Protein-Protein Interaction에 세포 내 위치 정보를 활용한 단백질 신호전달 경로 추출 알고리즘 연구

조미경*, 김민경*, 박현석*
*이화여자대학교 컴퓨터공학과
e-mail:edujmk@ewhain.net, minkk89@ewha.ac.kr, neo@ewha.ac.kr

Algorithm for extracting signaling pathways based on Protein-Protein Interaction and Protein location Information

Mi-Kyung Jo*, Min-Kyung Kim*, Hyun-Seok Park*
*Dept. of Computer Science and Engineering Ewha Womans University(소프트웨어응용)

요 약

단백질과 단백질의 상호작용은 최근 각광받고 있는 분야이다. 효모를 이용해 two-hybrid system의 실험으로 밝혀진 약 5,000여 개의 이스트 단백질의 위치정보를 이용하여 가중치를 부여고 단백질 신호 전달 경로 추출을 위한 LSPF 알고리즘을 최초로 제안 하였다. 세포 내 단백질 위치정보를 기반으로 제안한 LSPF 알고리즘에 의해 산출된 결과 중 의미적 상관도가 높은 것을 채택한 후 KEGG에서 제공하는 신호전달 경로와 같은 신호전달 경로를 추출하는지 비교분석 하였다. 최초로 제안된 단백질 위치정보를 이용한 신호전달 경로 찾기 연구가 발전되면 다양한 유전적 질병의 원인을 파악할 수 있고 치료제 개발에 단서를 얻을 수 있는 초석이 될 수 있다.

ABSTRACT

Intracellular signal transduction is achieved by protein-protein interaction. In this paper, we suggest performance algorithm based on Yeast protein-protein interaction and protein location information. We compare if pathways predicted with high valued weights indicate similar tendency with pathways provided in KEGG.

키워드 : Signaling Pathway, Localization, Algorithm, 상호작용, 위치정보, 신호전달, 경로추출, 알고리즘

Ⅰ. 서론

단백질 간 상호작용은 세포가 생명현상을 유지하기 위해 일어나는 필요한 현상이다. 효모와 초파리 등과 같은 대표적인 실험 모델에 있어서는 그 종이 가지고 있는 모든 단백질(n)에 대한 상호작용 여부를 실험하는 종 수준 데이터(n × n, genome scale)들이 출현하고 있다[1].

종 수준 단백질 상호작용 데이터는 단백질을 노드 상호작용을 에지로 표현하였을 때 소수의 연결도 높은 노드들이 존재하는 Scale Free Network[2]의 특징을 지니며 이러한 연결도 높은 노드들의 존재는 임의의 단백질 간 거리를 줄이는 역할을 한다.

또한 단백질 상호작용 네트워크는 Budding Yeast (Saccharomyces Cerevisiae)의 경우 제일 큰 클러스터 네트워크에 전체 단백질의 약 78%를 포함하는 것으로 알려져 있다[3].

신호 전달 경로란 세포가 외부 자극에 반응하여 새로운 단백질을 만들어 내야 할 필요가 있을 때 그 신호를 세포 표면으로부터 단백질 주형에 해당하는 DNA가 존재하는 세포핵 안쪽으로 전달하는 일련의 과정을 말한다. 이러한 신호 전달 과정은 필연적으로 단백질 간의 상호작용을 기반으로 한다. 이러한 신호 전달 과정을 통하여 외부 자극은 세포 전체로 그 자극을 확대하고 퍼뜨리는 것이 가능하다[4]. 신호 전달 경로는 전체 단백질 상호작용 네트워크의 부분 그래프에 해당한다. 생물학적 여러 의미 관계 구조들을 전산학적 관점에서는 네트워크로 접근이 가능하며 이론을 배경화한 주제로 다양한 그래프 이론들이 적용되고 있다[2]. 본 연구에서는 Yeast 단백질 간 상호작용 정보와 세포 내 위치 정보를 활용하여 세포막 단백질로부터 핵단백질까지 신호전달 경로를 찾는 알고리즘에 대해 제안하고자 한다.

Ⅱ. 신호전달 경로 추출 연구에 관한 선행연구

도메인 조합 기반 단백질 상호작용을 예측하는 기법을 재평가[7]하거나 확률 예측 틀[8] 연구나 온톨로지를 이용하

여 단백질 상호작용 네트워크를 개념적을 분유하여 레이아웃 하는 방법[9]이나 단백질 반응 데이터 품질향상[10]이나 데이터 신뢰도 향상[11]에 관한 연구가 있다. 그 외에 로컬 데이터베이스 자동 갱신[12]이나 데이터 모델링[13]에 관한 연구, 구조 유전체학의 대상이 될 단백질 단백질 상호작용을 선택하는 연구[14], 단백질 기능 모듈들과 정합될 수 있는 개념 모듈을 정의하고 탐색하는 연구[15], 네트워크 견고성의 핵심노드는 허브노드이지만 연결노드 또한 중요한 역할에 대한 연구[16], 도메인분석을 통한 단백질 기능발견 시스템[17]에 관한 연구, 단백질 경로 분석 시스템[18], 그래프 이론 기반의 데이터 분석 연구[19], 네트워크에서 상동성 기반 바이오 콤플렉스 예측[20]이나 템플릿 기반 동적 관리[21]에 관한 연구, 대사 경로 그래프 레이아웃을 위한 슈퍼노드화 연구[22], 연관속성개념 공간[23]이나 규칙[24]의 상호작용 예측에 관한 연구, 단백질 상호작용 예측 연구[25][26]들이 선행되어 연구 되었다. 또 다른 연구로는 사전 다른 정보를 사용하지 않고 그래프 알고리즘만을 적용하여 생물학적으로 의미 있는 부분을 추출 한 연구도 있다.

