韓國컴퓨터情報學會 論文誌
第14卷 第3號, 2009. 3.

2009-14-3-4-6

S. cerevisiae 단백질간 상호작용과 세포 내 위치 정보를 활용한 MAP Kinase 신호전달경로추출 및 예측을 위한 고성능 알고리즘 연구

조 미 경*, 김 민 경**, 박 현 석**

High performance Algorithm for extracting and redicting MAP Kinase signaling pathways based on S. cerevisiae rotein-Protein Interaction and Protein location Information

Jo Mi Kyung*, Kim Min Kyung**, Park Hyun Seok**

요 약

세포 내에서 일어나는 단백질 신호 전달 과정은 단백질간의 상호작용을 통해 수행되고 조절된다. Yeast 상호작용 정보와 녹색형광단백질(GFP)을 이용하여 밝혀진 약 5,000여 개의 Yeast 단백질 위치정보를 이용하여 가중치를 부여 하고 신호 전달 경로 추출 및 예측을 위한 고성능 LocSPF 알고리즘을 최초로 제안 하였다. 가중치 알고리즘에 의해 산출된 결과 중 의미 상관도가 높은 것을 채택한 후 KEGG에서 제공하는 신호전달 경로와 같은 신호전달 경로를 추출하는지 유사도 비교를 하였다. 한편 더 나아가 아직 실험을 통해 밝혀지지 않은 단백질 신호전달 경로를 예측하여 결과를 제시함으로써 본 연구를 통해서 알려지지 않은 새로운 신호전달 경로를 발견하거나 이전 경로에 참여 하지 않은 단백질들을 발견할 수 있는 가능성을 제시 하였다.

Abstract

Intracellular signal transduction is achieved by protein-protein interaction. In this paper, we suggest high 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. Furthermore, we suggest extracted results, which can imply a discovery of new signaling pathways that is yet proven through experiments. This will be a good basis for research to discover new protein signaling pathways and unknown functions of established proteins.

▸ Keyword : 상호작용(PPI), 위치정보(Localization), 신호전달(Signaling Pathway), 경로예측(Predicting), 알고리즘(Algorithm), 가중치(Weight)

• 제1저자 : 조미경
• 투고일 : 2008. 12. 19, 심사일 : 2009. 1. 2, 게재확정일 : 2009. 3. 30.
* 이화여자대학교 박사 ** 이화여자대학교 교수

I. 서 론

생명공학 기술이 발전함에 따라 인류는 질병이 감소하고 수명이 연장되는 등 다양한 혜택을 누리고 있다. 하지만 아직도 인류에는 많은 질병으로 인해 고통 받고 있으며 이를 극복하기 위한 노력을 지속하고 있다. 이와 같은 질병을 극복하기 위해서 필수적으로 수행되어야 할 연구는 인간의 유전체에 대한 연구이며 이를 통해 다양한 유전적 질병의 원인을 파악할 수 있고 치료제 개발에 단서를 얻을 수 있다. 현재 기능 유전체학을 비롯한 유전체학 분야의 연구가 활발히 진행되고 있으며 이를 바탕으로 유전자들 간의 상호작용을 연구하고 있다. 향후 질병과 유전자간의 연관성을 파악하는데 신호전달 정보는 가장 핵심적인 정보가 될 것이며 이에 대한 연구가 전 세계적으로 활발히 진행되고 있다. 단백질 간 상호작용은 세포가 생명현상을 유지하기 위해 일어나는 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][2][3]. 종 수준 단백질 상호작용 데이터는 단백질을 노드, 상호작용을 에지로 표현하였을 때, 소수의 연결도 높은 노드들이 존재하는 Scale Free Network[4]의 특징을 지니며 이러한 연결도 높은 노드들 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[5]. 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. 신호 전달 과정을 통하여 세포 전체로 외부 자극을 확대하고 퍼뜨리는 것이 가능하다[6]. 신호 전달 경로는 전체 단백질 상호작용 네트워크의 부분 그래프에 해당한다.

세포 내 단백질 위치는 단백질의 기능을 유추할 수 있는 중요 정보의 하나로 인식되고 있다. 진핵생물의 경우 막으로 둘러싸인 세포 내 소기관들을 가지고 있으며 구분된 기관들은 서로 다른 기능을 수행하도록 분화되어 있다. 따라서 특정 단백질의 세포 내 위치를 아는 것은 단백질 연구에 있어 기본 정보가 된다. 또한 세포 내 단백질 위치는 물리학적인 경계가 되기도 한다. 이는 서로 만날 필요가 없는 단백질 간에 불합리한 접촉을 피할 수 있어 효율성을 높이는 한 방법이 된다. 이처럼 생물학적 여러 의미 관계 구조들을 전산학적 관점에서는 네트워크로 접근이 가능하며 이론을 배경화한 주제로 다양한 그래프 이론들이 적용되고 있다[4].

본 연구에서는 Yeast 단백질 간 상호작용 정보와 세포 내 위치 정보를 활용하여 세포막 단백질로부터 핵 단백질까지 신호 경위 경로를 찾아 내는 알고리즘에 대해 제안하고자 한다.

II. 신호전달 경로 추출 연구에 관한 선행연구

기존에 상호작용 예측 관련 연구에는 도메인 조합을 기반으로 한 복수개의 단백질 쌍들을 대상으로 순위 부여하는 연구[18] 또는 상호작용 네트워크에서 상동성 기반 바이오 콤플렉스를 예측하거나 연관속성 개념을 이용한 연구가 있다. 또는 단백질 상호작용 데이터 신뢰도 향상에 대한 연구나 그래프 클러스터링 알고리즘(Graph Clustering Algorithm)을 적용하여 단백질 Complex와 Cellular Process에 해당하는 기능적 모듈을 찾는 연구도 존재한다[7].

전체 그래프에서 다른 부분에 비해 연결도가 높은 지역[8]을 찾아 그 결과가 앞의 이론과 비슷하게 단백질 복합체거나 동일한 생물학적 과정에 참여하는 기능적 단위였음을 밝힌 연구도 있다. 그 외 유전체 수준 대사 경로 그래프 레이아웃을 위한 슈퍼 노드화를 이용한 연구도 있으며 데이터베이스 자동 갱신 및 관리에 관한 연구 그리고 단백질 기능에 대한 연구 등 다양한 연구들이 진행되었다. 사전 다른 정보를 사용하지 않고 그래프 알고리즘만을 적용하여 생물학적으로 의미 있는 부분을 추출 한 연구도 있다.

III. 단백질 상호작용 정보와 세포 내 위치정보 활용

단백질-단백질 상호작용은 최근 대단위 실험의 증가로 각광받고 있는 분야이다. 효모를 이용한 two-hybrid system을 대상 단백질과 상호 작용하는 단백질을 조사하는데 널리 사용되고 있는 방법으로 대량의 실험을 수행할 수 있는 장점이 있다. 이렇게 얻어진 정보는 현재 MIPS나 DIP등의 데이터베이스로 구축되고 있다. 단백질간의 상호작용에 대한 정보 축적을 바탕으로 대사(metabolism) 및 신호전달(signal transduction), 그리고 세포주기(cell cycle)에 대한 정보를 수집할 수 있게 되었다. KEGG를 비롯하여 EcoCyc, SPAD

등 다양한 종류의 세포내 경로정보에 대한 데이터베이스가 구축되었으며, 현재 활발한 연구개발 활동이 진행되고 있다.

3.1 단백질 상호작용 데이터베이스

단백질 간 존재하는 상호작용을 이해하는 것은 단백질의 기능 연구는 물론 시스템적인 생명 현상 이해를 가능케 한다. 이는 한 종류의 단백질이 단독으로 작용하기 보다는 단백질 간의 상호작용으로 기능을 수행하기 때문이다. 단백질 상호작용 데이터는 서열로 기능을 추정하기 힘든 단백질의 기능 연구를 가능하게 하며 상호작용 네트워크에서 토폴로지 정보만으로 새로운 단백질 복합체를 발견하는 등의 일을 가능하게 한다. 연구에 필요한 정보를 DIP과 MIPS등의 데이터베이스로 통합하여 재구축 하였다. 이러한 데이터베이스들은 서로 다른 Id Number와 체계를 가지고 있기 때문에 통합된 정보를 가져오는 데는 어려움이 따르므로 NCBI의 정보를 이용한다.

