



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2011-0035716
(43) 공개일자 2011년04월06일

(51) Int. Cl.

G01N 33/68 (2006.01) G06F 19/00 (2011.01)
G06F 17/30 (2006.01) C12N 15/12 (2006.01)

(21) 출원번호 10-2009-0093533

(22) 출원일자 2009년09월30일

심사청구일자 2009년09월30일

(71) 출원인

이화여자대학교 산학협력단

서울 서대문구 대현동 11-1 이화여자대학교내

(72) 발명자

조미경

서울특별시 성동구 금호동4가 1517 16통 8반

박현석

서울특별시 용산구 청과동3가 이화맨션 201호

(74) 대리인

특허법인우인

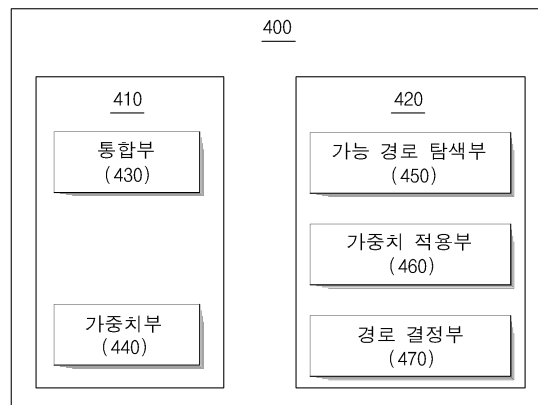
전체 청구항 수 : 총 19 항

(54) 유전자 온톨로지를 이용한 단백질 신호 전달의 경로 탐색 방법 및 그 시스템, 그리고 단백질 신호 전달 경로의 평가 방법

(57) 요약

유전자 온톨로지를 이용한 단백질 신호 전달의 경로 탐색 방법 및 그 시스템, 그리고 단백질 신호 전달 경로 평가 방법이 개시된다. 본 발명에 따른, 단백질 신호 전달의 경로 탐색 방법은, 시작 단백질, 종료 단백질, 경로길이 입력되면, DFS(Depth First Search) 방식으로 단백질의 신호 전달 가능경로를 탐색하는 단계; 단백질 위치에 대한 가중치 및 단백질 상호작용에 대한 가중치 중 적어도 하나를 이용하여 평가 함수를 산출하고, 상기 산출된 평가함수를 이용하여 상기 신호 전달 가능 경로를 평가하는 단계; 상기 평가 함수의 결과를 내림차순으로 정렬하여 가장 큰 값을 결정하는 단계; 및 상기 가장 큰 값을 가지는 신호 전달 가능 경로들에 대하여 유전자 온톨로지 정보를 이용하여 평가를 수행하는 단계를 포함하는 것을 특징으로 한다. 또한, 본 발명에 따른 단백질 신호 전달 경로 평가 방법은, 세포 구성체(Cellular Component, CC) 정보, 분자적 기능(Molecular Function, MF) 정보, 생물학적 작용(Biological Process, BP) 정보를 포함하는 유전자 온톨로지 정보를 이용하여 복수 개의 단백질 신호 전달 경로들에 대하여 평가 값을 계산하는 것을 특징으로 한다.

대표도 - 도4



특허청구의 범위

청구항 1

- (a) 시작 단백질, 종료 단백질, 경로길이가 입력되면, DFS(Depth First Search) 방식으로 단백질의 신호 전달 가능경로를 탐색하는 단계;
- (b) 단백질 위치에 대한 가중치 및 단백질 상호작용에 대한 가중치 중 적어도 하나를 이용하여 평가 함수를 산출하고, 상기 산출된 평가함수를 이용하여 상기 신호 전달 가능 경로를 평가하는 단계;
- (c) 상기 평가 함수의 결과를 내림차순으로 정렬하여 가장 큰 값을 결정하는 단계; 및
- (d) 상기 가장 큰 값을 가지는 신호 전달 가능 경로들에 대하여 유전자 온톨로지 정보를 이용하여 평가를 수행하는 단계를 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 방법.