III. 단백질 상호작용 정보와 세포 내 위치정보 활용

단백질간의 상호작용에 대한 정보축적을 바탕으로 대사(metabolism) 및 신호전달(signal transduction), 그리고 세포주기(cell cycle)에 대한 정보를 수집할 수 있게 되었다.

3.1 단백질 상호작용 데이터베이스

단백질 간 존재하는 상호작용을 이해하는 것은 단백질의 기능 연구는 물론 시스템적인 생명 현상 이해를 가능하게 한다. 기초자료 정보를 DIP과 MIPS등의 데이터베이스로 통합하여 재구축 하였다. 이러한 데이터베이스들은 서로 다른 Id Number와 체계를 가지고 있기 때문에 통합된 정보를 가져오는 데는 어려움이 따르므로 연구에서는 NCBI의 정보를 이용한다.

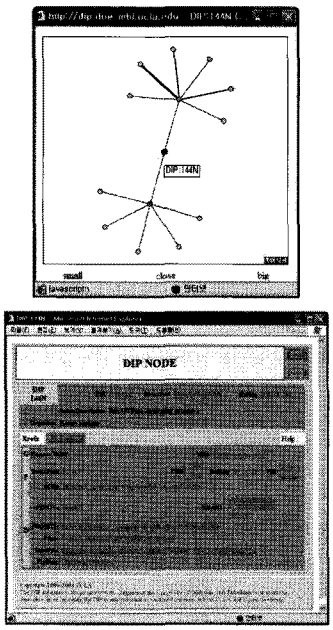
3.1.1 DIP

DIP은 단백질간의 상호작용에 초점을 맞춘 상호작용 데이터베이스이다. 여기에 수집된 자료는 모두 실험적인 근거를 기반으로 구축되었으며, 약 80개의 종으로부터 약 19,051개의 단백질과 55,732개의 상호작용을 기술 하였다. 또한 JDIP이라는 가시화 도구를 제공함으로써 사용자가 손쉽게 상호 작용하는 단백질을 시각적으로 볼 수 있으며 해당 단백질의 상세정보도 같이 볼 수 있는 기능을 제공하고 있다.

[표3-1] 단백질 상세 정보

ORGANISM	PROTEINS	INTERACTIONS	EXPERIMENTS
Drosophila melanogaster (fruit fly)	7052	20988	21085
Saccharomyces cerevisiae (baker's yeast)	4919	18224	22340
Escherichia coli	1831	7406	9051
Caenorhabditis elegans	2638	4030	4075
Helicobacter pylori	710	1425	1425
Homo sapiens (Human)	916	1407	2061
Mus musculus (house mouse)	202	292	397
Rattus norvegicus (Norway rat)	87	109	156
Others (102)	696		

또한 각 단백질의 정보를 Swiss-Prot, PDB, RefSeq, PIR, NCBI의 데이터베이스와 연결하여 관련 정보의 링크를 제공한다.



[그림 3-1] 단백질 상호작용 다이어그램

3.1.2 MIPS

MIPS(Munich Information Center of Sequences)[27]의 CYGD는 saccharomyces cerevisiae에 대한 정보를 제공하며 최근 Report, Graphical Displays, Genome의 특정 부분에 대한 Summary Table 등의 정보를 얻을 수 있다. 특히 Yeast에 관한 부분은 CYGD (The MIPS Comprehensive Yeast Genome Database, http://mips.gsf.de/genre/proj/Yeast/)에서 맡고 있다.

3.1.3 NCBI :

NCB(National Center for Biotechnology InformationI)에서 제공하는 모든 정보는 Entrez Search Engine에서 통합검색을 할 수 있다. PubMed는 MEDLINE Database 즉 의생명 관련 논문을 검색할 수 있

는 하나의 Search Engine으로서의 역할을 수행한다. Pubmed에 등록되어 있는Journal은 해당 Journal Database에서 제목 Subject Or Journal Title, Title Abbreviation, The NLM ID, The ISO Abbreviation, And Both The Print And Electronic International Standard Serial Numbers를 통해 검색할 수 있다. PubMed는 현재 1,400만 개 이상의 Citation을 갖고 있다.

〔표3-2〕 NCBI 정보 제공

GenBank	PubMed, PubMed Central
genome sequencing data 보유	biomedical research paper 및 기타 biotechnology 관련 자료 보유

3.2 단백질 상호작용 정보와 세포 내 위치정보의 활용

단백질 상호작용은 단백질 간 물리적으로 상당히 근접한 상태에서 일어나므로 이러한 반응을 이해하려면 단백질의 세포 내 위치 정보를 이해하는 것은 매우 중요하다 할 수 있다. 왜냐하면 막으로 구성된 위치에 존재하는 단백질은 그 위치 안에서 우선 상호작용이 일어날 수 있기 때문이다. 본 연구에서는 단백질 위치 정보의 중요성을 인식하고 세포 내 단백질 위치가 알려진 Budding Yeast를 대상으로 하여 단백질이 여러 위치에 존재할 수 있는 위치 정보를 이용한다.

