Signal transduction pathway extraction by information of protein-protein interaction and location

단백질 상호작용 정보와 위치정보를 활용한 신호 전달 경로추출

Eunha Kim [1*], Min Kyung Kim [1], Hyun Seok Park [1]

[1] Department of Computer Science Engineering, Ewha Womans university, Seoul, Korea

*To whom correspondence should be addressed. E -mail: milky78@ewha.ac.kr

Abstract

세포 내에서 일어나는 신호 전달 과정은 단백질간의 상호작용을 통해 수행되고 조절된다. 단백질 상호작용 데이터를 활용하여 수행된 연구로는 단백질의 기능을 유추하거나 전체 네트워크 중 다른 지역보다 더 조밀한 상호작용을 추출하여 complex 혹은 pathway를 발견하고 진화 과정을 이해하는 바탕이 되고 있다.

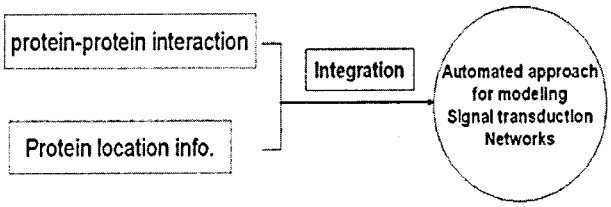
본 연구에서는 신호 전달 경로에 대한 사전 정보 없이 yeast 상호작용 정보와 녹색형광단백질(GFP)을 이용하여 밝혀진 4000여 개의 yeast 단백질 위치 분포 data를 이용하여 신호전달 경로를 찾는 방법을 시도했다. 기존 연구에 의해 밝혀진 yeast 내의 단백질 위치 분포 결과를 보면 21개의 category에 대해 각 단백질 상호작용 분포가 다양하게 나타나고, 특정 위치에서 상호작용 빈도수가 현저히 크다는 것을 알 수 있다. 특히 두 단백질이 같은 장소에 있을 경우 상호작용 확률이 높으며, 세포 내 소기관 사이에도 상호작용의 정도가 다양함이 알려져 있다. 따라서 이러한 분포상의 특성을 고려하여 상호작용을 기반으로 하여 세포막 단백질을 출발점으로, 핵에 있는 단백질을 도착점으로 잡고, 그 사이에 존재하는 다양한 가능 경로 중에서 단백질의 위치 정보를 가중치로 사용하여 그 중 최대 가능 경로를 찾도록 구현하였다.

이와 같은 pathway 모델링은 기존에 밝혀진 pathway와의 비교를 통해 알려지지 않은 새로운 경로를 발견하고, 이전에 경로에 참여하지 않은 단백질들을 발견할 수 있고, 이미 알려진 단백질들의 새로운 기능들에 대해서도 추론할 수 있을 것이라 기대한다.

Introduction

개체의 외부 자극에 대한 반응의 조절 및 항상성(homeostasis)에 필수적인 요건이라 할 수 있는 세포 내 신호 전달 과정은 세포 내 단백질간의 상호작용을 통해 수행된다. 예를 들어, MAPK signaling pathway의 경우 세포막에 존재하는 특이적인 수용체에 의해 인식되는 것을 기점으로 세포막으로부터 핵까지 단백질간의 상호작용의 신호 전달 과정을 통해 pheromone response, filamentous growth, maintenance of cell-wall integrity, high osmolarity 등과 같은 작용이 발생 된다. 이러한 신호 전달 과정은 세포 내 단백질간의 정교한 신호전달 네트워크를 기반으로 이루어 지게 되는데, 신호전달 네트워크를 구성하는 단백질들의 상호작용 데이터는 현재 high-throughput yeast two- hybrid와 같은 실험 등을 통해 종 수준에서의 상호작용이 밝혀질 수 있게 되었다. 단백질과 단백질간의 상호작용 데이터는 단백질의 기능을 유추하고[1,2] 신호전달 과정을 구성하는 전체 그래프 중 일부 부분 그래프를 추출하여 이전의 경로에 포함되지 않던 단백질을 새로이 신호전달 경로에 포함시키거나 새로운 단백질 복합체 등을 발견하는 것이 가능하다[3].