청구항 2

제 1항에 있어서,

상기 유전자 온톨로지 정보는, 세포 구성체(Cellular Component, CC) 정보, 분자적 기능(Molecular Function, MF) 정보, 생물학적 작용(Biological Process, BP) 정보 중 적어도 하나를 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 방법.

청구항 3

제 1항에 있어서,

상기 유전자 온톨로지 정보는, GO ID(id) 이름(NAME) 속한 온톨로지 도메인(NAMESPACE), 용어의 폐기 상태 여부(IS_ABSOLUTE), is_a 관계(IS_A), part-of 관계(RELATION), subset 정보(SUBSET), 바꿈 (REPLACED), 대체용어(ALT_ID), 관련된 단백질(ANNOTATIONID)을 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 방법.

청구항 4

제 2항에 있어서,

상기 (d) 단계는,

두 GO CC 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 두 단백질 사이의 평가 값을 계산하는 과정을 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 방법.

청구항 5

제 2항에 있어서,

상기 (d) 단계는,

두 GO MF 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 두 단백질 사이의 평가 값을 계산하는 과정을 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 방법.

청구항 6

제 2항에 있어서,

상기 (d) 단계는,

두 GO BP 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 상기 신호 전달 가능 경로를 구성하는 단백질에 대한 평가 값을 계산하는 과정을 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 방법.

청구항 7

제 1항에 있어서,

서로 다른 형태의 단백질 코드를 특정 단백질 코드로 단일화하는 단계; 및

단백질의 위치 분포도와 단백질의 위치간 상호작용 분포도를 이용하여 상기 단백질 위치에 대한 가중치와 상기 단백질 상호작용에 대한 가중치를 산출하는 단계를 더 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 방법.

청구항 8

시작 단백질, 종료 단백질, 경로길이가 입력되면, DFS(Depth First Search) 방식으로 단백질의 신호 전달 가능 경로를 탐색하는 탐색부;

단백질 위치에 대한 가중치 및 단백질 상호작용에 대한 가중치 중 적어도 하나를 이용하여 평가 함수를 산출하고, 상기 산출된 평가함수를 이용하여 상기 신호 전달 가능 경로를 평가하는 가중치 적용부; 및

상기 평가 함수의 결과를 내림차순으로 정렬하여 가장 큰 값을 결정하는 경로 결정부;

상기 유전자 온톨로지 정보를 저장하는 유전자 온톨로지 데이터베이스; 및

상기 가장 큰 값을 가지는 신호 전달 가능 경로들에 대하여 상기 유전자 온톨로지 정보를 이용하여 평가를 수행하는 유전자 온톨로지 적용부를 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 시스템.

청구항 9

제 8항에 있어서,

상기 유전자 온톨로지 정보는, 세포 구성체(Cellular Component, CC) 정보, 분자적 기능(Molecular Function, MF) 정보, 생물학적 작용(Biological Process, BP) 정보 중 적어도 하나를 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 시스템.

청구항 10

제 8항에 있어서,

상기 유전자 온톨로지 정보는, GO ID(id) 이름(NAME) 속한 온톨로지 도메인(NAMESPACE), 용어의 폐기 상태 여부(IS_ABSOLUTE), is_a 관계(IS_A), part-of 관계(RELATION), subset 정보(SUBSET), 바꿈 (REPLACED), 대체용어(ALT_ID), 관련된 단백질(ANNOTATIONID)을 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 시스템.

청구항 11

제 9항에 있어서,

상기 유전자 온톨로지 적용부는, 두 GO CC 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 두 단백질 사이의 평가 값을 계산하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 시스템.

청구항 12

제 9항에 있어서,

상기 유전자 온톨로지 적용부는, 두 GO MF 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 두 단백질 사이의 평가 값을 계산하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 시스템.

청구항 13

제 9항에 있어서,

상기 유전자 온톨로지 적용부는, 두 GO BP 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 상기 신호 전달 가능 경로를 구성하는 단백질에 대한 평가 값을 계산하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 시스템.

