미를 찾을 수 있다. 또한 중요한 허브 단백질 등을 찾아내어 신약개발에도 도움이 될 수 있다.

최단 경로 알고리즘은 시작 단백질로부터 목표 단백질까지의 최단거리 경로를 얻어 생체에서 진행되고 있는 경로를 예측하고 실험에 적용할 수 있는 실마리를 제공할 수 있다[3]. 기능 경로를 바탕으로 하는 알고리즘은 같은 기능을 수행하는 단백질 군을 발견하여 기능이 알려져 있지 않은 일련의 단백질의 기능을 유추할 수도 있다. 궁극적으로는 시행착오를 통해서 실시되는 많은 생화학 및 생명과학 실험을 좀 더 효율적으로 수행 할 수 있도록 돕는다. 또한 가중치를 고려한 검색은 먼저 실제 실험(In Vivo, In Vitro, Yeast 2 Hybrid)방법에 따라 가중치를 부여하는데 이러한 가중치는 단백질 사이의 상호작용에 대한 정확성 판단의 척도가 되어 보다 유용한 경로 분석을 가능하게 한다.

본 논문에서는 단백질-단백질 상호작용(PPI : Protein to Protein Interaction)정보가 들어 있는 데이터베이스로부터 단백질 네트워크를 구성하고, 구성된 네트워크를 이용하여 생물학적 의미를 찾을 수 있는 경로 분석 시스템을 구현한다. 사용자의 편의성을 위해서 시스템을 쉽게 사용할 수 있는 GUI(Graphic User Interface)를 제공함으로써 생화학자들이 쉽게 사용할 수 있도록 했다.

논문의 구성은 다음과 같다. Ⅱ장에서 생물정보학의 개념과 단백질 네트워크의 개념과 분석 필요성에 대해서 기술한다. Ⅲ장에서는 경로 분석 프로그램의 개요와 구조를 기술하고 프로그램에 사용된 여러 가지 알고리즘을 설명한다. 그리고 구현한 경로 분석 프로그램의 실행 예제와 사용 방법을 기술한다. Ⅳ장에서는 경로 분석 알고리즘을 실제 단백질 데이터와 상호작용 데이터로 구축 된 네트워크를 이용해서 성능을 평가한다. 마지막으로 Ⅴ장에서는 결론 및 향후 연구 계획을 제시한다.

II. 관련 연구

생물정보학[4]의 사전적 의미는 생물학, 전산학, 수학 및 통계학과 정보 기술이 결합된 학문 분야로 정의될 수 있다. 따라서 생물학에 필요한 여러 가지 전산학적, 수학적 접근을 통해서 궁극적으로 인간의 생물학적 원리 및 현상을 밝히는 분야라고 할 수 있다.

생물학자들은 생물학적 현상으로부터 도출된 기존의 많은 실험 데이터들에 접근이 가능하다. 하지만 대량의 정보로부터 생물학자들이 수작업으로 의미 있는 정보를 찾아내는 것은 쉬운 일이 아니다. 이를 위해 생물정보학에서는 기존의 많은 실험 데이터들로부터 생물학적 의미를 얻을 수 있는 모델과 가설을 세워 컴퓨터를 이용해 시뮬레이션을 함으로서 보다 효율을 높일 수 있다[3].

생물정보학에서는 다양한 형태의 연구 결과물을 제공

하는 매체들로부터 생물학적 정보를 추출하여 네트워크를 구성하게 된다. 예를 들면, [그림 1]과 같이 추출된 정보로 구성된 네트워크를 통해서 유전자 발현 조절관계,

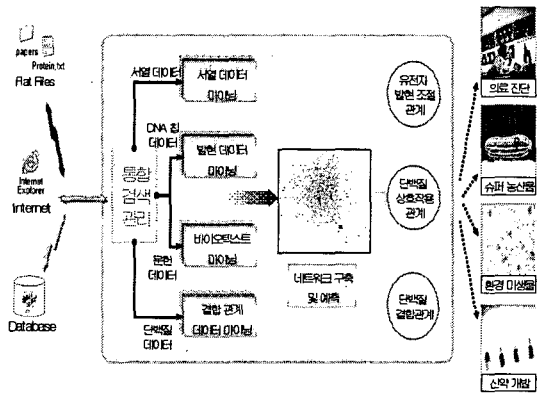

그림 1. 생물정보학 개요

단백질 상호작용관계, 단백질 결합관계 등을 분석하게 된다. 이러한 분석 결과를 통해서 쉽고 정확한 의료진단이 가능하고 인류 최대의 관심사 중에 하나인 식량부족을 해결할 수 있는 슈퍼 농산물의 생산도 가능하게 된다. 또한 환경을 깨끗하게 유지 해주는 환경 미생물의 발견 및 증식도 가능하게 되며 인류가 가지는 난치병을 고치는 신약 개발에도 큰 도움을 줄 수 있게 된다[5].