또한, 단백질 상호 작용을 이용하여 신호전달 경로를 구축하고자 하는 연구가 활발히 이루어지고 있는데, gene expression data를 이용하여 의미 있는 상호작용 데이터를 발견하여 사전 정보 없이 신호 전달 경로를 자동적으로 구축하고자 하는 시도 등이 이루어지기도



[Fig.1] modeling signal transduction networks by

하였다.[6] 그러나 이 방법은 상호작용과 발현 데이터 사이의 연관성이 존재함을 전제한 것인데 Gerstein 등에 의해 상호작용과 발현데이터 사이의 연관성은 상호작용을 이루는 permanent complex와 transient complex 중 permanent complex에만 존재함이 알려져 있다[5].

본 연구에서는 yeast를 대상으로 하여 [Fig.1] 에서 보는 바와 같이 상호 작용 데이터들 속에서 신호 전달 경로와의 연관성을 적용시키기 위해 세포 내 단백질들의 위치정보 데이터를 활용하였다. 임의의 두 노드 간 존재하는 다양한 경로 중 단백질 위치에 따른 상호작용 빈도를 가중치로 두어 가중치가 높은 경로를 선택하는 방식으로 경로를 추출하고, 이것이 알려진 신호전달 경로를 추출하는지 비교하고자 한다. 단백질 상호작용을 고려하기 위해 MIPS의 두 종류의 단백질 상호작용 데이터를 사용하였는데 하나는 10467개의 상호작용 셋으로 구성되었고, 또 하나의 경우는 15400개의 상호작용 셋으로 구성되어 있다. Yeast 단백질의 위치 정보는 nucleus, cell periphery, bud, lipid particle, early Golgi, ER to Golgi, cytoplasm, nucleolus, actin, peroxisome, late Golgi, Golgi, microtubule, bud neck, endosome,nuclear periphery,ER ,vacuolar membrane, spindle pole, vacuole,

[1] This work is supported by Ministry of Information and Communication

mitochondrion 등 21개의 위치로 분류된 4000여 개의 위치 데이터를 사용하였다.[4] 추출하고자 한 신호전달 경로는 yeast에서의 pheromone response, filamentous growth, cell wall integrity 기능과 연관되는 MAPKinase(mitogen-activated protein kinases) pathway를 대상으로 하였다.

Methods

출발점과 도착점을 각각 세포막과 핵에 존재하는 임의의 단백질로 선택하여, 출발점과 도착점 사이에 존재할 수 있는 모든 경로들을 우선적으로 찾도록 한다. 찾아낸 모든 가능 경로들에 세포 내 단백질의 위치정보와 위치에 따른 단백질간의 상호작용 빈도수를 이용하여 가중치를 적용하고, 각 경로 별 구해진 가중치 값에 따라 정렬된 경로들 중 신호전달 경로가 세포 내 위치의 적절한 순서에 따른 전달 경로일 확률이 높은 것을 선별하여 신호전달 패스웨이를 구성하도록 했다.

가능 경로 패스 찾기

사용한 위치정보 데이터

nucleus, cell periphery, bud, lipid particle, early Golgi, ER to Golgi, cytoplasm, nucleolus, actin, peroxisome, late Golgi, Golgi, microtubule, bud neck, endosome,nuclear periphery,ER ,vacuolar membrane, spindle pole, vacuole, mitochondrion 등 21개의 위치로 분류된 4100여 개의 위치 데이터를 사용하였다.[4] (http://yeastgfp.ucsf.edu) 이 데이터에 포함된 단백질들의 위치 별 분포를 살펴보면 표[1]과 같고, 이 중 cell periphery에서 nucleus까지의 신호전달 경로라고 가정했다. 이에 해당하는 단백질은 각각 1455개, 160개 이며 그 중 MAP kinase에 연관된 단백질로 한정하여 cell periphery에 속하는 Mid2, Wsc1, Wsc3, Sln1, Sho1 등의 단백질과 nucleus에 속하는 Ste12, Rlm1, Swi4, Swi6, Mcm1 등의 단백질을 출발점과 도착점으로 설정하였다.