청구항 14

제 1항에 있어서,

서로 다른 형태의 단백질 코드를 특정 단백질 코드로 단일화하는 통합부; 및

단백질의 위치 분포도와 단백질의 위치간 상호작용 분포도를 이용하여 상기 단백질 위치에 대한 가중치와 상기 단백질 상호작용에 대한 가중치를 산출하는 가중치부를 더 포함하는 것을 특징으로 하는 단백질 신호 전달의 경로 탐색 시스템.

청구항 15

(a) 복수 개의 단백질 신호 전달 경로 데이터를 입력받는 단계; 및

(b) 상기 복수 개의 단백질 신호 전달 경로들에 대하여, 세포 구성체(Cellular Component, CC) 정보, 분자적 기능(Molecular Function, MF) 정보, 생물학적 작용(Biological Process, BP) 정보 중 적어도 하나를 포함하는 유전자 온톨로지 정보를 이용하여 평가 값을 계산하는 단계를 포함하는 것을 특징으로 하는 단백질 신호 전달 경로의 평가 방법.

청구항 16

제15항에 있어서,

제 8항에 있어서,

상기 유전자 온톨로지 정보는, GO ID(id) 이름(NAME) 속한 온톨로지 도메인(NAMESPACE), 용어의 폐기 상태 여부(IS_ABSOLUTE), is_a 관계(IS_A), part-of 관계(RELATION), subset 정보(SUBSET), 바꿈 (REPLACED), 대체용어 (ALT_ID), 관련된 단백질(ANNOTATIONID)을 포함하는 것을 특징으로 하는 단백질 신호 전달 경로의 평가 방법.

청구항 17

제15항에 있어서,

상기 (b) 단계는, 상기 복수 개의 단백질 신호 전달 경로들 각각에 대하여, GO 비순환 방향 그래프(Direct Acyclic Graph, DAG) 상에서 GO CC 용어 간의 거리를 이용하여 두 GO CC 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 단백질 상호작용하는 두 단백질 사이의 평가 값을 계산하는 것을 특징으로 하는 단백질 신호 전달 경로의 평가 방법.

청구항 18

제15항에 있어서,

상기 (b) 단계는, 상기 복수 개의 단백질 신호 전달 경로들 각각에 대하여, GO 비순환 방향 그래프(Direct Acyclic Graph, DAG) 상에서 GO MF 용어 간의 거리를 이용하여 두 GO MF 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 단백질 상호작용하는 두 단백질 사이의 평가 값을 계산하는 것을 특징으로 하는 단백질 신호 전달 경로의 평가 방법.

청구항 19

제15항에 있어서,

상기 (b) 단계는, 상기 복수 개의 단백질 신호 전달 경로들 각각에 대하여, GO 비순환 방향 그래프(Direct Acyclic Graph, DAG) 상에서 GO BP 용어 간의 거리를 이용하여 두 GO MF 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 단백질 신호 전달 경로의 경로 구성 후보 단백질에 대한 평가 값을 계산하는 것을 특징으로 하는 단백질 신호 전달 경로의 평가 방법.

명세서

발명의 상세한 설명

기술 분야

본 발명은 단백질 신호 전달 경로에 관한 것으로, 보다 상세하게는 유전자 온톨로지 정보를 이용하여 단백질 신호 전달의 경로를 탐색하는 방법 및 그 시스템, 그리고 유전자 온톨로지 정보를 이용한 단백질 신호 전달 경로

[0001]

평가 방법에 관한 것이다.