단백질 상호작용 정보와 경로 분석은 생물정보학분야에서 중요하게 다뤄지고 있다. 인간에게 존재하는 단백질의 기능을 밝히는데 도움을 줄 수 있기 때문이다. 생화학자들에 의해 이루어지고 있는 실험 및 연구결과들을 데이터베이스로 구축하게 되면 네트워크를 통한 분석을 통해 단백질의 기능을 유추해 낼 수 있다.

현재 단백질 상호작용 정보를 데이터베이스로 구축하는 연구가 활발히 진행되고 있다. 대표적으로 DIP, MIPS, PathCalling, ProNet 등의 데이터베이스가 있다. 이들 데이터베이스에서는 실험적으로 결정된 단백질 상호작용 정보를 수집하고 제공한다. DIP 데이터베이스의 경우 초파리, 효모, 인간 등의 단백질 정보를 제공하고 있다[8-11].

[그림 2]는 실험을 통해서 얻은 자료들을 기존의 웹이나 데이터베이스로부터 추출하여 구축한 단백질 네트워크를 나타낸다. 단백질 경로 분석은 포스트 게놈 시대에 신약개발의 실마리와 각종 생화학적 현상에 대한 실마리를 제공할 수 있다는 것으로 중요하게 다루어진다. 기존에 연구되었던 단백질 경로 분석 관련 프로그램은 전자통신연구원의 생물정보학사업에서 개발한 '사용자 중심의 단백질 상호작용 네트워크 항해' 경로 분석 프로그램과 InterViewer[12]가 있다. 그밖에 해외에는 stratagene사의 PathwayAssist[13] 프로그램이 있다.

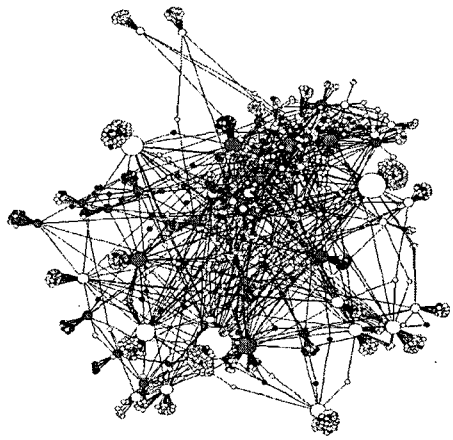

그림 2. 단백질 상호작용 네트워크

전자통신연구원에서 개발한 사용자 중심의 단백질 상호작용 네트워크 항해 프로그램은 텍스트 마이닝 도구를 제공하며 마이닝 도구로 추출한 단백질 상호작용 정보를 네트워크로 표현하여 사용자에게 시각적으로 보여주는 기능을 한다. 다양한 형태의 시각화 인터페이스를 제공하기 때문에 사용자가 사용하기에 쉽고 편한 반면 생물학적 의미를 찾을 수 있는 여러 가지 형태의 알고리즘을 제공하지 못한다. 현재는 최단경로와 기능경로 정도로 제한된다.

인하대학교 Web Intelligence 연구실에서 개발한 InterViewer는 단백질 상호 작용 정보를 시각화해서 사용자가 직관적으로 네트워크를 인식하고 원하는 경로를 찾을 수 있도록 제공하는 프로그램이다.

PathwayAssist는 해외에서 생화학자들이 필요한 프로그램을 전문적으로 개발하고 있는 Stratagene사에서 개발한 프로그램으로 단백질-단백질 사이의 상호작용

을 사용자에게 직관적으로 제공해주고 있다. 특별한 기능을 하는 단백질의 경우 일반적인 단백질과는 다른 형태로 보여주는 기능 등이 사용자가 네트워크를 직관적으로 인식하기 쉽게 도와준다. 하지만 아직까지 텍스트 마이닝 프로그램과의 연동이 이루어지지 않아서 자료를 수집하는데 어려운 점이 있다.