사용한 상호작용 데이터

두 지점 사이의 가능 경로를 구하기 위해 MIPS의 10500여 개의 상호작용 데이터를 이용했다. 이 상호작용 데이터들은 4300여 개의 단백질로 구성되며 실질적으로 가능 경로 찾기를 위한 출발점과 도착점을 입력받으면 이 10500여 개의 상호 작용 데이터들을 이용하여 연결성 있는 모든 가능 경로들을 찾게 된다. 수행해본 세포막 단백질로부터 핵단백질까지의 경로들은 Ras2-Ste12, Ste2-Ste12, Mid2-Swi6, Mid2-Swi4, Mid2-Rlm1, Sho1-Ste12, Wsc3-Rlm1 등이 있다.

Location	Number	Location	Number
nucleus	1455	Golgi	43
cell periphery	160	microtubule	20
bud	73	bud neck	98
lipid particle	23	endosome	49
early Golgi	55	nuclear periphery	61
ER to Golgi	6	ER	296
cytoplasm	1821	vacuolar membrane	60
nucleolus	164	spindle pole	66
actin	32	Vacuole	163
peroxisome	21	mitochondrion	527
late Golgi	46		

[표] 1 세포 내 단백질의 위치 별 분포

가능 경로 패스의 길이

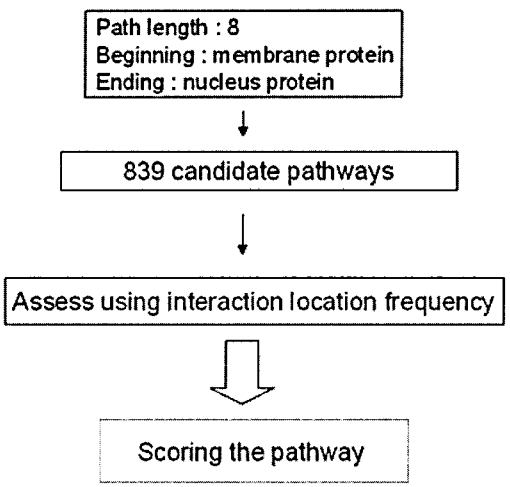
cell periphery에 속하는 Mid2, Wsc1, Wsc3, Sln1, Sho1 등의 단백질과 nucleus에 속하는 Ste12, Rlm1, Swi4, Swi6, Mcm1 등의 단백질을 출발점과 도착점으로 설정하였고, 그 가운데 존재할 수 있는 노드의 수는 6으로 한정하였으며 따라서 전체 선형 패스의 노드수는 8이 되게 하였다.

이는 이전의 연구인 NetSearch 알고리즘 에서 찾아낸 최적 선형 노드 수를 적용한 것이다[6]. 후에 일련의 가중치 값들을 처리한 뒤 시작 노드와 도착 노드가 같은 경우 병합을 하도록 한다. 따라서 한 패스웨이에 속하는 노드 수가 증가하게 되었고 단순한 선형 패스형태가 아닌 실재 패스웨이와 유사한 형태를 가지도록 하였다

가능 경로 패스 찾기

MAP kinase에 연관된 단백질들을 출발점과 도착점을 잡고 깊이 우선탐색 알고리즘(Depth First Search algorithm)을 이용하여 두 단백질 사이의 모든 가능 경로를 우선적으로 검색하였다. 즉, 경로길이 8을 갖는 세포막으로부터 핵까지 도달되는 단백질들의 모든 가능 경로를 검색하였다.

Mid2-Rlm1의 경우의 과정을 [fig.2]에서 보여주고 있다.



[fig. 2] scoring process(Mid2-Rlm1)

protein의 세포 내 위치정보의 활용

가능 경로 패스 가중치 부여

본 연구에서 사용한 가중치는 상호작용 하는 두 단백질의 위치에 따른 상호작용 확률 값을 사용한다. 단백질의 위치 정보로 yeast location 데이터를 보면 21개의 세포 내 위치에 대해 각 단백질 상호작용 분포가 다양하게 나타나고, 특정 위치에서 상호작용 빈도수가 현저히 크다는 것을 알 수 있다. 특히 두 단백질이 같은 장소에 있을 경우 상호작용 확률이 높으며, 세포 내 소기관 사이에도 상호작용의 정도가 다양함이 알려져 있다. 예를 들어,vacuolar membrane

에 위치하는 단백질과 Golgi에 위치하는 단백질의 상호작용 빈도수를 1로 보았을 때, 같은 vacuolar membrane에 위치하는 두 단백질의 상호작용 빈도수는 14로 더 높은 확률의 상호작용 가능성을 보인다[4].