배경 기술

- [0002] 1980년부터 시작된 DNA 염기서열결정방법 (Sequencing Method)의 개발로 유전체 정보에 대한 활발한 연구가 진행되고 있다. 그 결과 유전체 정보는 1977년 바이러스(Bacteriophage)(5.4 Kb)의 전체 염기서열을 해독된 이후, 1983년에는 람다(Lambda)(48.5Kb), 1997에는 인플루엔자(Haemophilus Influenzae)(1.8Mb)와 효모(Yeast)(12Mb), 1998에는 선충류인 C.elegans (98Mb)의 유전체 정보가 해독 되었다. 더욱이 2001년 인간 유전체 사업을 통해 인간의 유전자 지도(~3,000Mb)가 완성된 이후 많은 유전체 정보에 대한 연구가 진행되고 있다.
- [0003] 현재 효모, C.elegans, 초파리(Drosophila), 쥐(Mouse), 집쥐(Rat) 등 100종이상의 유전체 정보가 축적되고 있으며, 이 같은 정보는 데이터베이스에 계속 축적되고 있다.
- [0004] 최근에는 유전체 정보뿐만 아니라 단백질 상호작용에 대한 정보 또한 빠른 속도로 축적되고 있으며 이를 통해 대사(Metabolism) 및 신호전달(Signal Transduction), 그리고 세포주기(Cell Cycle)에 대한 정보를 수집할 수 있게되었다. KEGG를 비롯하여 EcoCyc, SPAD 등 다양한 종류의 세포 내 경로정보에 대한 데이터베이스가 그것이다. 특히 SPIDer는 대중적인 프로그램으로 효모의 유전체를 기반으로 단백질 상호작용을 데이터베이스로 정리하였다. SPIDer는 다양한 조건검색이 제공하며, 그래프를 이용하여 단백질 상호작용을 시각화 해주어 연구자들은 편리하게 광역적인 단백질 상호작용을 분석 할 수 있게 되었다.
- [0005] 최근 유전체 정보뿐만 아니라 단백질 상호작용에 대한 관심이 높아진 이유는 단백질 상호작용 경로를 구현하는 것이 생물학적 현상을 이해하는 것에 매우 중요하기 때문이다. 생명체를 구성하는 최소 단위인 세포의 생명활동은 세포 내의 다양한 종류의 단백질이 적재적소적소에서 정확하게 제 기능을 발휘함으로써 가능한 것이다.
- [0006] 특히 세포 안에서는 한 종류의 단백질이 단독적으로 작용하기보다는 기능적으로 연관되어 있는 단백질 상호작용이 정확하게 일어남으로써 대사작용 및 신호전달이 체계적으로 이루어진다.
- [0007] 그러나 현재 생물학적 측면에서는 외부자극에 대한 세포 내부의 복잡한 신호전달 경로가 정확히 어떻게 존재하는지 알아내기가 힘들다. 또한, 평가 대상이 되는 여러 신호 전달 경로들에 대하여 효과적인 평가 방법이 존재하지 않았다.

발명의 내용

해결 하고자하는 과제

- [0008] 본 발명은 유전자 온톨로지 정보를 이용하여 단백질 신호 전달 경로를 추출하는 시스템 및 방법, 그리고 유전자 온톨로지 정보를 이용하여 단백질 신호 전달 경로를 평가하는 방법을 제공하는 것이 목적이다.

과제 해결수단

- [0009] 상기 목적을 달성하기 위한 본 발명에 따른, 단백질 신호 전달 경로의 평가 방법은, (a) 복수 개의 단백질 신호 전달 경로 데이터를 입력받는 단계; 및 (b) 상기 복수 개의 단백질 신호 전달 경로들에 대하여, 세포 구성체(Cellular Component, CC) 정보, 분자적 기능(Molecular Function, MF) 정보, 생물학적 작용(Biological Process, BP) 정보 중 적어도 하나를 포함하는 유전자 온톨로지 정보를 이용하여 평가 값을 계산하는 단계를 포함하는 것을 특징으로 한다.
- [0010] 여기서, 상기 유전자 온톨로지 정보는, GO ID(id) 이름(NAME) 속한 온톨로지 도메인(NAMESPACE), 용어의 폐기 상태 여부(IS_ABSOLUTE), is_a 관계(IS_A), part-of 관계(RELATION), subset 정보(SUBSET), 바꿈 (REPLACED), 대체용어(ALT_ID), 관련된 단백질(ANNOTATIONID)을 포함할 수 있다.
- [0011] 또한, 상기 (b) 단계에서, 상기 복수 개의 단백질 신호 전달 경로들 각각에 대하여, GO 비순환 방향 그래프(Direct Acyclic Graph, DAG) 상에서 GO CC 용어 간의 거리를 이용하여 두 GO CC 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 단백질 상호작용하는 두 단백질 사이의 평가 값을 계산할 수 있다.
- [0012] 또한, 상기 (b) 단계에서, 상기 복수 개의 단백질 신호 전달 경로들 각각에 대하여, GO 비순환 방향 그래프(Direct Acyclic Graph, DAG) 상에서 GO MF 용어 간의 거리를 이용하여 두 GO MF 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 단백질 상호작용하는 두 단백질 사이의 평가 값을 계산할 수 있다.
- [0013] 또한, 상기 (b) 단계에서, 상기 복수 개의 단백질 신호 전달 경로들 각각에 대하여, GO 비순환 방향 그래프