앞에서 언급된 기존 단백질 경로 분석 프로그램에서는 다양한 형태의 직관적이고 시각화 된 네트워크를 보여준다. 하지만 생물학적인 의미를 찾을 수 있는 다양한 알고리즘을 제공하지 못하고 있다. 생화학자들이 인터페이스를 통해서 네트워크상에 단백질들을 일일이 찾기에는 너무나도 어려운 부분이다. 또한 하루에도 수 없이 나오게 되는 생화학 관련 정보들로부터 단백질 상호 작용을 쉽게 추출할 수 있는 텍스트 마이닝 프로그램과의 연동도 충분히 고려해야 한다.

III. 단백질 네트워크 경로 분석 시스템 설계

1. 개요

본 논문에서 구현하는 프로그램은 단백질 상호작용 네트워크를 구축하고 구축된 단백질 네트워크를 여러 가지 알고리즘으로 적절한 경로를 분석하여 출력하는 시스템이다. 일반적으로 단백질 네트워크는 [그림 5]의 1단계에서 보이는 것과 같이 텍스트 마이닝 프로그램을 이용해서 텍스트 데이터로부터 단백질 상호작용 정보를 추출하거나 기존에 구축된 대형 데이터베이스로부터 정보를 추출한다. 본 논문에서 사용한 데이터는 HPRD 사이트의 단백질 정보를 추출하여 단백질 상호작용 데이터베이스를 구축하였다. 그리고 자체 개발한 텍스트 마이너를 통해 문헌 정보로부터 새롭게 발견된 단백질 상호작용 정보 및 관련정보에 대한 업데이트가 가능하도록 하고 있다. 상호작용 데이터베이스가 구축이 되면 다음 단계로 경로 분석 프로그램을 실행시켜 단백질 네트워크를 구성한다[14].

경로 분석 시스템은 구축된 네트워크에 다양한 경로 분석 알고리즘을 적용하여 세포사멸, 세포생성 등의 다양한 생물학적 의미를 찾을 수 있는 경로를 찾는다. 분석한 결과는 InterViewer에서 지원하는 파일 형식인 pnm 확장자의 형태로 저장을 해서 검색결과를 화면에 출력한다.

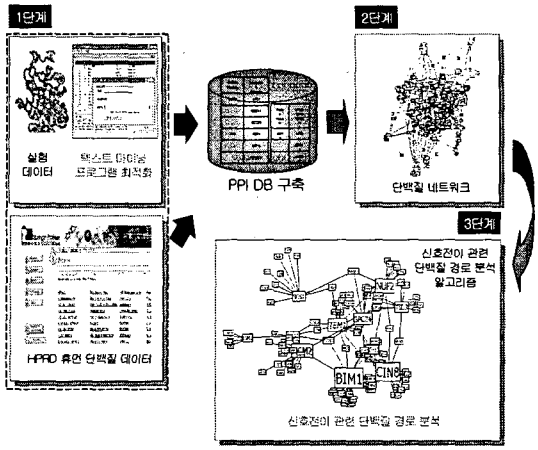

그림 5. 단백질 경로 분석 흐름

예를 들어, [그림 6]에서와 같이 구축해놓은 단백질 상호작용 데이터베이스로부터 경로 분석 프로그램이 데이터를 읽어와 네트워크를 구성하게 된다. 네트워크 구성이 완료되면 단백질 경로 분석을 수행하게 된다. 수행된 결과는 프로그램이 설치된 폴더에 pnm 확장자를 갖는 파일로 저장이 된. 저장된 파일은 경로 분석 프로그램에서 InterViewer를 호출하여 결과를 시각화 한다.

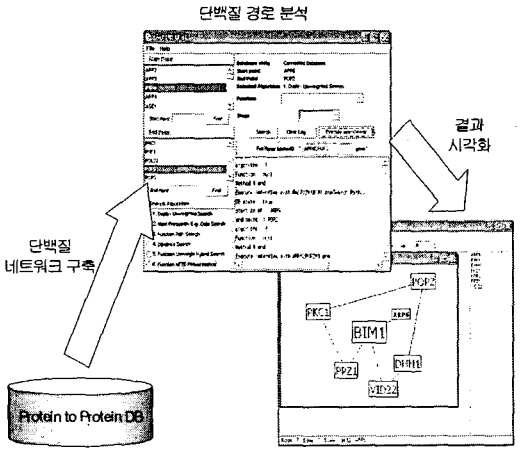

그림 6. 단백질 경로 분석 프로그램 실행 과정

2. 시스템 구조

단백질 경로 분석 시스템은 [그림 7]과 같이 여러 개의 모듈의 형태로 구현된다. 단백질 상호작용 데이터베이스, 네트워크 모듈, 경로 분석 모듈, 경로 분석 출력 모듈, 질의처리 모듈, 로그처리 모듈, 그리고 최상위 레벨에서 사용자의 입력을 받고 프로그램의 로그정보를 출력하는 인터페이스 모듈로 프로그램이 구성된다.