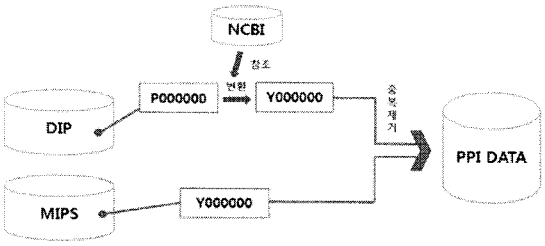


그림 1. DIP과 MIPS의 데이터베이스 통합
Fig 1. DIP and MIPS Database Integration

3.1.1 DIP

DIP (Database of Interacting Proteins) [9]은 Yeast Two-Hybrid나 면역침전 실험적 방법으로 얻은 단백질 간 상호작용 데이터를 제공한다. 일관성 단백질 상호작용 데이터베이스를 위해 Swiss-PROT과 다양한 리소스로부터 정보를 종합한다. DIP Data의 Core Subset은 상호작용 데이터 중 신뢰성이 높은 Protein-Protein Interaction 데이터를 제공하며 종별(Species-Specific) DIP 자료는 한 종에서 단백질들 간의 상호작용에 관련된 정보를 제공한다. 또한 〈DIP:nnnE〉형식의 고유 Identifier로 식별된다.

3.1.2 MIPS

MIPS(Munich Information Center of Sequences)[10]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 YPLD

YPLD(Yeast Protein Localization Database) [11]는 Yeast(saccharomyces cerevisiae) 단백질의 Subcellular Localization에 관련된 정보를 제공한다. SubCellular Localization 정보는 GFP로 표시된 단백질을 편광 현미경 혹은 공초점 현미경을 통해 얻어진다. YPL.db는 실험 조건, 이미지 정보, 단백질 데이터베이스에 대한 링크 등의 정보를 제공하며 세포 소기관에 대한 구조체 참조 데이터베이스도 포함한다.

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. 연구에서는 단백질 위치 정보의 중요성을 인식하고 세포 내 단백질 위치가 알려진 Budding Yeast를 대상으로 하여 단백질이 여러 위치에 존재할 수 있는 위치 정보를 이용한다.

선행연구에서 대사를 대상으로 신호전달 경로 추출 연구는 진행되어 왔으나 단백질 간 상호작용 신호전달 경로 추출 연구는 본 연구가 처음이며 특히 단백질 위치 정보를 이용한 연구 또한 최초로 발표하는 논문이기에 생물학자나 컴퓨터공학자 들에게 연구 기반을 제공할 수 있는 계기가 될 것으로 사료된다. 단백질 위치정보의 예를 살펴 보면 [그림 2]에서 YAL001C 단백질은 Cytoplasm과 Nucleus에 존재함을 알 수 있다.

	A	B	C	D	E	F
1	YAL001C	cytoplasm	nucleus			
2	YAL002W	endosome				
3	YAL005C	cytoplasm				
4	YAL007C	ER				
5	YAL029C	ambiguous	bud neck	cytoplasm	cell periphery	bud
6	YAL008W	mitochondrion				
7	YAL009W	nucleus				
8	YAL010C	mitochondrion				

그림 2. 단백질별 위치 정보
Fig 2. Information about Protein Position

선행연구로 밝혀진 세포내 단백질 위치정보는 [그림 3]에

서 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[14] 영역에 분포 시킨다.

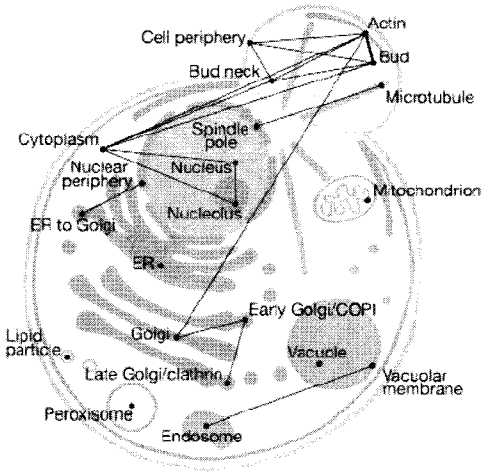


그림 3. 세포 내 21개의 단백질 위치 영역[33]
Fig 3. 21 of Protein Subcellular location
(출처: http://bric.postech.ac.kr/biotrend/scientist/ scientist_detail.php?nNum=282&nSID=1439)

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 [그림 4]에서 볼 수 있다.

Location	Number	Location	Number
nucleus	2,001	Golgi	180
cell periphery	209	microtubule	117
bud	88	bud neck	99
lipid particle	26	endosome	57
early Golgi	55	nuclear periphery	62
ER to Golgi	6	ER	553
cytoplasm	2,666	vacuolar membrane	95
nucleolus	207	spindle pole	81
actin	32	Vacuole	220
peroxisome	47	mitochondrion	939
late Golgi	46	7,786	

그림 4. 세포 내 위치별 단백질 분포표
Fig 4. protein Subcellular location Distribution table

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. 단백질 데이터 현황은 아래와 같다.

(단위: 개수)

DIP_DataBase	MIPS_DataBase	중복_Data 삭제	고유_Interaction
34,797	30,429	19,153	46,073

IV. 단백질 상호작용 정보와 세포 내 위치정보 활용 알고리즘

연구에서 제안한 위치기반 알고리즘을 적용한 시스템을 LocSPF(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 분류할 수 있다. "단백질 간 상호작용 경로 찾기"단계는 입력 값으로 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 적용하여 구한 단백질 상호작용 경로의 최종 값을 내림차순으로 정렬하여 동일한 순위 내에서 최대값 포함 노드를 추출하여 KEGG 신호전달 경로와 유사도를 비교분석 한다.

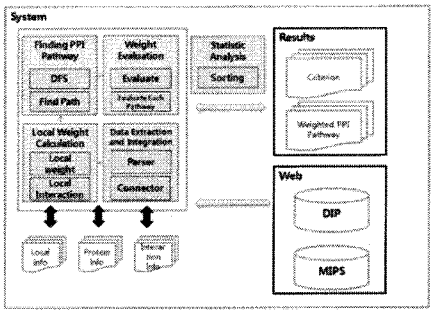


그림 5. LocSPF의 구성도
Fig 5. Configuration of LocSPF

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

단 계	세 부 사 항
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.

그림 6. 알고리즘의 단계별 세부사항
Fig 6. Details of the step-by-step algorithm

단백질 위치정보를 기반으로 하여 제안한 LocSPF 알고리즘 흐름도는 [그림 7]와 같다.

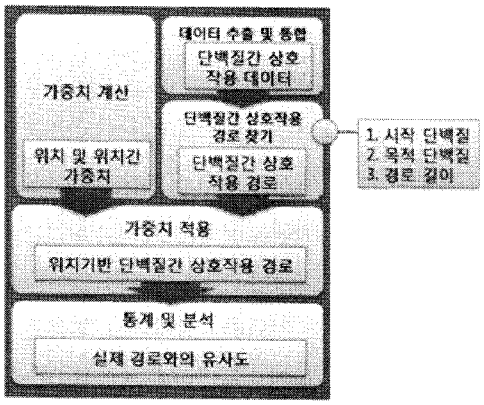


그림 7. LocSPF의 알고리즘 흐름도
Fig 7. LocSPF flow of the Algorithm

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 단백질간의 상호작용 여부를 나타내며 [그림 8]에서 인접행렬로 나타내었다. 예를 들어 행 A6셀에 데이터 YAL008W와 열 E1셀에 데이터 YAL007C와는 단백질 간 상호작용이 일어나고 있음을 나타낸다.

	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		

그림 8. 46,121개 단백질 상호작용 인접행렬 예
Fig 8. Protein Interaction Adjacency Matrix Ex.

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 시에 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. 이것을 수식으로 표현해 보자. 단백질1의 가중치는

$$\ln\left(\left(\left(n(\{i \mid i \in \text{proteins}, a \in i\text{의 위치}\}) \Big/ \sum_{i \in \text{proteins}} n(i\text{의 위치})\right) \times 100\right) \times 100\right)$$

에 의해 산출하고 단백질1과 단백질2의 상호작용 확률은

$$\left(\left(\left(n(\{i \mid i \in \text{proteins}, a \in i\text{의 위치}\}) \Big/ \sum_{i \in \text{proteins}} n(i\text{의 위치})\right) \times 100\right) \times 100\right)$$

에 의해 산출 한다

4.3 단백질 간 상호작용 경로 찾기

신호전달 경로는 세포내 세포막에 위치하는 단백질로부터 출발하여 핵에 존재하는 단백질까지 도달하는 신호전달 과정

이다. 실험에서는 이스트 단백질 MAPK의 "High osmolarity" 기능을 대표 모델로 하여 신호전달 경로를 예측하는 프로그램을 구현 하였다. 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 [그림 9]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.