단백질 위치정보의 예를 살펴보면 〔표3-3〕에서 YDR098C 단백질은 Nucleus와 Cytoplasm에 존재함을 알 수 있다.

〔표3-3〕 단백질별 위치 정보

	A	B	C	D	E
1	1) 단백질의 위치 정보				
2	YKL175W : nucleus|vacuole|				
3	YDR130C : nucleus|spindle pole|cytoplasm|				
4	YDR473C : nucleus|				
5	YDR098C : nucleus|cytoplasm|				
6	YHR069C : nucleus|nucleolus|cytoplasm|				
7	YDR239C : cytoplasm|				
8	YDR533C : nucleus|er|cytoplasm|				

선행연구로 밝혀진 세포내 단백질 위치정보〔27〕에서 알 수 있는 것처럼 21개로 분류되고 있으며 연구에서도 그 분류 정보를 이용한다. 또한 단백질 정보는 DIP과 MIPS에서 제공하는 단백질을 통합하여 고유한 단백질 5,495 개가 21개의 단백질 위치 영역에 분포 시킨다. 이때 단백질 하나에 대해 여러 위치 정보를 가질 수 있기 때문에 고유 단백질 5,495개를 다중 위치에 분포 시킨 후 7,786개의 위치정보 데이터를 사용 하였다. 단백질 다중위치 분포도를 〔표3-4〕에서 볼 수 있다.

〔표3-4〕 세포 내 위치별 단백질 분포표

	Cell내 단백질 위치	단백질 개수
1	bud	88
2	nucleus	2,001
3	lipid particle	26
4	early golgi	55
5	er to golgi	6
6	cytoplasm	2,666
7	nucleolus	207
8	actin	32
9	peroxisome	47
10	late golgi	46
11	golgi	180
12	microtubule	117
13	bud neck	99
14	endosome	57
15	nuclear periphery	62
16	er	553
17	vacuolar membrane	95
18	cell periphery	209
19	spindle pole	81
20	vacuole	220
21	mitochondrion	939
	total	7,786

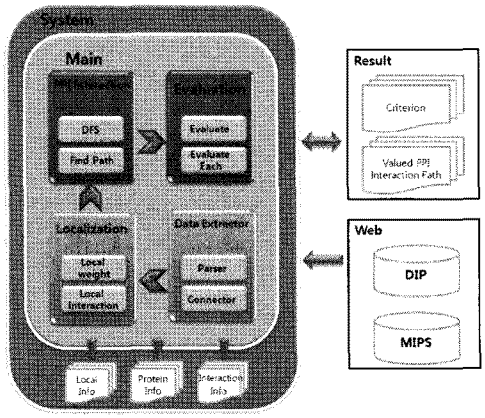
단백질 상호작용 데이터는 DIP에서 제공하는 34,797개와 MIPS에서 제공하는 30,429개를 통합한다. 그 중 중복 데이터 19,153개를 제거하여 고유 상호작용 단백질 46,073개와 논문을 통해 밝혀진 의미 있는 데이터 48개를 추가하여 총 46,121개의 단백질 상호작용 데이터를 입력 데이터로 사용하였다.

IV. 단백질 상호작용 정보와 세포내 위치정보 활용 알고리즘

연구에서 제안한 위치기반 알고리즘을 적용한 시스템을 LSPF(Local Signaling Pathway Finder)라 명명하였다. 시스템 구성을 살펴보면 시스템 전체의 기초데이터인 단백질 상호작용 데이터를 "데이터추출 및 통합"단계에서 DIP과 MIPS에서 다운한 데이터를 NCBI에서 제공되는 정보를 이용하여 MIPS코드 형식으로 코드 단일화를 한다.

"가중치 계산"단계에서는 단백질 21개 위치에 따른 단백질 위치별 분포도와 단백질 간 상호작용 분포도를 이용하여 확률값 테이블을 작성한다. 이 단계까지를 기초 작업 단계로 분류 할 수 있다. "단백질 간 상호작용 경로 찾기"단계는 입력값으로 시작 단백질과 목표 단백질 그리고 깊이를 입력 받아 DFS(Depth First Search) 방식으로 모든 가능 경로를 찾는다. "가중치 적용"단계에서는 구한 모든 상호작용 경로에 위치별 가중치와 상호작용 가중치로 평가를 실시한다. 마지막 단계인 "통계 및 분석"단계에서는 가중치 알고리즘을 적용

하여 구한 단백질 상호작용 경로의 최종 값을 내림차순으로 정렬하여 동일한 순위 내에서 최대값 포함 노드를 추출하여 KEGG 신호전달 경로와 유사도를 비교분석 한다.



[그림4-1] LSPF의 구성도

이중 핵심 단계는 DFS를 통해 경로를 찾는 "단백질 간 상호작용 경로 찾기" 단계와 평가를 적용하는 "가중치 적용" 단계이다.

[표4-1] 알고리즘의 단계별 세부사항

단계	세부사항
데이터 추출 및 통합	DIP 데이터 형식과 MIPS 데이터 형식을 MIPS 데이터 형식으로 통일하기 위하여 NCBI 데이터를 활용한다.
가중치 계산	단백질의 위치 정보, 단백질간 위치에 따른 상호작용 정보를 이용하여 위치 별 확률과 위치에 따른 상호작용 확률로 가중치를 구한다.
단백질간 상호작용 경로 찾기	주어진 시작단백질, 목표단백질, 최장경로 길이를 기준으로 깊이 우선 탐색을 통해 모든 단백질간 상호작용 경로를 찾는다.
가중치 적용	구해진 상호작용 경로를 위치 별 가중치와 상호작용 가중치로 평가한다.
통계 및 분석	가중치 적용을 거쳐 구한 위치기반 단백질간 상호작용 경로를 정렬하여 실제 경로와의 유사도를 비교 분석한다.