네트워크 모듈은 데이터베이스의 상호작용 정보를 통해서 단백질 네트워크를 구성하는 모듈이다. 단백질 경로 분석을 위한 기초가 되는 네트워크가 네트워크 모듈을 통해서 구성이 되면 실제 사용자의 질의를 받고 실행된 결과를 화면에 보여주는 인터페이스 모듈을 통해서 각종 로그와 경로 분석 결과 그리고 필요한 입력을 받는다.

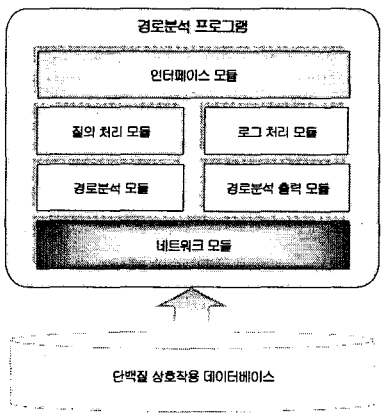

그림 7. 단백질 경로 분석 프로그램 구조

사용자의 다양한 요청이 들어오게 되면 인터페이스 모듈은 질의처리 모듈을 호출한다. 인터페이스 모듈로부터 호출된 경로 분석 모듈은 입력데이터의 타당성을 검사한 뒤 경로 분석 모듈의 적당한 알고리즘을 호출한다. 경로 분석 모듈의 실행이 종료되면 결과로 나오는 경로는 네트워크상에 기록이 된다. 그리고 분석된 결과의 출력을 위해서 경로 분석 출력 모듈을 호출한다. 경로 분석 출력 모듈은 호출된 알고리즘에 따라 적절한 출력 기능을 이용하여 InterViewer와 같은 프로그램을 통해서 보여줄 수 있는 형태의 파일구조로 출력을 한다. 그리고 요청에 따라서 InterViewer를 호출하여 질의처리 결과 경로를 출력한다. 이와 같은 작업이 수행되는 중간에는 로그처리 모듈을 통해서 사용자 인터페이스에 계속해서 프로그램 로그들을 출력하여 사용자가 프로그램의 현재 상태를 파악할 수 있도록 한다.

3. 경로 분석 알고리즘의 설계

단백질 경로 분석 알고리즘은 생물학적인 의미를 지니는 경로를 찾는 것이 목적이므로 일반적인 검색 알고리즘과는 다른 접근이 필요하다. 경로 분석 알고리즘은 두 단백질 노드 사이의 최단경로를 빠르게 찾아주는 양방향 최단 경로검색 알고리즘과 생화학적 실험 방법에 따라 신뢰성에 대한 가중치를 설정하고 그 가중치에 따라서 경로를 설정하는 가중치 검색 알고리즘 그리고 단백질의 생화학적 기능 경로를 따라서 검색을 하는 기능 중심 경로 검색 알고리즘으로 나누어 볼 수 있다. 본 논문에서 제안하는 경로분석 알고리즘은 생화학 전공 교수에 의해 타당성을 검토 받았다.

3.1 양방향 최단 경로검색 알고리즘

상호작용 하는 두 단백질 사이의 최단경로를 검색한다. 경로비용이 모두 같다는 전제로 이루어지는 검색으로 가장 적은 개수의 노드를 거치는 경로를 탐색한다. 검색 속도 향상을 위해서 양방향으로 검색을 진행한다.

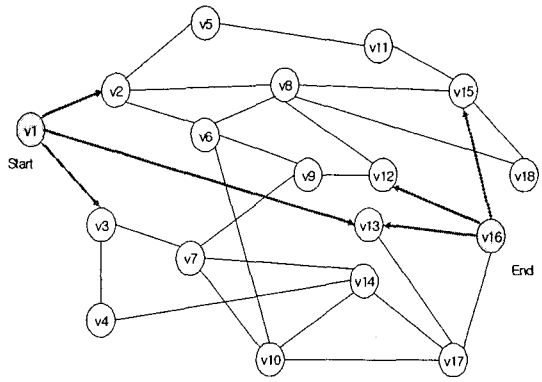

그림 8. 양방향 최단 경로검색 알고리즘

[그림 8]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 8]에서 v13노드를 통한 경로 v1, v13, v16이 최단 경로로 검색 된다.