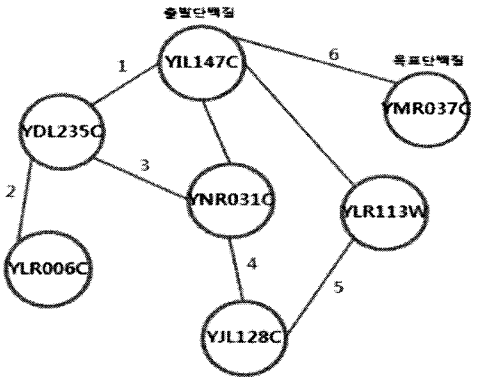


그림 9. DFS알고리즘을 수행 예
Fig 9. DFS Algorithm Practice Ex.

단백질 간 상호작용 알고리즘을 살펴보자. 화면상으로 받은 입력 값을 가지고 노드배열(Vertex Array)에서 시작 단백질의 인덱스를 찾아 스택(Stack)에 삽입한다. 스택이 최종적으로 비워질 때까지 [그림 10]과 같은 작업을 반복하면서 프로그램이 수행된다.

```
PPI Interaction Path Search
Class PathFind{
        InputDataReader reader
            //reading node & edge info.
                from input file
        Input start-protein, target-protein and
        node maximum length
        Save this input info. into the matrix.
        Dfs (start-protein, target-protein,
            node maximum length)
}

//by Depth-First Search, find the possible path.

1: Dfs(start-protein, target-protein,
      node maximum length)
2: {
3:      push (start-protein index, link_info);
4:      for all   e∈ existing adjacent nodes to
                the source-protein   do
5:      for (current stack depth<
            node-maximum length) do
6:      find adjacent vertex and push the value.
7:      if (finding the target protein) break;
8:      end if
9:      end for
10:     pop the discovered possible path.
11:     end for
12: }
```

그림 10. 단백질 간 상호작용 알고리즘
Fig 10. Protein Interaction Algorithm

[그림 10]의 슈도코드(Pseudo Code)에서 스택의 Top 포인터가 가리키는 가장 상위의 데이터 값인 인덱스에 해당하는 노드가 단말노드(Leaf Node)인 경우 즉, 인접한 노드(Vertex)가 없을 경우 방문배열(Vistied Array)의 방문 표시를 모두 초기화 한다. 그리고 스택에서 Top 포인터가 가리키는 가장 상위의 인덱스를 꺼낸 뒤 동시에 삭제하고(Pop) 다시 작업을 반복 수행한다. 스택의 Top 포인터가 가리키는 가장 상위의 데이터 값인 인덱스에 해당하는 노드가 단말노드가 아닐 경우 그 노드에 인접한 노드를 방문 하게 된다. 이때 인접한 노드들을 대상으로 첫째 목표 단백질 여부와 둘째 Circle을 만드는지 여부 그리고 셋째 경로의 깊이가 입력받은 최대깊이 값을 넘는지 체크하며 방문한다.

방문 프로세스는 다음과 같다. 첫째 목표 단백질의 경우 모든 스택 내용을 결과 파일에 저장 한다. 둘째 Circle을 만들거나 셋째 경로의 깊이가 입력받은 최대깊이를 넘을 때는 인접 노드의 개수를 하나 줄인다. 넷째 위의 세 가지 경우 모두에 해당 사항이 없을 경우 인접 노드의 개수를 하나 줄이고 스택에 그 인접 노드를 삽입하며 다시 작업을 반복 수행하게

된다. 이렇게 수행된 최종 결과 산출물은 [그림 11]에 제시된 바와 같이 조건에 적합한 모든 경로를 찾는 것이다.

	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	YMR058W	YDL029W	YHR030C	YPL089C
8	...							
9	(중략)							

그림 11. 단백질 간 상호작용 LocSPF 알고리즘 결과
Fig 11. Result of Interactions between Proteins LocSPF Algorithm

4.4 가중치 적용

단백질 간 상호작용이 막으로 구성된 위치에 존재하는 단백질은 그 위치 안에서 우선 상호작용이 일어날 수 있기 때문에 단백질 위치정보는 매우 중요하다.

그래서 단백질 위치 정보의 중요성을 인식하고 위치 정보를 활용 하였다. "단백질 간 상호작용 경로 찾기"단계에서 산출된 결과를 기반으로 모든 경로에 위치별 가중치와 상호작용 가중치를 이용한 평가함수를 적용하여 평가한다. 단백질 위치 정보를 활용한 평가 함수 적용 알고리즘은 [그림 13]에서 살펴 볼 수 있다. 또한 가중치를 적용한 두 가지 평가함수는 다음에 의해 구해진다.

$$\prod_{(i,j)\in \text{PPI Interaction Path}} (\text{Protein } i\text{와 Protein } j\text{의 Interaction Criterion 확률})$$

"단백질 간 상호작용 경로 찾기"를 평가하기 위해서는 패스를 이루는 단백질 앞 뒤 순서 간 모든 평가 값을 곱한다. 예를 들어 YLR332W-YOL109W-YBR069C-YLR372W -YMR058W-YDL029W-YHR030C-YPL089C를 평가해 보자. YLR332W-YOL109W, YOL109W-YBR069C, YBR069C-YLR372W, YLR372W-YMR058W, YMR058W-YDL029W, YDL029W-YHR030C, YHR030C-YPL089C 단백질 간 모든 상호작용 평가 값을 구하고 그 값들을 곱 한다. 단백질 간 상호작용의 값을 평가하기 위해 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 사용한다.

```
1    YKL175W : nucleus|vacuole|
2    YDR130C : nucleus|spindle pole|cytoplasm|
3    YDR473C : nucleus|
4    YDR098C : nucleus|cytoplasm|
5    YHR069C : nucleus|nucleolus|cytoplasm|
6    YDR239C : cytoplasm|
7    YDR533C : nucleus|er|cytoplasm|
...
5486 YMR124W : bud|cytoplasm|cell periphery|bud neck|
5487 YDR197W : mitochondrion|
5488 YJR072C : cytoplasm|
5489 YIL135C : cytoplasm|
5490 YDL124W : nucleus|cytoplasm|cell periphery|
5491 YDR198C : cytoplasm|
5492 YGR115C :
5493 YIL134W : nucleus|mitochondrion|
5494 YBR230C : mitochondrion|
5495 YAL029C : bud|mitochondrion|microtubule|cytoplasm|cell periphery|bud neck|
```

그림 12 단백질 5,495개의 위치정보
Fig 12. Protein 5,495 Position Information

```
Evaluation
Class Evaluation{
        Reading paths, criterion formed by two
        dimensional array from input file
        Input maximum length
        Evaluate (node maximum length)
}

//Evaluate the score of the path.

Evaluate(node maximum length)
{
for all   path ∈ paths   do
for   (protein1, protein2) ∈
        path until maximum length do
        Multiply score with value of
        EvaluationEach(protein1, protein2)
                        &
        Save into score variable
        end for
        end for
}

//Evaluate the score between two proteins

EvaluateEach(protein1, protein2)
{
        if protein1.locations = no location or
           protein2.locations = no location
        return  1.0f
        else if   protein1.locations ≠ no location and
                  protein2.locations ≠ no location
                  choose max value from criterion
                  each protein1.locations
                  and protein2.locations
                  return   max value
}
```

그림 13. 단백질 위치정보 이용 평가 LocSPF 알고리즘
Fig 13. Protein Position Information Evaluation LocSPF Algorithm

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 [그림 14]에서 제시 한다.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Y5843A2-YGL178W											
2												
3	YAL001C-YBR123C											
4	nucleus(1)-nucleus(1)mitochondrion(20)-nucleus(1)cytoplasm(5)-nucleus(1)]											
5	YAL002W-YKR026C											
6	endosome(13)-cytoplasm(5)]											
7	YAL002W-YLR149W											
8	endosome(13)-golgi(16)endosome(13)-vacuole(18)endosome(13)-vacuolar membrane(16)endosome(13)-endosome(13)]											
9	YAL002W-YLR396C											
10	endosome(13)-vacuole(19)endosome(13)-vacuolar membrane(16)endosome(13)-cytoplasm(5)]											
11	YAL002W-YMR231W											
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)]											

그림 14 위치 인터랙션의 모든 경우의 수
Fig 14. All Case of Position Interaction

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. 단백질I과 단백질J의 상호작용 가중치 확률 값(1)은

$$\max_{(a,b)\ \text{aProtein i}\text{로 최종 xProtein j}\text{로 최종}} (\text{Local a와 Local b의 Interaction 확률})$$

에 의해서 구하며 단백질I와 단백질J의 상호작용 가중치 확률 값(2)는

$$\max_{(a,b)\ \text{aProtein i}\text{로 최종 xProtein j}\text{로 최종}} (\text{Local a와 Local b의 Interaction 확률} \times \text{Local a의 위치 가중치} \times \text{Local b의 위치 가중치})$$

에 의해 산출 한다. 단백질1과 단백질2의 단백질 간 상호작용 값을 평가하기 위해서 단백질1과 단백질2의 단백질 위치 정보를 가져와 산출한다.