4.1 데이터추출 및 통합

실험 데이터로는 이스트 단백질을 대상으로 한다. 이때 상호작용 데이터를 제공하는 DIP과 MIPS 데이터베이스에서 단백질 상호작용과 단백질 위치정보를 다운한다. 다운한 정보의 단백질 코드는 각각 다른 형태로 제공된다. DIP 코드는 P로 시작되는 코드이며 MIPS 코드는 Y로 시작되는 코드 형태이다. 이때 코드 단일화를 위해 NCBI에서 제공되는 정보를 이용하여 MIPS코드 형식으로 변환 한다.

변환된 단백질 상호작용 데이터 약 46,121개는 각 단백질 간의 상호작용 여부를 나타내며 [표4-2]에서 인접행렬로 나타내었다. 예를 들어 행 A6셀에 데이터 YAL008W와 열 E1셀에 데이터 YAL007C와는 단백질 간 상호작용이 일어나고 있음을 나타낸다.

[표4-2] 46,121개 단백질 상호작용 인접행렬 예

	A	B	C	D	E	F	G	H	I
1	구분	YAL001C	YAL002W	YAL005C	YAL007C	YAL008W	YAL009W	YAL010C	YAL011W
2	YAL001C			1	1				
3	YAL002W	1		1	1				
4	YAL005C	1	1		1	1	1	1	1
5	YAL007C		1	1		1		1	1
6	YAL008W		1	1	1			1	1
7	YAL009W	1	1	1	1	1		1	1
8	YAL010C	1	1		1		1		1
9	YAL011W	1	1	1	1	1	1	1	
10	YAL012W	1			1	1	1		
11	YAL014C	1			1	1	1		
12	YAL015C	1		1	1	1	1		

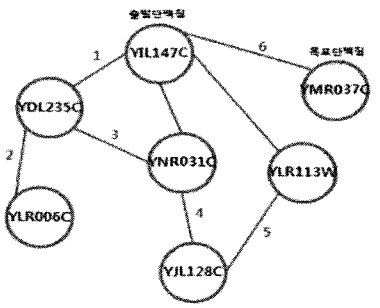
4.2 가중치 계산

단백질의 21개 위치정보에 따른 위치별 단백질 분포도와 두 단백질 간 상호작용 분포도를 이용하여 단백질 가중치 확률 값과 두 단백질 간 상호작용 확률 값을 가중치 적용단계의 알고리즘 수행 때 테이블로 작성된다.

이때 A와 B를 단백질이라 가정한다. 단백질의 위치 정보와 단백질의 상호작용 정보를 이용하여 위치 가중치 확률 값과 위치 간 상호작용 확률 값을 구한다. A위치의 가중치란 단백질 위치 정보를 이용해서 Log((A위치에 속하는 단백질의 경우의 수 / 모든 단백질이 해당하는 위치수의 합 * 100) * 100)으로 구한다. 만약 그 값이 0.0f 일 경우 곱하는 함수를 이용하기 때문에 전체 값이 0이 되는 것을 방지하기 위해 0.1f 값을 부여 하였다. A-B위치 상호작용 확률이란 모든 단백질 상호작용 정보와 단백질 위치 정보를 이용해서 Log ((위치 A 와 위치 B Interaction 경우의 수/ 모든 위치 경우의 수 * 100) * 100)으로 구한다.

4.3 단백질 간 상호작용 경로 찾기

프로그램 실행 화면에서 시작 단백질(예:YIL147C)과 종료 단백질(예:YMR037C) 그리고 깊이(예:8)를 실행화면의 입력 값으로 준다. 단백질 간 상호작용 데이터 46,121개를 입력 자료로 하여 DFS(Depth First Search) 알고리즘을 수행하며 이때 입력받은 시작 단백질로부터 출발하여 제한 깊이를 비교해 가며 핵에 존재하는 목표 단백질까지 모든 신호전달 가능 경로를 찾게 된다. DFS 알고리즘을 이용하여 탐색하는 예를 [그림4-2]에서 제시 한다. 예를 들어 탐색경로를 살펴보면 YIL147C을 출발 단백질로 하여 YDL235C-YLR006C-YNR031C-YJL128C-YLR113W를 찾고 마지막 목표 단백질인 YMR037C의 순으로 신호전달 경로를 추출한다.