이렇게 다양성을 갖는 위치 별 상호작용 확률 값을 이용하여 각 패스에 가중치를 부여하였다. 하나의 경로를 이루는 8개의 단백질들 사이에 존재하는 상호작용 관계는 7가지 존재하고, 각 관계에 대해 단백질이 갖는 위치에 다른 확률 값을 모두 합하여 그 경로에 해당하는 가중치 값을 구하도록 한다. 예를 들어, SHO1 - BOI2 - BOI1 - BEM1 - STE5 - KSS1 - DIG1 - DIG2 - STE12 의 경로를 살펴보면 각 단백질들의 위치를 살펴보면 SHO1은 bud neck, cytoplasm, cell periphery, BOI2은 bud neck, cytoplasm, cell periphery, bud, BOI1은 bud neck, cytoplasm, cell periphery, bud, BEM1은 bud neck, cell periphery, bud, STE5은 cytoplasm, nucleus, KSS1은 cytoplasm, nucleus, DIG2은 cytoplasm, nucleus, STE12은 nucleus과 같다.

신호 전달 경로를 고려해보면 단백질들의 전달 과정은 세포 내 같은 위치에서의 전달뿐만 아니라 세포막에서 핵까지 도달하기 위해 위치가 변하는 shuttle 단백질에 의해 이루어지게 된다. 따라서 검색한 가능 경로 패스의 가중치를 계산할 때 위치가 변화하는 정도도 반영하기 위해 하나의 단백질이 존재하는 위치가 여러 개일 경우 각각의 경우에 대해서 확률을 구하고 그 값을 모두 더해주는 방식으로 경로에 대한 가중치를 계산하였다. 예를 들어, SHO1과 BOI2사이의 가중치를 계산해보면 SHO1의 위치 정보인 bud neck, cytoplasm, cell periphery와 BOI2의 위치 정보인 bud neck, cytoplasm, cell periphery, bud 를 이용하여 두 단백질 사이에 생성될 수 있는 모든 상호작용 경우에 대한 가중치의 합을 구하면 62.78이 된다. 또한 단백질의 위치 정보가 존재하지 않을 경우에는 갭이 발생하는 것으로 처리하여 그 다음 단백질과의 상호작용 값으로 대치하게 된다. 따라서 이때에는 이에 대한 패널티 점수를 부여한다. 이외에도 다이렉트한 상호작용 단백질의 위치 사이에 상호작용이 발생할 경우가 0인 경우에 한해서 그 상호작용을 제외시키도록 하였다.

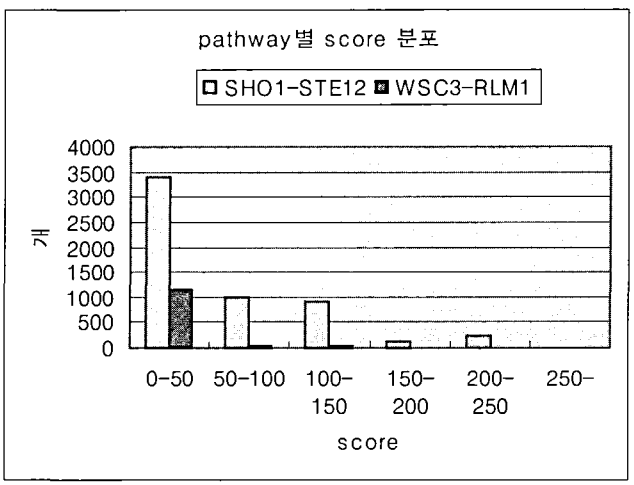
위의 과정을 거쳐서 줄어든 경로들은 일정 순위 경우만을 추려서 시작노드와 도착노드가 같은 경우 병합을 하도록 한다. 따라서 한 패스웨이에 속하는 노드 수가 증가하게 되었고 단순한 선형 패스 형태가 아닌 실제 패스웨이와 유사한 형태를 가지도록 하였다.