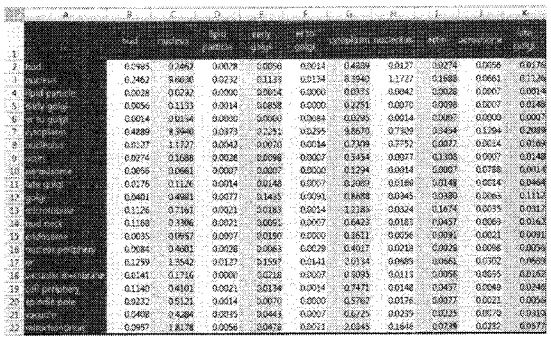


그림 15. 위치 상호작용 확률 값
Fig 15. Position Interaction Probability Values

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.

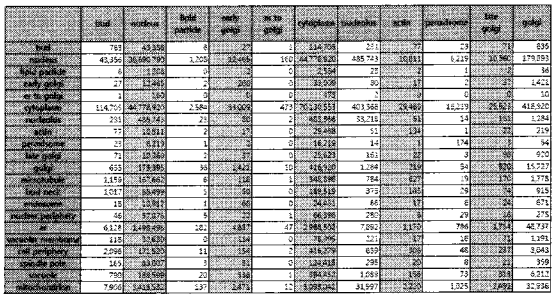


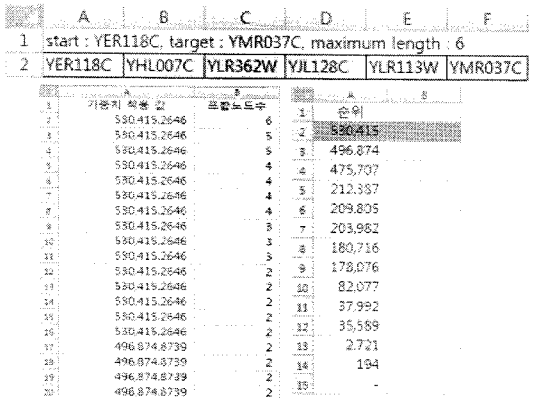
그림 16. A 위치의 가중치 * B 위치의 가중치 * A-B위치 인터랙션 확률 값
Fig 16. A Weight * B Weight * A-B Interaction Probability Value

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 최대 값의 포함노드를 포함하고 있는 값을 추출하여 KEGG신호전달 경로와의 유사도를 비교 분석 한다.

YER118C-YHL007C-YLR362W-YJL128C-YLR113W-YMR037C인 단백질 신호전달 경로 추출을 목표로 하여 시작 단백질을 YER118C로 목표 단백질을 YMR037C로 하

고 깊이를 6으로 하는 프로그램을 실행한 후 가중치 알고리즘을 적용하였다. 이때 산출된 결과 값을 내림차순 정렬하여 중복 값을 제거한 후 유일한 값에 대한 순위를 부여하고 최대 값을 선출한다. 이때 가중치 산출 결과 값 리스트에서 최대 값에 해당하는 값을 〔그림 17〕에서 살펴보면 최대 값 530,415.2646에 해당하는 값의 중복이 15개로 산출 되었다.



	A	B	C	D	E	F
1	start : YER118C, target : YMR037C, maximum length : 6					
2	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YMR037C

	A	B
1	가중치 적용 값	프랍노드수
2	530,415.2646	6
3	530,415.2646	5
4	530,415.2646	5
5	530,415.2646	4
6	530,415.2646	4
7	530,415.2646	4
8	530,415.2646	4
9	530,415.2646	3
10	530,415.2646	3
11	530,415.2646	3
12	530,415.2646	2
13	530,415.2646	2
14	530,415.2646	2
15	530,415.2646	2
16	530,415.2646	2
17	496,874.8739	2
18	496,874.8739	2
19	496,874.8739	2
20	496,874.8739	2

	A
1	순위
2	530,415
3	496,874
4	475,707
5	212,387
6	209,805
7	203,982
8	180,716
9	178,076
10	82,077
11	37,992
12	35,589
13	2,721
14	194
15	-

그림 17. 가중치 알고리즘 적용 산출 결과 및 순위
Fig 17. Apply Weight Algorithm Result and Rank

이때 중복 값이 산출된 원인으로는 단백질이 여러 위치 정보를 가지고 있을 때 평가 대상에서 가장 큰 값을 가지고 있는 위치 값을 선택하기 때문이다. 다시 말하면 Nucleus, Cytoplasm, Er, Mitochondrion 등에 단백질이 집중하고 있어 평가 대상에서 선택될 확률이 다른 위치 보다 많기 때문인 것으로 사료된다. 이후 최고 값을 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 실시 한다.

V. 적용 및 평가

5.1 MAPK

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. 세포막에서 신호를 전달받은 단백질들은 두 개가 하나의 단위로 작용하는 이합체(Dimer)가 됨으로써 활성부위가 노출된다. 이로써 Tyrosine Kinase로서 활성화되어 단백질들을 인산화 하는 일련의 반응을 유도한다. 이렇게 차례로 인산화 되는 신호화 과정을 전달받은 MAP Kinase는 세포질에서 핵으로 이동하며 세포가 필요로 하는 새로운 단백질을 만들 수 있도록 한다. 따라서 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 〔그림 18〕과 같다.

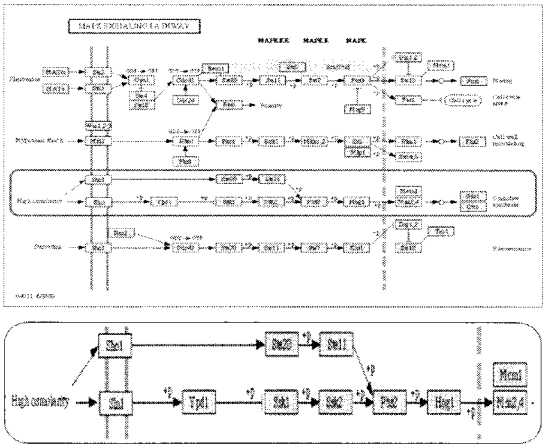
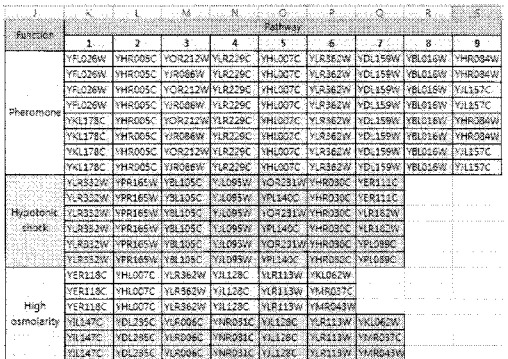


그림 18. MAPK 신호전달 경로(이스트)
Fig 18. MAPK Signaling Pathway (S.cerevisiae)

KEGG MAPMA깊이9를 제외한 전체 신호전달 경로 수행 결과표Signaling Pathway는 4개의 Function으로 구분할 수 있으며 각 해당 Pathway는 〔그림 19〕와 같이 총 21개의 Pathway로 구분할 수 있다.