(Direct Acyclic Graph, DAG) 상에서 GO BP 용어 간의 거리를 이용하여 두 GO MF 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 단백질 신호 전달 경로의 경로 구성 후보 단백질에 대한 평가 값을 계산할 수 있다.

- [0014] 상기 목적을 달성하기 위한 본 발명에 따른, 단백질 신호 전달의 경로 탐색 방법은, (a) 시작 단백질, 종료 단백질, 경로길이가 입력되면, DFS(Depth First Search) 방식으로 단백질의 신호 전달 가능경로를 탐색하는 단계; (b) 단백질 위치에 대한 가중치 및 단백질 상호작용에 대한 가중치 중 적어도 하나를 이용하여 평가 함수를 산출하고, 상기 산출된 평가함수를 이용하여 상기 신호 전달 가능 경로를 평가하는 단계; (c) 상기 평가 함수의 결과를 내림차순으로 정렬하여 가장 큰 값을 결정하는 단계; 및 (d) 상기 가장 큰 값을 가지는 신호 전달 가능 경로들에 대하여 유전자 온톨로지 정보를 이용하여 평가를 수행하는 단계를 포함하는 것을 특징으로 한다.
- [0015] 여기서, 상기 유전자 온톨로지 정보는, 세포 구성체(Cellular Component, CC) 정보, 분자적 기능(Molecular Function, MF) 정보, 생물학적 작용(Biological Process, BP) 정보 중 적어도 하나를 포함할 수 있다.
- [0016] 또한, 상기 유전자 온톨로지 정보는, GO ID(id) 이름(NAME) 속한 온톨로지 도메인(NAMESPACE), 용어의 폐기 상태 여부(IS_ABSOLUTE), is_a 관계(IS_A), part-of 관계(RELATION), subset 정보(SUBSET), 바꿈 (REPLACED), 대체용어(ALT_ID), 관련된 단백질(ANNOTATIONID)을 포함할 수 있다.
- [0017] 또한, 상기 (d) 단계는, 두 GO CC 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 두 단백질 사이의 평가 값을 계산하는 과정을 포함할 수 있다.
- [0018] 또한, 상기 (d) 단계는, 두 GO MF 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 두 단백질 사이의 평가 값을 계산하는 과정을 포함할 수 있다.
- [0019] 또한, 상기 (d) 단계는, 두 GO BP 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 상기 신호 전달 가능 경로를 구성하는 단백질에 대한 평가 값을 계산하는 과정을 포함할 수 있다.
- [0020] 또한, 상기 단백질 신호 전달의 경로 탐색 방법은, 서로 다른 형태의 단백질 코드를 특정 단백질 코드로 단일화하는 단계; 및 단백질의 위치 분포도와 단백질의 위치간 상호작용 분포도를 이용하여 상기 단백질 위치에 대한 가중치와 상기 단백질 상호작용에 대한 가중치를 산출하는 단계를 더 포함할 수 있다.
- [0021] 상기 목적을 달성하기 위한 본 발명에 따른, 단백질 신호 전달의 경로 탐색 시스템은, 시작 단백질, 종료 단백질, 경로길이가 입력되면, DFS(Depth First Search) 방식으로 단백질의 신호 전달 가능경로를 탐색하는 탐색부; 단백질 위치에 대한 가중치 및 단백질 상호작용에 대한 가중치 중 적어도 하나를 이용하여 평가 함수를 산출하고, 상기 산출된 평가함수를 이용하여 상기 신호 전달 가능 경로를 평가하는 가중치 적용부; 및 상기 평가 함수의 결과를 내림차순으로 정렬하여 가장 큰 값을 결정하는 경로 결정부; 상기 유전자 온톨로지 정보를 저장하는 유전자 온톨로지 데이터베이스; 및 상기 가장 큰 값을 가지는 신호 전달 가능 경로들에 대하여 상기 유전자 온톨로지 정보를 이용하여 평가를 수행하는 유전자 온톨로지 적용부를 포함하는 것을 특징으로 한다.
- [0022] 여기서, 상기 유전자 온톨로지 적용부는, 두 GO CC 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 두 단백질 사이의 평가 값을 계산할 수 있다.
- [0023] 또한, 상기 유전자 온톨로지 적용부는, 두 GO MF 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 두 단백질 사이의 평가 값을 계산할 수 있다.
- [0024] 또한, 상기 유전자 온톨로지 적용부는, 두 GO BP 용어 간의 평가 값을 계산하고, 상기 계산된 평가 값을 이용하여 상기 신호 전달 가능 경로를 구성하는 단백질에 대한 평가 값을 계산할 수 있다.
- [0025] 또한, 상기 단백질 신호 전달의 경로 탐색 시스템은, 서로 다른 형태의 단백질 코드를 특정 단백질 코드로 단일화하는 통합부; 및 단백질의 위치 분포도와 단백질의 위치간 상호작용 분포도를 이용하여 상기 단백질 위치에 대한 가중치와 상기 단백질 상호작용에 대한 가중치를 산출하는 가중치부를 더 포함할 수 있다.