3.2 가중치 중심 경로 검색 알고리즘

상호작용에서 가중치가 높은 노드들을 순회하며 검색한다. 본 연구에서 가중치로 사용하고 있는 정보는 단백질 상호작용 정보를 추출하기 위해 사용한 실제 실험(In Vivo, In Vitro, Yeast 2 Hybrid) 방법의 종류에 따라 다른 가중치를 부여한다. 그리고 여러 가지 실험 방법에 의해 추출된 상호작용 정보는 그만큼 높은 가중치가 설정된다. 이것은 상호작용 관계의 정확성 판단의 척도가 되어 보다 유용한 경로 분석을 가능하게 한다. 즉, 다양한 실험의 결과로 등장하는 상호작용은 그만큼 신뢰도가 있다는 반증이 되므로 신뢰성에 기반을 둔 경로 검색이 필요할 경우 사용하게 되는 검색 알고리즘이다.

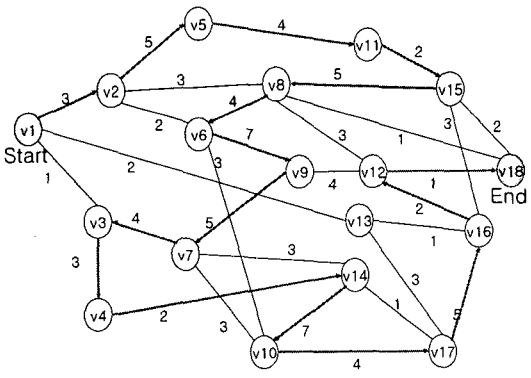

그림 9. 가중치 중심 경로 검색

[그림 9]에서 각각의 노드를 연결하는 간선의 숫자가 가중치를 의미한다. 가중치 중심 경로 검색은 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.

3.3 기능 중심 경로검색 알고리즘

시작점으로 주어진 단백질과 연결된 단백질 중 시작점에 해당하는 단백질의 기능과 일치하는 노드들을 따라서 경로를 검색한다.

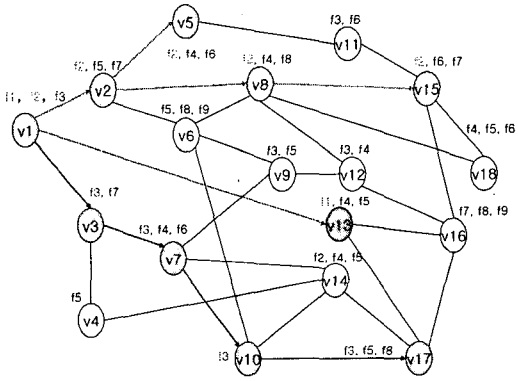

그림 10. 기능 중심 경로 검색

[그림 10]에서 v1이 시작 노드다. v1은 f1, f2, f3의 기능을 가지고 있으므로 먼저 순서상 검색을 원하는 기능을 선택을 하게 되면 선택된 기능에 대한 경로를 탐색하게 된다. [그림 10]에서와 같이 각각 f1기능을 가진 경로와 f2, f3의 기능 경로를 모두 검색할 수 있게 된다.

3.4 단방향 기능 검색 알고리즘

기능 경로 검색과 단 방향 검색 알고리즘을 병행한 알고리즘으로 일정 단계까지의 기능 경로를 볼 수 있는 알고리즘이다.

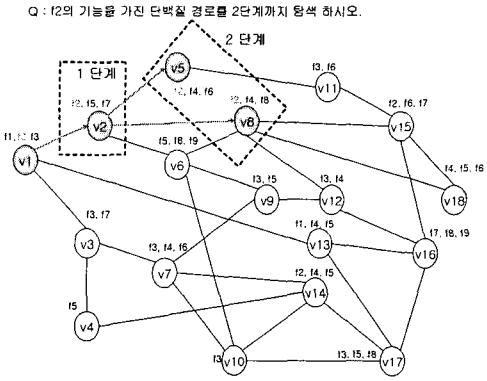

그림 11. 단방향 기능 검색

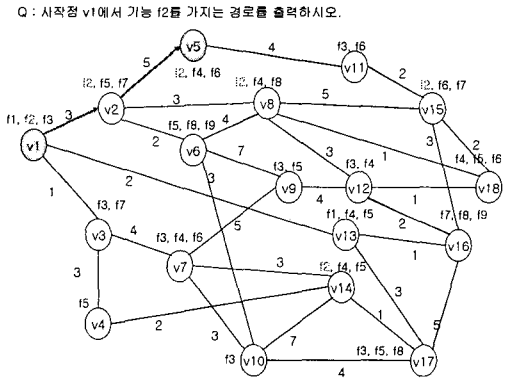

그림 12. 가중치 기능고려 경로 검색 알고리즘

검색시에 검색을 원하는 기능과 단계를 명시하게 되면 [그림 11]과 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 v1로부터 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 반복한다.