Function	Pathway 1	2	3	4	5	6	7	8	9
Pheromone	YFL026W	YHR005C	YOR212W	YLR229C	YHL007C	YLR362W	YDL159W	YBL016W	YHR084W
	YFL026W	YHR005C	YIR086W	YLR229C	YHL007C	YLR362W	YDL159W	YBL016W	YHR084W
	YFL026W	YHR005C	YOR212W	YLR229C	YHL007C	YLR362W	YDL159W	YBL016W	YJL157C
	YFL026W	YHR005C	YIR086W	YLR229C	YHL007C	YLR362W	YDL159W	YBL016W	YJL157C
	YKL178C	YHR005C	YOR212W	YLR229C	YHL007C	YLR362W	YDL159W	YBL016W	YHR084W
	YKL178C	YHR005C	YIR086W	YLR229C	YHL007C	YLR362W	YDL159W	YBL016W	YHR084W
	YKL178C	YHR005C	YOR212W	YLR229C	YHL007C	YLR362W	YDL159W	YBL016W	YJL157C
	YKL178C	YHR005C	YIR086W	YLR229C	YHL007C	YLR362W	YDL159W	YBL016W	YJL157C
Hypotonic shock	YLR332W	YPR165W	YBL105C	YJL095W	YOR231W	YHR030C	YER111C		
	YLR332W	YPR165W	YBL105C	YJL095W	YPL140C	YHR030C	YER111C		
	YLR332W	YPR165W	YBL105C	YJL095W	YOR231W	YHR030C	YLR162W		
	YLR332W	YPR165W	YBL105C	YJL095W	YPL140C	YHR030C	YLR162W		
	YLR332W	YPR165W	YBL105C	YJL095W	YOR231W	YHR030C	YPL089C		
	YLR332W	YPR165W	YBL105C	YJL095W	YPL140C	YHR030C	YPL089C		
High osmolarity	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YKL062W			
	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YMR037C			
	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YMR043W			
	YJL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	YKL062W		
	YJL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	YMR037C		
	YJL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	YMR043W		
Starvation	YER118C	YLR229C	YHL007C	YLR362W	YDL159W	YGR040W			

그림 19. MAPK의 Function별 신호전달 경로
Fig 19. MAPK signaling path of each Function

각 신호전달 경로를 Pathway 길이로 분류하여 보면 단백

질을 9개 포함하는 경로인 경우가 8개, 단백질을 7개 포함하는 경로인 경우가 9개 그리고 단백질이 6개 포함하는 경로인 경우가 4개로 구분 된다. 이렇게 시작 단백질과 목적 단백질이 알려져 있는 데이터를 LocSPF의 입력 값으로 하여 방법1과 방법2로 가중치를 다르게 하여 수행하였다.

단백질이 9개를 포함하는 경로에 대해서도 실험을 하였으나 "단백질 간 상호작용 찾기"단계 산출시간이 Dell XPS710 (Intel coreTM2 CPU, 2.4GHz)에서 대략 240시간 이상 실행되었고 파일 사이즈도 하나의 Pathway가 78.15M이상 되기 때문에 현실적으로 효율성이 떨어진다고 판단하여 연구결과 분석에서 제외 하였다.

Function	Pathway 1	2	3	4	5	6	7	방법1_발견노드수	방법2_발견노드수
Hypotonic shock	YLR382W	YPR165W	YBL105C	YJL095W	YCR231W	YMR035C	YER111C	4	4
	YLR332W	YPR165W	YBL105C	YJL095W	YP_140C	YMR035C	YER111C	5	5
	YLR332W	YPR165W	YBL105C	YJL095W	YCR231W	YMR035C	YLR182W	4	5
	YLR332W	YPR165W	YBL105C	YJL095W	YP_140C	YMR035C	YLR182W	5	5
	YLR332W	YPR165W	YBL105C	YJL095W	YCR231W	YMR035C	YPL059C	2	2
	YLR332W	YPR165W	YBL105C	YJL095W	YP_140C	YMR035C	YPL059C	3	3
High osmolarity	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YKL062W		6	6
	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YMR037C		6	6
	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YMR043W		6	6
	YIL147C	YDL235C	YLR006C	YNR051C	YJL128C	YLR113W	YKL062W	7	7
	YIL147C	YDL235C	YLR006C	YNR051C	YJL128C	YLR113W	YMR037C	7	7
	YIL147C	YDL235C	YLR006C	YNR051C	YJL128C	YLR113W	YMR043W	7	7
Starvation	YER118C	YLR232W	YHL007C	YER062C	YDL159W	YGR042W		5	5

그림 20. MAPK의 신호전달 경로 찾기 성능 평가표
Fig 20. Find the MAPK's Signaling Pathway Performance Test

이미 밝혀진 상호작용 경로에 대해 LocSPF에서 수행한 결과를 5.2.1에서 제시한다. 5.2.2에서는 아직 실험을 통해 단백질 간 상호작용이 밝혀지지 않은 단백질이 신호전달 경로에 포함되어 있다고 가정하고 실험한 결과와 그 의미를 제시한다.

5.2 적용 및 결과분석

"가중치 적용"단계에서 구한 모든 상호작용 경로에 위치별 가중치와 상호작용 가중치로 평가를 실시하며 마지막 단계인 "통계 및 분석"단계에서 가중치를 적용하여 구한 단백질 상호작용 경로 값을 내림차순으로 정렬하여 동일한 순위 내에서 최대 포함노드를 추출하여 KEGG 신호전달 경로와의 유사도를 알아본다. LocSPF 알고리즘을 적용한 결과를 [그림 21]에서 제시 한다.

	A	B	C	D	E	F	G
1	KEGG_단백질수	실험_단백질수	발견_단백질수	발견1_Pathway수	발견1(%)	발견2_Pathway수	발견2(%)
2	6	6	6	3	75%	3	75%
3			5	1	25%	1	25%
4			합계	4	100%	4	100%
5	7	8	7	3	33%	3	33%
6			5	3	33%	4	44%
7			4	2	22%	1	11%
8			2	1	11%	1	11%
9			합계	9	100%	9	100%

그림 21. 깊이9를 제외한 전체 신호전달 경로 수행 결과표
Fig 21. All Signaling Pathway Performance Result Table Except Depth 9

KEGG MAPK Signaling Pathway의 경로 즉 [표5-1]에서 제시된 경로들 중 시작 단백질과 목표 단백질이 같은 경로들은 유일한 경로 하나로 통일하여 프로그램 실행에 소요된 시간표를 [그림 22]에서 제시한다.

	A	B	C	D	E	F	G	H
1	시작단백질	종료단백질	KEGG-Depth	LC System	Program_Start	Program_End	OutputFile_Size	소요시간
2	YER118C	YMR037C	6	5	2008-11-11 5:42 PM	2008-11-11 6:33 PM	2.12K	0시간 51분
3	YER118C	YKL062W	6	6	2008-11-11 5:45 PM	2008-11-11 6:43 PM	104K	0시간 58분
4	YER118C	YMR043W	6	6	2008-11-11 5:47 PM	2008-11-11 6:38 PM	54.3K	0시간 51분
5	YER118C	YER062W	6	6	2008-11-11 5:49 PM	2008-11-11 6:41 PM	175K	0시간 52분
6	YLR182W	YPL059C	7	8	2008-11-04 4:23 AM	2008-11-04 4:38 PM	6.9M	12시간 15분
7	YLR332W	YER111C	7	8	2008-11-04 4:31 AM	2008-11-04 6:26 PM	3.15m	13시간 55분
8	YLR332W	YLR182W	7	8	2008-11-04 4:27 AM	2008-11-04 4:39 PM	2.15m	12시간 12분
9	YIL147C	YMR037C	7	8	2008-11-04 4:15 AM	2008-11-06 10:06 PM	215K	17시간 51분
10	YIL147C	YKL062W	7	8	2008-11-04 4:16 AM	2008-11-05 9:55 AM	506K	5시간 39분
11	YIL147C	YMR043W	7	8	2008-11-04 4:21 AM	2008-11-05 10:54 AM	2.12m	6시간 33분

그림 22. 신호전달 경로 프로그램 실행 시간표
Fig 22. Signaling Pathway Program Practice Time

5.2.1 밝혀진 상호작용 경로 찾기

LocSPF의 성능을 알아보기 위해 깊이가 6인 상호작용 데이터에 대해 실험을 실시 하였다.