[그림4-2] DFS알고리즘을 수행 예

〔표4-3〕 단백질 간 상호작용 LSPF 알고리즘 결과

	A	B	C	D	E	F	G	H
1	start : YLR332W, target : YPL089C, maximum length : 8							
2	YLR332W	YOL109W	YBR069C	YPR198W	YHR026W	YJR091C	YDR167W	YPL089C
3	YLR332W	YOL109W	YBR069C	YLR372W	YBR054W	YJR091C	YDR167W	YPL089C
4	YLR332W	YOL109W	YBR069C	YLR372W	YJR010C	YJR091C	YDR167W	YPL089C
5	YLR332W	YOL109W	YBR069C	YLR372W	YML123C	YJR091C	YDR167W	YPL089C
6	YLR332W	YOL109W	YBR069C	YLR372W	YHR026W	YJR091C	YDR167W	YPL089C
7	YLR332W	YOL109W	YBR069C	YLR372W	YMR055W	YDL029W	YHR030C	YPL039C
8	...							
9	(중략)							

4.4 가중치 적용

단백질 위치 정보의 중요성을 인식하고 위치 정보를 활용하였다. "단백질 간 상호작용 경로 찾기"단계에서 산출된 결과를 기반으로 모든 경로에 위치별 가중치와 상호작용 가중치를 이용한 평가함수를 적용하여 평가한다. "단백질 간 상호작용 경로 찾기"를 평가하기 위해서는 패스를 이루는 단백질 앞 뒤 순서 간 모든 평가 값을 곱한다. 단백질의 위치정보와 위치 간 확률 값을 담은 이차원 배열 정보를 사용 하였다. 이때 단백질 위치 가중치 계산은 21개의 위치에 따른 각 위치별 단백질 개수의 확률 값을 사용 한다. 단백질 간 상호작용 확률 값은 배열 (21위치X21위치)을 이용하여 두 단백질이 갖는 모든 위치의 상호작용 경우를 계산한 확률 값을 사용 하였다. 예를 들어 단백질의 위치정보가 YMR088C : vacuolar membrane, cytoplasm YLL040C: mitochondrion, endosome, cytoplasm 일때 YMR088C와 YLL040C의 모든 상호작용 경우의 수는 위치 개수 2*3한 값인 6개가 발생하는 것이다. 산출물은 〔표4-4〕에서 제시 한다.

〔표4-4〕 위치 인터랙션의 모든 경우의 수

	A	B	C	D	E	F	G	H	I	J	K
1	YSAJA2-YGL178W										
2											
3	YAL001C-YBR123C										
4	nucleus(1)-nucleus(1)/mitochondrion(20)-nucleus(1)/cytoplasm(5)-nucleus(1)										
5	YAL002W-YDR026C										
6	endosome(13)-cytoplasm(5)										
7	YAL003W-YLR149W										
8	endosome(13)-golgi(10)/endosome(13)-vacuole(19)/endosome(13)-vacuolar membrane(16)/endosome(13)-endosome(13)										
9	YAL005W-YLR386C										
10	endosome(13)-vacuole(19)/endosome(13)-vacuolar membrane(16)/endosome(13)-cytoplasm(5)										
11	YAL002W-YMR021W										
12	endosome(13)-vacuole(19)/endosome(13)-vacuolar membrane(16)/endosome(13)-endosome(13)										
13	YAL002W-YPL045W										
14	endosome(13)-vacuole(19)/endosome(13)-vacuolar membrane(16)										
15	YAL003W-YKL081W										
16	cytoplasm(5)-mitochondrion(20)/cytoplasm(5)-cytoplasm(5)										

단백질 위치정보와 단백질 상호작용 값을 이용하여 평가함수인 두 가지 형태의 함수를 정립하였다. 첫째는 A-B 위치 상호작용 확률 값 둘째는 A위치의 가중치 * B위치의 가중치 * A-B 위치 상호작용 확률 값이다. 단백질1과 단백질2의 단백질 간 상호작용 값을 평가하기 위해서 단백질1과 단백질2의 단백질 위치 정보를 가져와 산출한다.

첫째 만약 단백질1의 위치가 없거나 단백질2의 위치가 없을 경우 1.0f의 값을 평가한다. 둘째 단백질1과 단백질2 모두 하나 이상의 위치를 가질 경우 그 위치 간 쌍의 모든 경우의 수 중에서 평가함수 정보를 통해 산출한 값이 가장 큰 값을 선택 한다. 이렇게 선택된 단백질 간 상호작용 값들을 곱함으로써 한 개의 "단백질 간 상호작용 경로 찾기" 단계를 평가하게 된다.

4.5 통계 및 분석

가중치를 적용하여 구한 단백질 상호작용 경로 값을 내림차순으로 정렬하여 동일한 순위 내에 여러 개의 값이 존재할 경우 그 동일 순위의 값 중 최대값의 포함노드를 포함하고 있는 값을 추출하여 KEGG신호전달 경로와의 유사도를 비교분석한다.

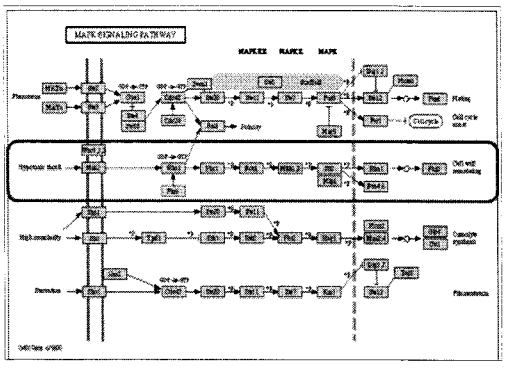
YLR332W-YPR165W-YBL105C-YJL095W-YOR231W-YHR030C-YER111C인 단백질 신호전달 경로 추출을 목표로 하여 시작 단백질을 YLR332W로 목표 단백질을 YER111C로 하고 깊이를 7로 하는 프로그램을 실행한 후 가중치 알고리즘을 적용하였다. 이때 산출된 결과 값을 내림차순 정렬하여 중복 값을 제거한 후 유일한 값에 대한 순위를 부여하고 최대값을 선출한다. 이때 최대값에 해당하는 것이 여러 개의 중복 값이 산출된 원인으로는 단백질이 여러 위치 정보를 가지고 있을 때 평가 대상에서 가장 큰 값을 가지고 있는 위치 값을 선택하기 때문이다. 이후 최고값을 갖는 여러 경로들을 대상으로 하여 깊이가 가장 큰 값을 갖는 경로를 추출한다. 이때 추출된 경로가 비교 기준 대상인 KEGG 신호전달 경로와 일치하는 지에 대한 유사도 성능 평가를 실시한다.