Results

위 과정의 결과 찾아낸 신호전달 가능 경로 수는 **[표2]**와 같다. cell wall integrity와 관련된 단백질인 MID2 - RLM1 사이에서 839개인 것으로 나타났고, pheromone response 관련 단백질인 STE2 - STE12 사이에서는 736개, Filamentous Growth Invasion관련 단백질인 SHO1 - STE12 사이에서는 5667개의 가능 경로가 검색됐다. 이렇게 얻어진 모든 가능 경로에 대해 다시 경로 중 상호작용의 가능성이 0이 포함된 가능 경로를 삭제하여 얻어진 결과는 MID - RLM1, STE2 - STE12, SHO1 - STE12 에서 각각 467개, 261개, 3609개로 줄어들게 된다.

START PROTEIN	END PROTEIN	ALL POSSIBLE PATH	PATH AFTER ZERO-DELETION
RAS2	STE12	1477	935
STE2	STE12	736	261
MID2	SWI6	895	361
MID2	SWI4	654	422
SHO1	STE12	5667	3609
WSC3	RLM1	1141	739
MID2	RLM1	839	467

[표2] signaling pathway result



[표3] signaling pathway의 SCORE분포

단백질 상호작용 발생 확률값이 0인 것을 제외하고 처리한 가능 경로 데이터들을 score 값으로 정렬해보면 SHO1 - STE12, WSC3 - RLM1 pathway의 경우, [표3]과 같은 분포를 나타낸다.

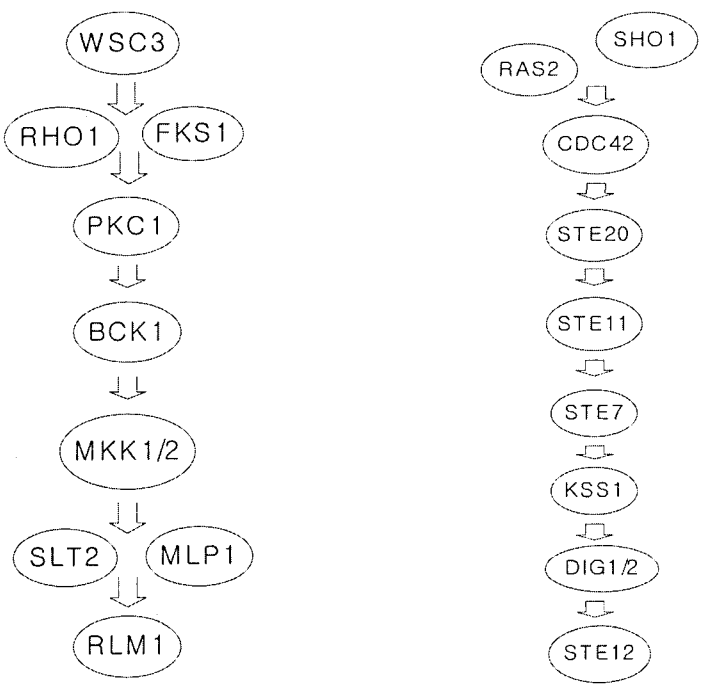
0-50 사이의 score를 갖는 pathway가 가장 많고, 100이상의 score를 갖는 pathway수는 현저히 줄어든다. 높은 score를 갖는 pathway들의 경우에도 단백질 전달 경로 순서를 고려해 볼 때, cell periphery 에서 nucleus로 직접 연결되는 상호작용은 실질적으로 상호작용의 가능성이 희박하는 전제하에 제외시켰다. 왜냐하면, 세포막으로부터 핵까지의 신호 전달 과정은 인접한 위치로의 단백질 상호작용을 통해 이루어질 가능성이 높기 때문이다. 선별된 pathway들 중 출발점과 도착점이 같은 경로 8의 선형 pathway들을 병합하여 한 pathway에 속하는 노드 수가 증가하게 되었고 단순한 선형 패스 형태가 아닌 실재 pathway와 유사한 형태를 가지도록 하였다. 이러한 과정을 통해 만들어진 signal transduction pathway를 [fig4]와 [fig5]를 통해 볼 수 있다. [fig4]는 KEGG에 있는 이미 알려진 신호 전달 경로 중 hypotonic shock 관련 기능을 하는 WSC3-RLM1 pathway와 Starvation관련 기능을 하는 SHO1-STE12 pathway의 신호 전달 경로 그래프이다. [fig5]는 본 연구에서 수행한 단백질 상호작용 데이터 가운데 단백질 위치에 따른 상호작용 빈도수를 가중치화 하여 구한 WSC3-RLM1 pathway와 SHO1-STE12 pathway의 신호 전달 경로 그래프이다. [fig4]와 [fig5]를 비교해보면 WSC3-RLM1 pathway의 경우 기존의 KEGG pathway database와 겹치는 노드는 WSC3, MKK1/2, FKS1, PKC1, SLT2, RLM1 이며, 새로 편입된 노드로는 PEX13/14/17, PLB3, JSN1, VMA6, SLG1, KEL1등이 있다. SHO1-STE12 pathway의 경우 기존의 KEGG pathway database와 겹치는 노드는 SHO1, STE20, STE11, CDC42, KSS1, DIG1/2, STE12 이며, 새로 편입된 노드로는 BOI1/2, BEM1, DSE1, CDC28, BIK1, STE5, SRP1, RGA1, ADY3, FUS3 등이 있다.