5.2.1.1 알고리즘 적용

단백질 간 상호작용의 데이터와 상호작용 관계를 표현한 인접행렬을 가지고 시작 단백질에서 목적 단백질까지의 깊이가 6인 신호전달 경로를 찾아 보았다. [표5-3]에서 알 수 있듯이 KEGG 신호전달 경로 4의 패스에 대해 실험한 결과 3개의 패스는 KEGG 신호전달 경로와 완전하게 일치하는 결과을 얻었으며 나머지 1개의 패스는 신호전달 경로에서는 KEGG 신호전달 경로에서 제시하는 6개 중 5개의 단백질을 포함하는 패스를 찾아 주었다. 이때 실험에는 "LocSPF의 알고리즘 흐름도" 중 "가중치 적용"단계에서 실행하는 두 가지 평가함수 알고리즘을 적용하였다. 첫 번째 방법은 상호작용 확률을 가중치로 적용한 것이고 두 번째 방법은 상호작용 확률뿐만 아니라 위치별 가중치를 적용한 것이다. 실험 결과로는 위에서 제시한 것과 같은 결과를 얻었으며 알고리즘 두 가지의 경우 모두 동일한 결과를 얻었다. 이때 사용한 알고리즘

은 첫 째는 A-B 위치 상호작용 확률 값 둘째는 A위치의 가중치 * B위치의 가중치 * A-B 위치 상호작용 확률 값이다.

5.2.1.2 밝혀진 경로 평가 분석

LocSPF 알고리즘을 적용한 성능을 분석하여 보면 "High osmolarity" 기능에 대해 경로1의 LocSPF_방법1에 대해서는 찾은 가중치를 내림차순으로 정렬하였을 때 상위3%이내 수준에서 또한 경로1의 LocSPF_방법2, 경로2와 경로3의 LocSPF_방법1과 LocSPF_방법2 모두 찾은 가중치를 내림차순으로 정렬하였을 때 Top스쿼어에서 KEGG에서 제공하는 신호전달 경로와 완전하게 동일한 패스를 찾고 있음을 〔그림 23〕에서 보여 준다.

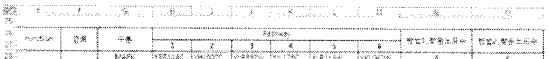


그림 23. 깊이6일 때 적용한 알고리즘 결과표
Fig23. Depth 6 Apply Algorithm Result

또한 "Starvation"기능에 대해서는 두가지 알고리즘의 경우 모두 KEGG에서 제공하는 신호전달 경로 노드6 패스에 대해 노드5가 일치하는 패스를 찾고 있다. 결과를 〔그림 24〕과 〔그림 25〕에 결과를 제시 하였다.

YER118C	YLR229C	YHL007C	YLR362W	YDL159W	YGR040W

	J	K	L	M	N	O	P	Q
1	1 node수	start : YER118C, target : YGR040W, maximum length : 6						
2	15576.61	5	YER118C	YER114C	YHL007C	YLR362W	YDL159W	YGR040W
3	15576.61	5	YER118C	YHL007C	YLR362W	YLR313C	YDL159W	YGR040W
4	15576.61	5	YER118C	YHL007C	YLR362W	YLL021W	YDL159W	YGR040W
5	15576.61	5	YER118C	YHL007C	YLR362W	YDR103W	YDL159W	YGR040W
6	15576.61	5	YER118C	YHL007C	YLR362W	YDL159W	YDR103W	YGR040W
7	15576.61	5	YER118C	YHL007C	YLR362W	YBL016W	YDL159W	YGR040W

그림 24. 방법1 결과 : 깊이6
Fig 24. Method1 Result : Depth 6

	J	K	L	M	N	O	P	Q
1	1 node수	start : YER118C, target : YGR040W, maximum length : 6						
2	530415.3	5	YER118C	YER114C	YHL007C	YLR362W	YDL159W	YGR040W
3	530415.3	5	YER118C	YHL007C	YLR362W	YLR313C	YDL159W	YGR040W
4	530415.3	5	YER118C	YHL007C	YLR362W	YLL021W	YDL159W	YGR040W
5	530415.3	5	YER118C	YHL007C	YLR362W	YDR103W	YDL159W	YGR040W
6	530415.3	5	YER118C	YHL007C	YLR362W	YDL159W	YDR103W	YGR040W
7	530415.3	5	YER118C	YHL007C	YLR362W	YBL016W	YDL159W	YGR040W

그림 25. 방법2 결과 : 깊이6
Fig 25. Method2 Result : Depth 6

KEGG에서 제공되는 코드와 실제 실험에서 사용된 MIPS 데이터베이스 제공코드가 서로 표현이 상이하다. 그래서 〔그림 26〕에서 코드 대조표를 제시한다.

KEGG	MIPS	KEGG	MIPS	KEGG	MIPS
BCK1	YJL095W	MLP1	YKR095W	SSK1	YLR006C
CDC42	YLR229C	MSN2	YMR037C	SSK2	YNR031C
FAR1	YJL157C	MSN4	YKL062W	STE11	YLR362W
FUS3	YBL016W	PBS2	YJL128C	STE12	YHR084W
GPA1	YHR005C	PKC1	YBL105C	STE18	YJR086W
HOG1	YLR113W	RAS2	YNL098C	STE2	YFL026W
KSS1	YGR040W	RHO1	YPR165W	STE20	YHL007C
MCM1	YMR043W	RLM1	YPL089C	STE3	YFL026W
MID2	YLR332W	SHO1	YER118C	STE4	YOR212W
MKK1	YOR231W	SLN1	YIL147C	STE7	YDL159W
MKK2	YPL140C	SLT2	YHR030C	YPD1	YDL235C

그림 26. MIPS와 KEGG 단백질 코드 대조표
Fig 26. MIPS and KEGG Protein Code Comparison

5.2.2 미지의 단백질 포함 상호작용 경로 예상하여 찾기

아직 실험을 통해 밝혀 지지 않은 미지의 단백질이 무수히 많음을 인지하고 그 단백질들이 앞으로 밝혀질 것을 예상하여 본 실험에서 적용해 보았다. KEGG에서 제공하는 신호전달 경로 패스가 7일 경우 본 실험에서는 8로 하여 미지의 단백질을 실제 KEGG의 신호전달 경로 패스에 삽입하였을 경우에도 KEGG의 신호전달 경로 패스를 동일하게 찾아 주는지 실험하여 그 유사도를 제시 한다.

5.2.2.1 알고리즘 적용

KEGG MAPK의 "High osmolarity" 기능에서 제공하는 단백질 신호전달 경로 깊이7에 대해 깊이8로 하여 실험을 통해 아직 밝혀지지 않은 미지의 단백질이 패스 중간에 삽입 되어 있음을 전제하고 실험을 실시하였다. 방법과 과정은 앞의 깊이6일 때와 동일하여 생략하며 결과만 제시한다. LocSPF 알고리즘 실험 결과는 〔그림 27〕과 같다.

5.2.2.2 밝혀지지 않은 경로 평가 분석

"High osmolarity"기능에서 제공하는 신호전달 경로 패스에 미지의 단백질이 포함되어 있음을 전제로 하여 실험을 한 결과 KEGG에서 제공하는 실제 신호전달 경로 패스와 완전히 일치하게 결과를 찾아 주어 완성도 높은 알고리즘임을 증명하여 주었다. 고성능 알고리즘을 입증할 수 있었던 배경이 무엇인지 살펴보자. 앞의 실험 깊이6인 경우와 동일하게 단백질 위치정보의 중요성이라 할 수 있다. 〔그림 28〕에서 살펴보면 단백질이 Cytoplasm, Nucleus에 집중되어 있음을 알 수 있다. 다시 말해서 단백질 위치정보가 여러 가지일 경우 단백질 위치 값이 최고인 값을 선택하는 알고리즘의 특성상 여러 단백질 위치의 후보군 중에 대상자로 선택될 확률이 매우 높다는 것이다.