V. 적용 및 평가

5.1 MAPK에 적용

MAP(Mitogen-Activated Protein) Kinase 신호전달 경로는 세포막 단백질에 세포분열 유도물질인 Mitogen이 결합하며 시작된다. 이러한 신호전달 결과 세포 분화, 분열, 생존, 사망 등의 현상이 일어난다. 이러한 MAP Kinase 신호전달 경로는 세포막에서 발생한 신호를 핵 안쪽까지 증폭하면서 전달하고, 각 단계에서 어느 단백질을 활성화 하느냐에 따라 다양한 반응을 나타낼 수 있다. KEGG MAPK Signaling Pathway는 〔그림5-1〕과 같다.

그림 115



〔그림 5-1〕 MAPK Signaling Pathway (S.cerevisiae)

KEGG MAPK Signaling Pathway는 4개의 Function으로 구분할 수 있다. 그 중 실험에서는 "Hypotonic shock" 기능에 대해 실험하여 그 결과를 분석한다.

5.2 평가 및 분석

LSPF 알고리즘을 거쳐 산출된 평가 값을 내림차순 하여 가장 큰 값을 선정한 후 그 중 가장 많은 노드를 포함하고 있는 패스를 선정하여 비교 기준 대상인 KEGG에서 제공하는 MAPK"Hypotonic shock" 신호전달 경로를 실험 결과와 비교 하였다.

〔표5-1〕 LSPF 알고리즘 수행을 위한 입력 데이터

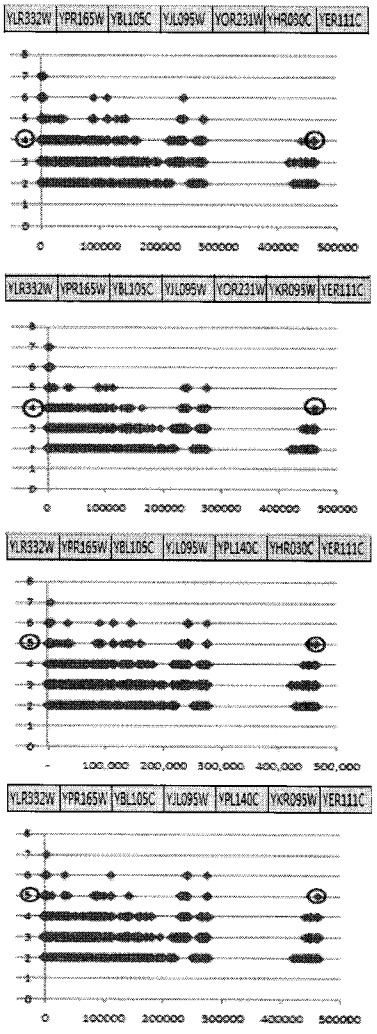
P	Q	R	S	T	U	V	W
start : YLR332W, target : YER111C, maximum length : 8							
YLR332W	YOL109W	YBR069C	YLR372W	YMR058W	YDL029W	YHR030C	YER111C
YLR332W	YOL109W	YBR069C	YLR372W	YER118C	YLR096W	YHR030C	YER111C
YLR332W	YOL109W	YBR069C	YLR372W	YER118C	YHL007C	YPL256C	YER111C
YLR332W	YOL109W	YBR069C	YLR372W	YDR456W	YGL008C	YHR030C	YER111C
YLR332W	YOL109W	YBR069C	YLR372W	YCR024C	YGL008C	YHR030C	YER111C
YLR332W	YOL109W	YBR069C	YLR372W	YGL054C	YNL201C	YPL153C	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YJR091C	YBR160W	YLR182W	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YJR091C	YBR160W	YPL256C	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YJR091C	YBR160W	YPR119W	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YJR091C	YBR160W	YKR095W	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YJR091C	YPL140C	YHR030C	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YJR091C	YPL140C	YKR095W	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YJR091C	YCL014W	YPR119W	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YJR091C	YHR193C	YHR030C	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YCL052C	YLR268W	YHR030C	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YLR447C	YGR086C	YHR030C	YER111C
YLR332W	YOL109W	YBR069C	YJR010C	YLR447C	YOR231W	YHR030C	YER111C

"Hypotonic shock" 기능 중 4개의 패스를 대상으로 LSPF 알고리즘 방법1과 방법2를 이용하여 신호전달 경로 추출 실험을 한 결과 〔표5-2〕와 같은 결과를 얻었다. 다시 말해 4개의 패스 중 2개의 패스는 목표노드 7개 중 4개 노드가 일치하는 패스를 찾았고 나머지 2개의 패스는 5개 노드가 일치하는 패스를 찾았다. 여기서의 특징은 LSPF 알고리즘 방법1과 방법2가 모두 일치하는 결과를 얻었다. 이 결과를 통해 알 수 있는 것은 방법1과 방법2 알고리즘 모두 단백질 위치정보 기반에서 착안한 알고리즘으로 방법이 매우 유사한 알고리즘이라 할 수 있겠다.