한편, 신호 전달 가능 경로의 결과가 단백질 상호작용 데이터 개수에 따라 바뀌는 지 알아보기 위해 기존의 데이터 수보다 약

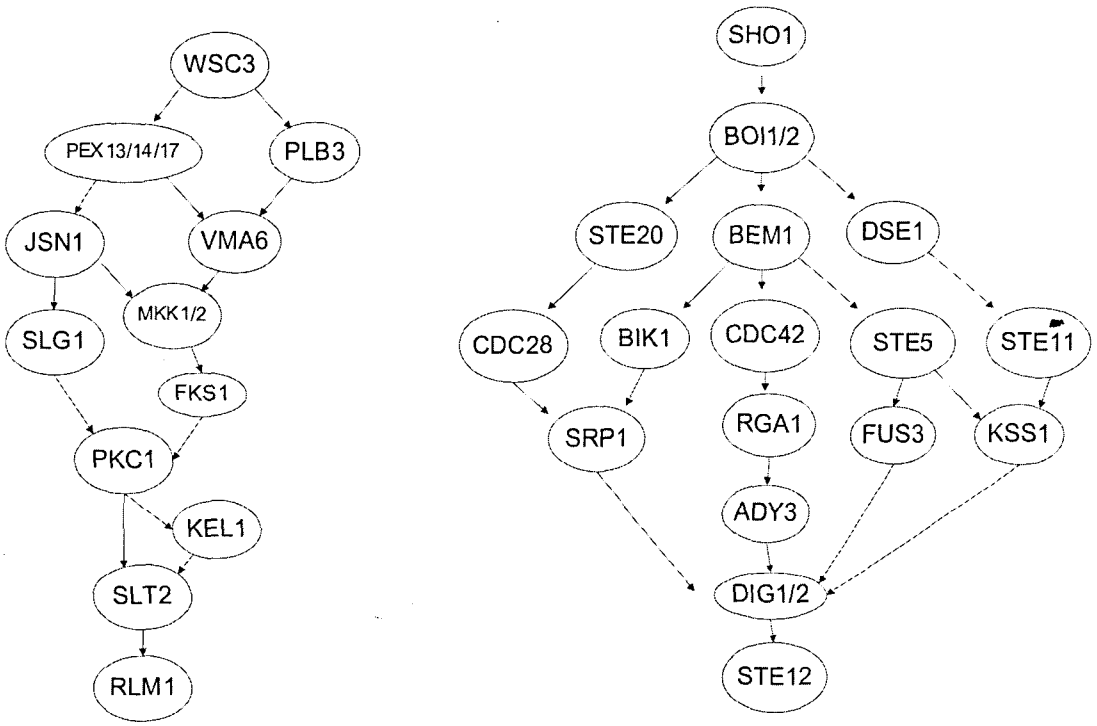
5000개를 늘린 15400개 상호작용 데이터로 STE2-STE12의 신호전달 가능 경로를 구해보았다. 그 결과로 나온 pathway의 개수는 [표4]와 같다. 가중치 값이 0인 경로를 제외한 후의 pathway의 개수는 이전의 261개보다 약 1200개 늘어난 1492개였다.