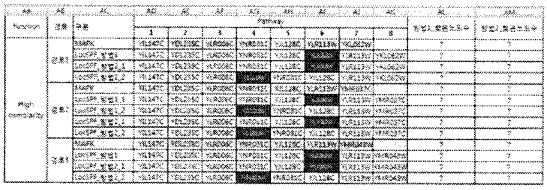


그림 27. 깊이8일 경우 산출결과표
Fig 27. Depth8 Result Table

	A	B	C	G	H	I	J
1		Kegg_Path	mips_code	Location1	Location2	Location3	Location4
261	260	SLN1	YIL147C	cell periphery	0	0	0
262	261	YPD1	YDL235C	cytoplasm	nucleus	0	0
263	262	SSK1	YLR006C	cytoplasm	0	0	0
264	263	SSK2	YNR031C	cytoplasm	0	0	0
265	264	PBS2	YJL128C	cytoplasm	0	0	0
266	265	HOG1	YLR113W	cytoplasm	0	0	0
267	266	MSN2	YMR037C	cytoplasm	nucleus	0	0

그림 28. 시작단백질:SLN1 종료단백질:MSN2
깊이:7인 단백질 위치정보
Fig 28. StartProtein : SLN1 TargerProtein : MSN2 Depth : 7 Protein Position Information

5.3 결과 분석

첫 번째, 알고리즘의 기반이 되는 단백질 위치정보의 중요도가 매우 높음을 결과를 통해 알 수 있었다. 알고리즘의 방법1과 방법2는 단백질 위치 정보의 값을 기반으로 작성 되었다. 이때 실험에서 사용된 단백질들의 위치를 살펴보면 Nucleus와 Cytoplasm에 많이 위치하는 정보를 가지고 있다. 알고리즘 수행 중 단백질이 여러 위치 정보를 가지고 있을 때 여러 경우의 평가 값 중 가장 큰 값을 선택하여 프로그램을 실행하게 한다. Nucleus와 Cytoplasm의 단백질 위치 정보의 값은 매우 큰 값을 가진다. 즉 실제 가중치에 사용 될 위치 값으로 선택될 확률이 다른 위치정보를 가진 단백질 보다 매우 높다는 의미이다. 두 번째, 알고리즘 방법1과 방법2는 매우 유사한 알고리즘 이라고 판단할 수 있다. 다시 말해서 방법1이나 방법2 알고리즘 모두 단백질 위치 정보 기반에서 착안된 알고리즘이라는 것을 증명한 것이다. [그림 21]에서 제시해 주는 결과를 살펴보면 깊이6이나 깊이8로 신호전달 경로 찾기 프로그램을 실행했을 때 방법1이나 방법2 알고리즘 수행 결과의 유사도가 비슷함을 결과를 통해 알 수 있다.

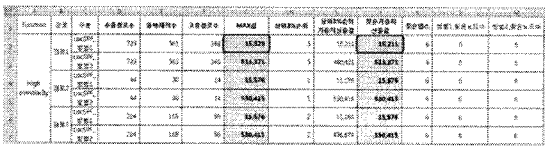


그림 29. 깊이6일 경우 평가 값 현황
Fig 29. Depth 6 Case Evaluation value State

[그림 29]에서 보면 깊이6일 경우 신호전달 경로에 평가 값 적용 현황을 제시 한다. 이때 경로1에서 LocSPF_방법1 알고리즘 결과는 가중치 적용 결과 값을 내림차순으로 정렬하여 상위 3%수준에서 신호전달 가능 경로 패스를 정확하게 찾아 주는 것을 [그림 30]에서 알 수 있다. 그 외 5가지에 대해 LocSPF 알고리즘을 적용한 결과는 Top스퀴어에서 완전한 패스를 모두 찾아 알고리즘의 우수성을 입증해 주고 있다.

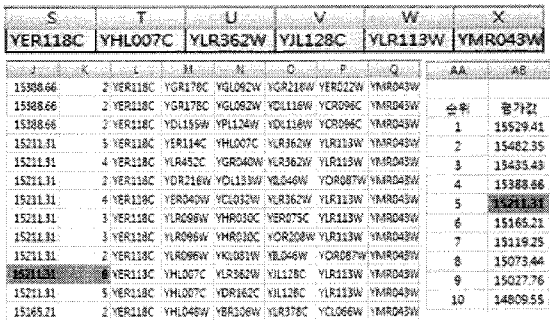


그림 30. 깊이6의 LocSPF_방법1 평가 값 및 위치
Fig 30. Depth 6 LocSPF_Method1 Evaluation and Position

깊이8일 경우 신호전달 경로에 평가 값 적용 현황을 [그림 31]에서 제시 한다. LocSPF 알고리즘을 적용한 결과는 Top스퀴어에서 모두 일치하는 패스를 찾아 주어 알고리즘의 고성능을 증명해 주고 있다.

그림 31. 깊이8일 경우 평가 값 현황
Fig 31. Depth 8 Case Evaluation value State

"MAPK High osmolarity 기능 패스 도식"에서는 총6개의 신호전달 경로를 병합해 놓은 트리 구조 데이터로 생물학자들이 실험을 통해 신호전달 경로를 밝혀 놓은 경로 이다. 실험에서는 이것을 표본으로 비교분석 하고자 한다. 표본 트리 구조를 크게 왼쪽 그룹 TREE와 오른쪽 그룹 TREE로 분리하여 증명 하겠다. LocSPF의 알고리즘을 적용하여 실행한 결과 "MAPK High osmolarity기능 패스 도식"과 완전히 일치하는 결과를 얻었다.

KEGG에서 제공하는 MAPK을 기준으로 "High osmolarity"기능을 모델로 하여 "가중치 적용"단계를 거쳐 최

종 산출 결과를 가지고 도식으로 〔그림 32〕에 보여 주고 있다. 이때 연구의 모델인 "High osmolarity" 기능의 표본 6개의 Pathway를 병합하여 실제 KEGG 신호전달 경로에서 제시해준 생물학적 의미 있는 신호전달 경로와 동일한 구조를 형성하고 알고리즘의 정확도를 측정하여 보았으나 앞서 논의한 것처럼 동일한 결과를 얻을 수 있었다.

〔그림 33〕과 〔그림 34〕를 통해 보면 지향하는 최고 높은 깊이 값에 대응하는 가중치 부가 산출 값 또한 최고 높은 값에 존재하고 있음을 증명해 보여 주고 있다. 이 결과를 통해 알 수 있듯이 연구에서 제안한 알고리즘의 정확도가 매우 높음을 증명해 준다.

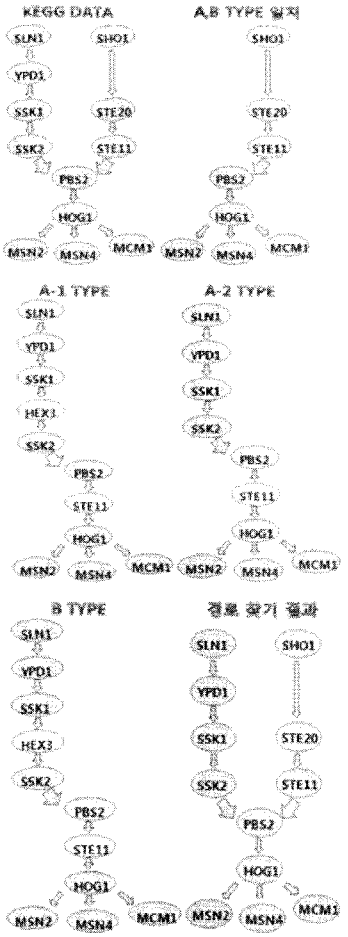


그림 32. MAPK High osmolarity 기능 패스 도식
Fig 32. MAPK High osmolarity Function Figure

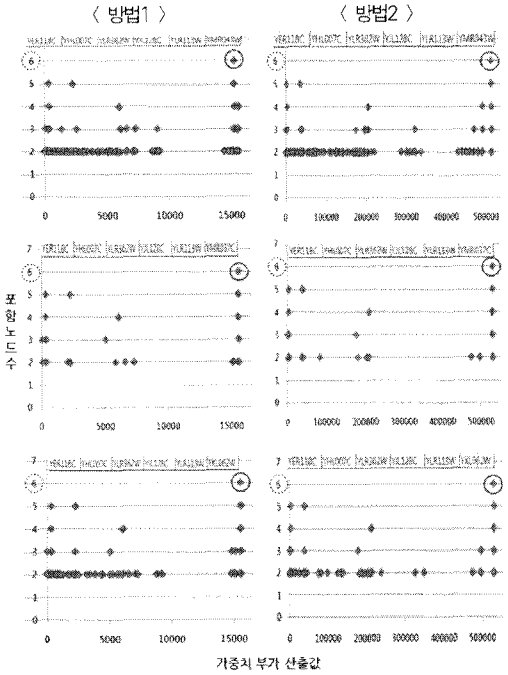


그림 33. KEGG-깊이6, 목표-깊이6, 방법1-방법2
Fig 33. KEGG-Depth6, Targer-Depth6, Method1-Method2