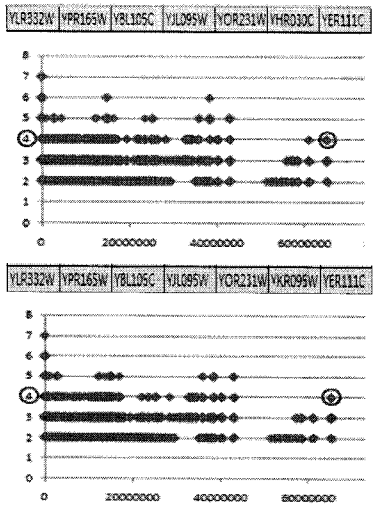
〔표5-2〕 "Hypotonic shock" 경로 및 수행결과

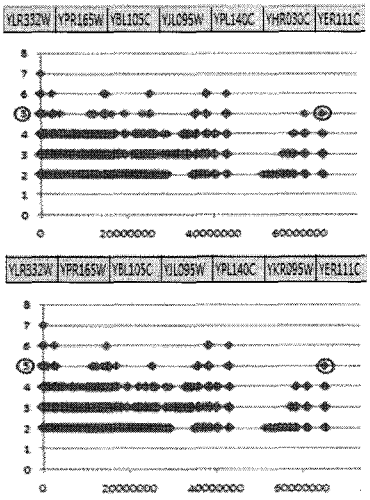
Function	Pathway							방법1,KEGG	방법1(%)	방법2,KEGG	방법2(%)
	1	2	3	4	5	6	7				
Hypotonic shock	YLR332W	YPR165W	YBL105C	YJL095W	YOR231W	YHR030C	YER111C	4	57%	4	57%
	YLR332W	YPR165W	YBL105C	YJL095W	YOR231W	YKR095W	YER111C	4	57%	4	57%
	YLR332W	YPR165W	YBL105C	YJL095W	YPL140C	YHR030C	YER111C	5	71%	5	71%
	YLR332W	YPR165W	YBL105C	YJL095W	YPL140C	YKR095W	YER111C	5	71%	5	71%

〔표5-2〕의 결과를 살펴보면 실험 대상인 4개의 패스 중 2개의 패스는 비교 기준인 KEGG와의 유사도 측정에서 57%의 정확도를 얻었으며 나머지 2개의 패스는 71%의 정확도를 얻었다. 단백질의 위치정보를 기반으로 한 이 실험은 최초로 제안하는 연구로 생물학자나 컴퓨터 바이오인포매틱스를 연구하는 학자들에게 앞으로 단백질 신호전달 경로 연구에 초석이 되기를 기대한다. 실험 결과를 그래프로 표현하여 〔그림5-2〕와 〔그림5-3〕에서 보여준다.



〔그림5-2〕 4개 패스에 방법1 알고리즘 수행결과





〔그림5-3〕 4개 패스에 방법2 알고리즘 수행결과

VI. 결론 및 향후 계획

선행연구로는 신호전달 경로 추출에 있어서 투 하이브리드를 통해 얻은 단백질 상호작용 단백질과 DNA Microarray Data 실험으로부터 얻은 발현 프로파일을 사용 한다. 그러나 좋은 결과를 얻지 못한다. 그래서 본 연구에서는 Yeast 상호작용 데이터 46,121개의 단백질 정보와 약 5,000여 개의 Yeast(S. cerevisiae) 위치 정보를 이용하여 신호 전달 경로 추출 알고리즘을 제안 하였다. 상호작용을 기반으로 출발점을 세포막 단백질로 하고 도착점을 핵에 있는 단백질로 하였으며 신호전달 경로를 추출 하였다. 시뮬레이션 결과를 통하여, KEGG에서 제공하는 MAPK "Hypotonic shock" 기능 신호전달 경로와 유사도를 측정한 결과 방법1과 방법2 알고리즘 모두 동일하게 4개의 패스 중 2개의 패스는 57%의 정확도를 얻었으며 나머지 2개의 패스는 71%의 정확도를 얻었다.

신호전달 경로 패스 찾기가 최초로 발표하는 연구 이며 또한 단백질 위치정보를 이용한 신호전달 경로 패스 찾기도 최초이다. 많은 생물학자나 컴퓨터공학자들 사이에 집중이 되는 연구이기는 하지만 아직까지 활발한 연구가 되지 못하는 이유 중에 하나는 바이오인포매틱스 연구자가 많지 않았기 때문인 것으로 사료된다. 차기 연구에서는 MAPK의 기능을 좀 더 확대하여 실험함으로써 실험 정확도를 더욱 높이고 아직 실험을 통해 밝혀 지지 않은 단백질들이 많음을 감안 할 때 미지의 단백질을 고려한 신호전달 경로 찾기에 대한 연구가 필요할 것으로 사료된다.