10467 개 MIPS PPID		15400 개 MIPS PPID	
ALL POSSIBLE PATH	PATH AFTER ZERO-DELETION	ALL POSSIBLE PATH	PATH AFTER ZERO-DELETION
736	261	2192	1492

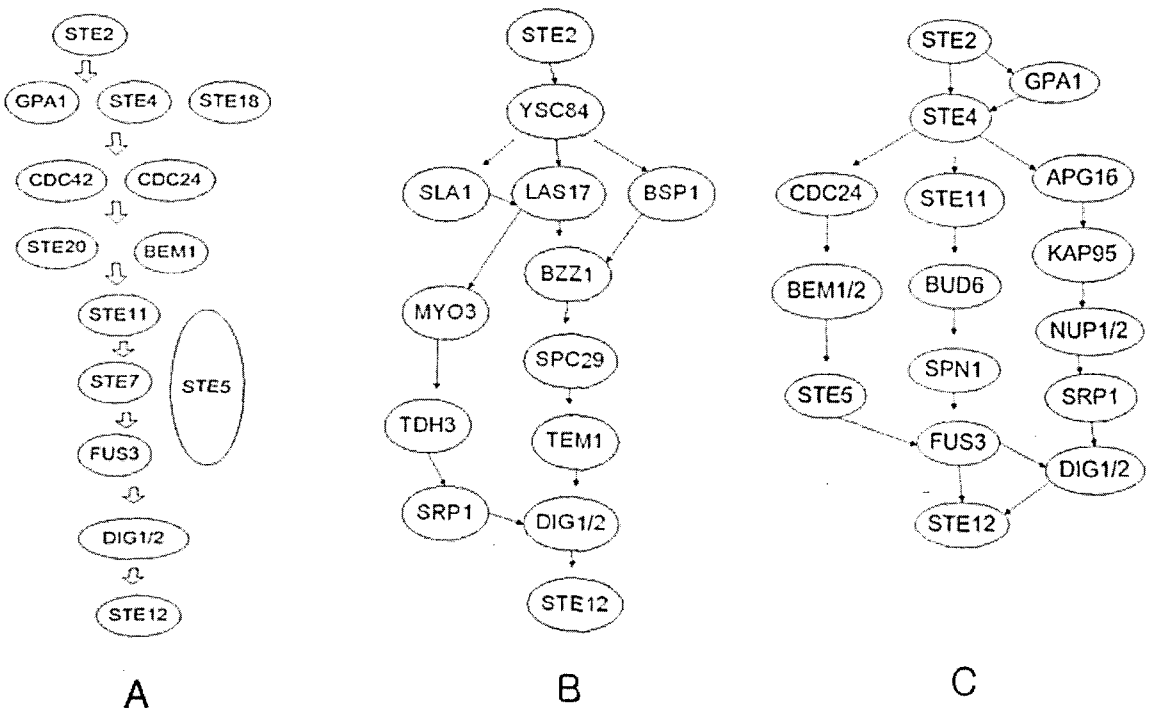
[표4] STE2 – STE12 signaling pathway



[fig.4] KEGG signal transduction pathway



[fig.5] signal transduction pathway by PPID & location info.



[fig.6] STE2-STE12 pathway

(A. KEGG pathway, B. pathway (PPID = 10467) , C. pathway (PPID = 15400)