3.5 가중치 기능 고려 검색 알고리즘

단백질 노드의 기능 경로와 가중치 검색 알고리즘이 복합된 형태의 알고리즘이다. [그림 12]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 개인 경우 노드와의 간선의 가중치를 비교한 후 가중치가 높은 쪽 경로를 선택하게 된다. 같은 방식으로 경로를 설정하다가 만약 원하는 기능의 노드가 존재 하지 않는 경우에는 종료하고 경로를 출력하게 된다. 종료 조건은 질의에서 명시한 단계에 이르게 되거나 같은 기능을 가지는 연결된 모든 노드를 방문한 경우에 종료하게 된다.

Ⅳ. 단백질 경로 분석 시스템 구현

1. 구현 환경

본 논문에서 제안한 알고리즘은 Java 2 언어를 이용하여 구현하였다. 구현에 사용된 시스템은 펜티엄-Ⅳ 2.6c 프로세서 512Mbytes의 메모리를 가지며, 운영체제는 Windows 2000 Professional Service Pack 2 버전을 사용하였고, Java 2 SE 1.4.02버전을 사용하여 구현하였다. 단백질 정보를 저장하는 데이터베이스는 MS Access와 MySQL 3.23을 모두 지원하기 위해서 두 가지 데이터베이스를 모두 사용했다.

2. 구현 결과

구현한 시스템은 크게 6개의 영역으로 나누어지며 실행 화면은 [그림 13]과 같다. [그림 13]의 ①번 영역과 ②번 영역은 데이터베이스에 있는 단백질 리스트를 보여준다. 각각의 알고리즘에 맞게 원하는 단백질을 선택하면 된다. ①번과 ②번 영역을 통해서 단백질을 선택한 후에 ③번 영역에 있는 분석 알고리즘을 선택하고 만약 단백질 기능의 이름을 이용한 검색인 경우에는 ④번 영역이 활성화 되어 기능을 선택할 수 있다.

경로 분석을 위한 선택을 마치고 Search 아이콘을 클릭하게 되면 알고리즘에 맞는 적절한 결과를 파일로 출력하게 된다. 이때 각종 분석과정에서 발생하는 이벤트들이 ⑤번 영역의 로그 출력 창에 출력된다. 경로 분석 결과를 직관적으로 보고자 할 때는 ⑥버튼을 누르면 자동으로 시각화 툴인 InterViewer를 호출한다. 예를 들면, [그림 13]의 ⑥번을 통해 해당하는 결과를 출력한다.

분석 결과는 기본적으로 파일로 출력된다. 확장자는 pnm으로 InterViewer가 지원하는 네트워크 파일이다.

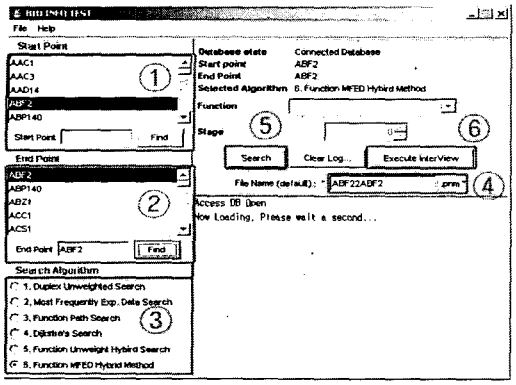

그림 13. 경로분석 프로그램 인터페이스

파일로 출력이 되고 나면 언제든지 파일을 읽어서 결과를 볼 수 있다. 따라서 분석한 결과가 여러 가지인 경우 단순히 파일을 실행해서 언제든지 경로검색 결과를 재확인 할 수 있다. [그림 14]는 경로 분석 결과로 출력된 파일을 InterViewer로 열어본 화면이다. 각각의 경로가 연결된 형태로 잘 표현이 되고 있다. 중간에 특별히 라벨이 크게 표시되는 단백질 노드는 특별히 연결된 단백질의 수가 많은 경우 크게 표시가 되어 허브 단백질을 쉽게 찾을 수 있다.