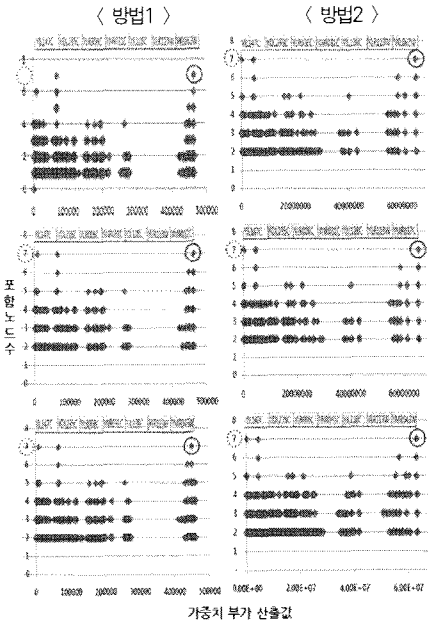


그림 34. KEGG-깊이7, 목표-깊이8, 방법1-방법2
Fig 34. KEGG-Depth7, Targer-Depth8, Method1-Method2

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. 이때 가장 밀접하게 연관된 발현 프로파일 데이터집합을 이용하여 새로운 신호 전달 경로를 예측하고 상호 작용하는 단백질들의 작은 클러스터를 찾는 것이 가능하다고 밝힌다[16]. 그러나 이 방법은 상호작용과 발현 데이터 사이의 연관성이 존재함을 전제한 것인데 Gerstein등에 의해 밝혀진 상호작용과 발현데이터 사이의 연관성은 단백질들이 세포막에서 핵까지 전달되는 과정에서 지속적으로 존재하는 퍼머넌트 콤플렉스(permanent complex)와 전달과정 중 잠시 존재했다가 사라지는 트랜지언트 콤플렉스(transient complex) 중 퍼머넌트 콤플렉스(permanent complex)에만 존재함이 알려졌다[17]. 그러나 실제 신호전달 상호작용에는 잠시 존재했다가 사라지는 트랜지언트 콤플렉스(transient complex)에서도 상호작용이 일어날 확률이 높기 때문에 좋은 성능을 보여주지 못한다. 그래서 Yeast 상호작용 데이터 46,121개의 단백질 정보와 약 5,000여 개의 Yeast(S. cerevisiae) 위치정보를 이용하여 신호 전달 경로 추출 및 예측을 위한 새로운 알고리즘을 제안 하였다. 상호작용을 기반으로 출발점을 세포막 단백질로 하고 도착점을 핵에 있는 단백질로 하였으며, 신호전달 경로를 추출할 때 단백질 위치 정보를 활용하여 임의의 두 단백질 간 다양하게 작용하는 상호작용 빈도수와 단백질 위치에 따른 단백질 빈도수를 가중치로 부여하는 방식으로 신호전달 경로를 추출 하였다. 시뮬레이션 결과를 통하여, KEGG에서 제공하는 MAPK High osmolarity 기능 Pathway 경로를 정확도 높게 찾아 증명하였다. 본 논문에서는 단백질의 위치정보를 사용해서 깊이9를 제외한 13가지 신호전달 경로 패스에 대해 단백질 상호작용 경로 추출에 적용한 결과 의미 있는 결과가 도출 되었고 그 중 특히 6개의 상호작용 경로 추출에서 가장 높은 값을 찾아 제안한 알고리즘의 우수한 성능을 나타내었고 경로 찾기에서 단백질 위치정보의 중요성을 입증하였다. 시뮬레이션 대상이었던 13가지 중 특히 6가지 신호전달 경로 찾기에서 완전히 일치하는 경로를 찾아 본 논문에서 제안한 알고리즘의 중요성을 입증하였다.

6개를 제외한 나머지 경로에 대해서는 위치정보가 Top스코어을 기록하지 못한 이유는 향후 고찰할 연구내용이지만 현재의 시점에서 그 이유를 유추해 보자면 단백질 상호작용이 같은 위치에 있을 때 서로 상호작용이 일어날 확률이 높고 거리가 가까운 곳에서 상호작용이 일어날 확률이 높지만 생물학적 관점에서 접근해 보면 단백질 상호작용은 한곳에 정지하며 단백질 간 상호작용이 일어나기도 하지만 때로는 이동하면도 상호작용이 일어나기 때문에 신호전달 과정 중 잠시 존재했다가 사라지는 트랜지언트 콤플렉스(transient complex)에서도 상호작용이 일어나는 것이다. LocSPF 알고리즘은 단백질 위치정보 기반 알고리즘이기 때문에 미세한 트랜지언트 콤플렉스를 세밀하게 반영 해 주지는 못한다. 위치정보를 이용한 신호전달 경로 찾기로서는 최초로 발표하는 연구 결과이다. 많은 생물학자나 컴퓨터공학자들 사이에서 집중이 되는 연구이기는 하지만 아직까지 활발한 연구가 되지 못하는 이유 중에 하나는 바이오인포매틱스 연구자들도 많지 않았고 앞에서 제기한 트랜지언트 콤플렉스 문제를 어떻게 해결할 것인지에 대한 대안연구가 진행 중 이기 때문인 것으로 사료된다. 차기 연구에서는 아직 밝혀 지지 않은 단백질 상호작용을 고려하여 제안한 알고리즘의 성과가 생물학 적으로 어떤 의미가 있는지를 분석하여 신호전달 예측 시스템의 정확도를 높이기 위한 노력이 필요할 것으로 사료된다.

참고문헌

[1] Schikowski B, Uetz P and Fields S, "A network of Protein-Protein Interaction in Yeast," Nat Biotechnol, 18:1257-1261, 2000

[2] Uets P,Giot L, Cagney G, Mansfield TA, Jusdson RS, Knight JR, Lock-shon D, Narayan V, Srinivasan M and Pochart P, "A comprehensive analysis of Protein-Protein Interactions in Saccharomyces Cerevisiae,"

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

[4] Reuven Cohen, Shlomo Havlin, "Scale-Free Networks Are Ultrasmall," Vol. 90, 90-94, 2003

[5] lbert-Laszlo arabasi, Zoltan N. Oltvai, "Understanding the Cell's Funcrional Organization," Nature, vol. 5, 101-103, 2004

[6] Silvia D. M. Santos, Peter J. Verveer, Philippe I. H. Bastiaens, "Growth Factor-Induced MAPK Network Topology Shapes Erk Sponse Determining PC-12 Cell Fate," Nature, Vol. 9,

pp. 324-330, 2007

[7] Jose B. Pereira-Leal , Anton J. Enright , Christos A. Ouzounis, "Detection of Functional Modules from Protein Interaction Networks", Vol. 54, 49-57, 2004

[8] Victor Spirin , Leonid A. Mirny, "Protein Complexes and Functional Modules in Molecular Networks", Vol. 100, No. 21, pp. 12123-12128, 2003

[9] DIP, *http://dip.doe-mbi.ucla.edu*

[10] MIPS, *http://mips.gsf.de*

[11] YPLD, *http://ypl.uni-graz.at/pages/home.html*

[12] Martin steffen, Allegra Petti, John Aach, Patrik D'haeseleer,George Church, "Automated Modeling of Signal Transduction Networks," BioMed Central, Nov. 2002

[13] KEGG, *http://www.genome.jp/kegg/*

[14] Won-ki Huh, James V.Falvo et al., "Global Anlaysis of Protein Localization in Budding Yeast," Nature, 2003

[15] NCBI, *http://www.ncbi.nlm.nih.gov*

[16] Martin steffen, Allegra Petti, John Aach, Patrik D'haeseleer,George Church, "Automated modeling of signal transduction networks," BioMed Central, Nov. 2002

[17] Ronald Jansen, Dov Greenbaum, Mark Gerstein, "Relating Whole-Genome Expression Data with Protein-Protein Interactions," Genome Research, 2001

[18] 조미경, 김민경, 박현석, "Protein-Protein Interaction에 세포 내 위치 정보를 활용한 단백질 신호전달 경로 추출 알고리즘 연구," 한국컴퓨터정보학회 동계학술발표대회 논문집, 제2권, 제1호, 77~84쪽, 2008년 12월.

저 자 소 개

조 미 경

소속 : 이화여자대학교

학위 : 이화여자대학교 공과대학 공학박사

관심분야 : Natural Language, Processing, Bioinformatics, Genome Informatics

김 민 경

소속 : 이화여자대학교

학위 : 서울대학교 의과대학 의학박사

관심분야 : 바이오인포매틱스(단백질 상호작용)

박 현 석

소속 : 이화여자대학교부교수

컴퓨터공학전공주임교수

컴퓨터 전자공학부장

과학기술경영연계전공주임교수

학위 : Ph.D. (영)Univ. of Cambridge (Computer Science and Engineering)

관심분야 :Natural Language Processing Bioinformatics, Genome Informatics, Human Computer Interaction