이 결과를 그래프로 비교해보면 [fig.6]과 같다. A는 KEGG에 있는 이미 밝혀진 STE2-STE12 pathway이고 B는 상호작용 데이터수가 적을 때(10467개), C는 상호작용 데이터 수가 많을 때(15400개)의 pathway이다. 기존의 KEGG데이터와 비교해 봤을 때 상호작용의 개수가 더 많은 C의 경우에 더 많은 단백질이 함께 발견됨을 알 수 있다. 이러한 결과는 본 연구에서 밝힌 신호전달 경로 찾는 과정이 사용된 상호작용 데이터의 개수에 의해 영향을 받는 다는 것을 알려준다. 즉, STE2에서 STE12로 가는 상호 작용 데이터의 missing이 하나라도 있는 경우엔 이러한 경로는 찾아지지 않게 되는 특징이 있기 때문이다. [Fig.6]의 경우에는 10467의 데이터에 STE2-STE4,STE2-GPA1 상호 작용 데이터가 포함되지 않으므로 최종 pathway결과에 나오지 않게 된다. 따라서 상호작용 데이터가 증가하고 이들의 신뢰도가 증가한다는 가정하에 본 연구에서 수행된 상호작용 데이터와 위치 정보를 활용한 pathway 추출 모델링은 기존에 밝혀진 pathway와의 비교를 통해 알려지지 않은 새로운 경로를 발견하고, 이전에 경로에 참여하지 않은 단백질들을 발견할 수 있고, 이미 알려진 단백질들의 새로운 기능들에 대해서도 추론할 수 있을 것이라 기대한다.

Discussion

본 논문에서는 신호 전달 경로에 대한 사전 정보 없이 yeast 단백질 상호작용 정보와 녹색형광단백질(GFP)을 이용하여 밝혀진 4000여 개의 yeast 단백질 분포 데이터를 이용하여 신호 전달 경로를 찾는 방법을 시도하였다. 단백질 상호 작용이 일어나는 세포 내 위치를 가중치로 두어 단백질 상호작용 데이터만으로 찾아낸 세포막 단백질과 핵 단백질간의 경로에 반영하였고, 이를 통해 만들어진 신호 전달 경로 그래프를 구성하는 단백질 중에는 기존에 밝혀진 KEGG pathway database의 단백질이 다수 등장한다는 것을 알았고, 상호 작용 데이터의 수가 많고, 데이터의 신뢰성이 높아질수록 그 결과는 더 높은 정확도를 보여준다는 것을 밝혔다. 이와 같은 신호 전달 경로의 모델링은 기존에 밝혀진 신호전달 경로와의 비교를 통해 알려지지 않은 새로운 경로를 발견할 가능성을 제시하고, 이전에 경로에 참여하지 않은 단백질을 발견할 수 있으며, 이미 알려진 단백질들의 새로운 기능들에 대해서도 추론할 수 있을 것이라 기대한다. 또한 단순히 실험을 통해 얻어진 상호 작용 데이터에 세포막으로부터 핵까지 신호가 전달되는 세포 내 위치를 고려하여 착안한 방법으로서 기존의 실험 방법들과 함께 적용된다면 더 정확한 신호 전달 경로 발견에 응용될 수 있을 것이다.

현재의 시스템은 최고 점수를 갖는 경로가 최적의 경로로 선별되는 것은 아니나 알려진 경로가 높은 점수를 갖는 패스에 속해있는 상태이다. 따라서 이를 보다 최적화시키는 방안에 대한 연구를 지속할 예정이다.

향후 연구로는 단백질의 세포 내 위치를 통해 pathway뿐 아니라 protein complex의 경우, 같은 위치에 존재할 가능성이 높다는 가정하에 동일 위치상에 존재하며 상호 작용을 하는 단백질들을 조사하여 실제 protein complex의 구성과의 유사성을 비교해 볼 계획이다.

Acknowledgements

1. 이 논문은 정보통신부 IMT2000 농생물 생체 정보경로지도 자동구축용 툴 개발 과제(AB-05)의 연구지원 하에 이루어졌음.
2. 단백질 위치에 따른 상호작용 확률표를 제공해주신 허원기 박사님께 감사드립니다.

References

[1] Schikowski B, Uetz P and Fields S A network of protein-protein interaction in yeast. Nat Biotechnol 2000, 18:1257-1261

[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] Victor Spirin and Leonid A.Mirny Protein complexes and functional module s in molecular networks. PNAS 2003, vol,100 no. 21 12123-12128

[4] Won-ki Huh, James V.Falvo, Luke C.GERKE, Adam S.Carroll, Russell W.Howson, Jonathan S.Weissman, Eric K.O' Shea Global anlaysis of protein localization in budding yeast. Nature 2003

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

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