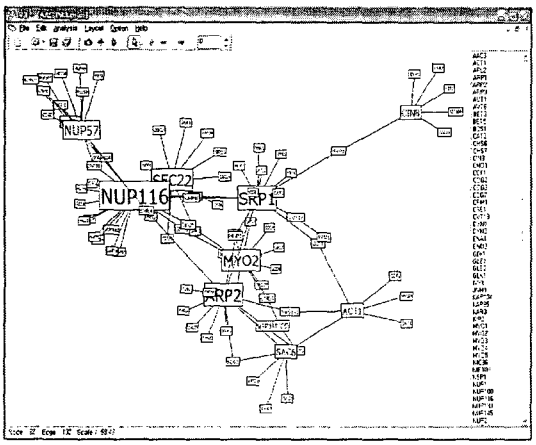

그림 14. 경로 분석 결과 출력 화면

3. 성능 평가

본 논문에서 구현한 여러 가지 알고리즘을 지원하는 경로 분석 시스템은 생물학 분야에서 생물학적으로 필요한 의미를 찾기 위한 분석 알고리즘 이므로 일반적인 최소비용 검색 알고리즘과 수행 시간을 비교하여 비슷한 응답시간을 가진다면 성능 관점으로 보았을 경우에 사용하는데 큰 부담이 없으므로 타당한 성능을 가진다. 따라서 구현한 5가지 알고리즘을 일반적인 최소비용 알고리즘인 Dijkstra[14] 알고리즘을 비교 모델로 구현하여 알고리즘의 검색 응답시간을 측정하였다. 각각의 알고리즘은 Java 2 1.4 버전을 기반으로 구현 했으며 성능 평가를 위한 데이터는 HPRD(Human Protein Reference Database)에서 웹로봇을 이용해 추출한 단백질 데이터 22,240개와 각각의 단백질이 가질 수 있는 기능정보를 유지한 기능 데이터 10,555개를 이용하여 각각을 평가했다.

표 1. 성능 평가 환경

평가환경	
CPU	Pentium-IV 2.6c
RAM	512 Mbytes
OS	Windows 2000 Pro. sp2

표 2. 성능 평가 데이터

데이터	
단백질	22,240 개
기 능	10,555 개

위와 같은 환경에서 시스템의 응답시간을 평가 해 보았다. 평가결과 [그림 15]에서 보듯이 일반적인 경로 분석 알고리즘에 해당하는 Dijkstra알고리즘의 응답시간보다 빠른 것으로 측정되었다.

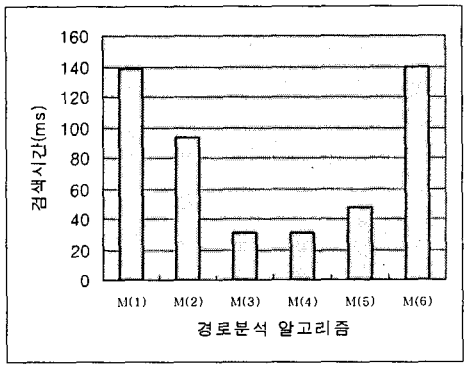

그림 15. 경로 분석 알고리즘 응답시간

Dijkstra는 모든 경로를 분석하기 위해서 모든 단백질 노드를 방문한다. 양방향 경로 분석에서는 양쪽 방향에서 확장하기 때문에 최단경로는 전체노드를 방문하기 이전에 검출된다. 또한 네트워크를 구성할 때 필요한 정보(예: 방문 플래그) 등을 부가적으로 사용하여 양방향 검색 시 확장노드 집합에 대한 공통집합을 검색해야하는 추가적인 처리비용은 없다.

같은 위치에서 종료조건이 만족하는 경우 방문하지 않았던 단백질 노드들을 모두 접근할 필요가 없으므로 가장 빠른 응답시간을 보였다.

표 3. 알고리즘 응답시간

알고리즘	응답시간(ms)
가중치 중심 경로검색 M(1)	138
기능중심 경로검색 M(2)	93
양방향 최단 경로검색 M(3)	31
단방향 기능 검색 알고리즘 M(4)	31
가중치 기능고려 경로검색 알고리즘 M(5)	47
dijkstra 알고리즘 M(6)	140

반면 가중치중심 경로 분석 알고리즘의 경우 Dijkstra의 경로 분석 방식이 거의 유사 하므로 비슷한 응답시간을 보였다. 기능경로 분석 알고리즘의 경우 모든 노드들을 방문하지는 않지만 양방향 경로 분석의 경우보다는 많은 노드들을 접근해야 하기 때문에 가중치 중심 경로 분석과 양방향 경로 분석의 응답시간의 중간 정도의 응답시간이 소요됐다. 각각의 알고리즘은 응답시간을 분석해 본 결과 일반적인 경로 분석 알고리즘에 해당하는 Dijkstra 알고리즘 보다 비슷하거나 빠른 응답시간을 보여 주었다.