참 고 문 헌

〔1〕 L. Giot, J. S. Bader, C. Brouwer et al., "A Protein Interaction Map of Drosophila melanogaster", Science, Vol. 302, No. 5651, 1727-1736, 2003

〔2〕 Reuven Cohen, Shlomo Havlin, "Scale-Free Networks Are Ultrasmall", vol. 90, 90-94, 2003

〔3〕 Ibert-Laszlo arabasi, Zoltan N. Oltvai, "Understanding the Cell's Funcrional Organization ", Nature, vol. 5, 101-103, 2004

〔4〕 Silvia D. M. Santos, Peter J. Verveer, Philippe I. H. Bastiaens, "Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate", Nature, vol9, 324-330, 2007

〔5〕 Jose B. Pereira-Leal , Anton J. Enright , Christos A. Ouzounis, "Detection of functional modules from protein interaction networks", vol. 54, 49-57, 2004

〔6〕 Victor Spirin , Leonid A. Mirny, "Protein complexes and functional modules in molecular networks", vol. 100, no. 21, 12123-12128, 2003

〔7〕 한동수, 김홍숙, 장우혁, 이성독, "도메인 조합 기반 단백질 상호작용 가능성 순위 부여 기법", 정보 과학회 논문지:컴퓨팅의 실제, 제11권, 제5호, 427~435, 2005

〔8〕 한동수, 서정민, 김홍숙, 장우혁, "도메인 조합 기반 단백질 단백질 상호작용 확률 예측 틀", 정보 과학회 논문지:컴퓨팅의 실제, 제10권, 제4호, 299~308, 2004

〔9〕 방선이, 최재훈, 박종민, 박수준, "단백질 상호작용 네트워크의 개념 분류 레이아웃", 한국 컴퓨터 종합 학술대회 논문집, Vol.33, No.1(A), 61~63, 2006

〔10〕 장희선, 원민영, 김태경, 조완섭, "온톨로지 기반의 단백질 반응 데이터 품질향상 연구", 한국 컴퓨터 종합 학술대회 논문집, Vol.33, No.1(C), 94~96, 2006

〔11〕 이민수, 박승수, 이상호,"특징 추출과 분석 기법에 기반한 단백질 상호작용 데이터 신뢰도 향상 시스템",정보처리학회논문지, V13B, no.110, 679-688, 2006

〔12〕 김기봉, "단백질 상호작용 데이터의 효율적 관리와 자동 갱신을 위한 시스템 설계와 구현", Journal of Life Science, Vol.18, No. 3, 318~322, 2008

〔13〕 박지숙, 백은옥, 이공주, 이상혁, 이승록, 양갑석, "세포 신호전달 경로 데이터베이스를 위한 데이터 모델링", 정보 과학회 논문지:데이타베이스, 제30권, 제6호, 573~584, 2003

〔14〕 박황서, "구조 유전체학의 새로운 대상로서의 단백질-단백질 상호작용", 8~9, Korea Genome Organization

〔15〕 박종민, 최재훈, 박수준, 양재동, "단백질 상호작용 네트워크에서의 개념 기반 기능 모듈 탐색 기법", 정보 과학회 논문지:시스템 및 이론, 제34권, 제10호, 474~492, 2007

〔16〕 안명상, 고정환, 유재수, 조완섭,"단백질 상호작용 네트워크에서 연결노드 추출과 그 중요도 측정",한국 산업

정보학회 논문지, 제12권, 제5호, 1~13, 2007
[17] 강태호, 류제운, 유재수, 김학용, "단백질 허브 네트워크에서 도메인분석을 통한 단백질 기능발견 시스템", 한국콘텐츠학회논문지, Vol.8, No.1, 259~271, 2008
[18] 이재권, 강태호, 이영훈, 유재수, "단백질 경로 분석 시스템의 설계 및 구현", 한국콘텐츠학회논문지, Vol.5, No.6, 31~40, 2005
[19] 진희정, 윤지현, 조환규, "그래프 이론 기반의 단백질 단백질 상호작용 데이터 분석을 위한 시스템", 정보 과학회 논문지:시스템 및 이론, 제33권, 제5호, 267~281, 2006
[20] 최재훈, 박종민, 박수준,"단백질 상호작용 네트워크에서 상동성 기반 바이오 콤플렉스 예측", 한국 전자통신 연구원, Vol.33, No.1(A), 64~66, 2006
[21] 박종민, 최재훈, 박선희, "단백질 상호작용 네트워크를 위한 템플릿 기반 동적 관리", 한국 컴퓨터 종합 학술대회 논문집, Vol.32, No.1(B), 289~291, 2005
[22] 송은하, 김민경, 이상호, "유전체 수준 대사 경로 그래프 레이아웃을 위한 슈퍼노드화 방안에 관한 연구", 한국 컴퓨터 종합 학술대회 논문집, Vol.33, No.1(A), 58~60, 2006
[23] 엄재홍, 장병탁, "연관속성개념공간으로의 사상을 이용한 단질 상호작용 예측",한국 컴퓨터 종합 학술대회 논문집, Vol.33, No.1(A), 73~75, 2006
[24] 엄재홍, 장병탁,"최적 연관 속성 규칙을 이용한 비명시적 단백질 상호작용의 예측",정보 과학회 논문지:소프트웨어 및 응용, 제33권, 제4호, 365~377, 2006
[25] 한동수, 김홍숙, 장우혁, 이성독, "단백질 상호작용 예측 서비스 시스템", 정보과학회논문지:컴퓨팅의 실제, 제11권, 제6호, 503~513, 2005
[26] 이미경, 김기봉, "단백질 상호작용 추론 및 가시화 시스템", 정보과학회논문지:소프트웨어 및 응용, 제31권, 제12호, 1602~1610, 2004
[27] Won-ki Huh, James V.Falvo et al., "Global anlaysis of Protein localization in budding Yeast." Nature, 2003