Ⅳ. 결 론

본 논문에서는 단백질-단백질 네트워크를 구축하고 구축한 단백질 네트워크를 통해서 단백질 경로 분석을 할 수 있는 시스템을 설계하고 구현하였다. 단백질 경로 분석 시스템에서는 두 단백질 사이의 최단 경로검색, 상호관계 특성에 따라 부여된 가중치가 높은 경로를 분석, 단백질의 생물학적인 기능에 따른 경로 분석 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 InterViewer와 같은 GUI툴과 연동 할 수 있도록 구현했다. 전체 상호작용 관계를 네트워크로 구축함으로써 이를 통해 [그림 14]의 NUP116, SKP1, MYO2, ARP2 등과 같이 단백질 상호작용의 핵심 허브 노드를 중심으로 클러스터가 형성되는 것을 확인하는 것이 가능하며, 다양한 알고리즘을 통한 분석 방법을 제공함으로서 다양한 생물학적 의미를 예측할 수 있게 하였다.

향후 연구방향으로는 구축된 네트워크를 통해서 생물학적 의미를 찾을 수 있는 다양한 알고리즘을 추가 개발하고 사용자가 직관적으로 파악하면서 다양한 분석을 수행 할 수 있는 시각화 모듈을 개발하고자 한다.

참 고 문 헌

[1] http://hprd.org/

[2] http://www.genomesonline.org/

[3] Q. Jacobson, E. Rotenberg, and J. Smith, "Path-based Next Trace Prediction," Proc. 30th. Annual IEEE/ACM Intl. Symp. on Microarchitecture, pp.14-23, 1997.

[4] D. Benton, "Bioinformatics-principles and potential of a new multidisciplinary tool," Trends in Biotechnology, Vol.14, No.8, pp.261-272, 1996.

[5] P. G. Baker and A. Brass, "Recent development in biological sequence databases," Current Opinion in Biotechnology, Vol.9, pp.54-58, 1998.

[6] E. Alm and P. Arkin, "Biological networks," Curr Opin Struct Biol, Vol.13, No.2 pp.193-202, 2003.
[7] T. Ideker, et al, "Integrated genomic and proteomic analyses of a systematically perturbed metabolic networks," Science Vol.292, pp.929-934, 2001.
[8] http://dip.doe-mbi.ucla.edu/dip/Main.cgi.
[9] http://mips.gsf.de/
[10] http://portal.curagen.com/pathcalling_portal/
[11] http://bioinfo.sarang.net/wiki/ProNet.
[12] http://interviewer.inha.ac.kr/
[13] http://www.stratagene.com/
[14] Adam Drozdek, *Data Structures and Algorithms in Java*, BROOKS/COLE, USA, 2001.

저 자 소 개

이 재 권(Jae-Kwon Lee) 준회원

- 2003년 : 충북대학교 정보통신공학과(공학사)
- 2005년 : 충북대학교 정보통신공학과(공학석사)
- 2005년 3월~현재 : 충북대학교 정보통신공학과 박사과정

<관심분야> : 데이터베이스 시스템, XML, 생물정보학, 분산 객체 컴퓨팅 등

강 태 호(Tae-Ho Kang) 정회원

- 1999년 : 호원대학교 정보통신공학과(공학사)
- 2002년 : 충북대학교 정보산업공학과(공학석사)
- 2003년 3월~현재 : 충북대학교 정보통신공학과 박사과정

<관심분야> : 데이터베이스 시스템, 웹 컨텐츠 관리 시스템, 데이터마이닝, 생물정보학 등

이 영 훈(Young-Hoon Lee) 준회원

- 2004년 : 충북대학교 정보통신공학과(공학사)
- 2004년 3월~현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 데이터베이스 시스템, 모바일 동시성 제어, 생물정보학 등

유 재 수(Jae-Soo Yoo) 종신회원

- 1989년 : 전북대학교 컴퓨터공학과(공학사)
- 1991년 : 한국과학기술원 전산학과(공학석사)
- 1995년 : 한국과학기술원 전산학과(공학박사)
- 1989년~1996년 : 목포대학교 전산통계학과 전임 강사
- 1996년~현재 : 충북대학교 전기전자컴퓨터공학부 부교수

<관심분야> : 데이터베이스 시스템, 정보검색, 멀티미디어 데이터베이스, 분산 객체 컴퓨팅 등