효 과

- [0026] 본 발명에 의하면, 유전자 온톨로지 정보를 이용하여 단백질 상호작용 정보와 단백질 위치정보를 기반으로 용이하게 신호전달 경로를 추출할 수 있고, 유전자 온톨로지 정보를 이용하여 단백질 신호 전달 경로를 평가할 수 있다.
- [0027] 나아가 신호전달 경로가 추출됨으로써 질병과 유전자간의 연관성 파악이 용이해질 수 있다. 또한, 신호전달 과

정을 분석함으로써 신약 물질의 대상 단백질을 결정하는데 유용하게 이용될 수 있을 뿐만 아니라, 신약에 대한 정확한 대상 단백질을 선별하는 것은 신약개발과정시 불필요한 예산낭비를 줄여줄 수 있다.

[0028] 게다가, 유전자 온톨로지 정보를 이용하여 신호 전달 가능 경로들에 대하여 평가를 수행함으로써 신호 전달 경로 추출의 정확도를 높일 수 있다.

발명의 실시를 위한 구체적인 내용

[0029] 이하에서는 도면을 참조하여 본 발명을 보다 상세하게 설명한다.

[0030] 세포를 구성하는 여러 물질(Molecules) 중 단백질은 매우 중요하다. 단백질은 생명체를 구성하는 성분으로, 효소 및 항체를 형성하여 생명현상을 유지시키기 때문이다. 단백질들은 각자의 기능을 독립적으로 하는 것이 아니라, 각각의 기능을 분담하고 협력하여 하나의 생물학적 기능을 수행한다.

[0031] 예를 들어, 호르몬이 세포에 영향을 주기 위해서는, 먼저 세포 표면에 존재하는 수용체에 호르몬이 결합하여야 한다. 이 결합은 Receptor Dimerization 같은 수용체의 구조적인 변화를 수반하기도 하며, 일련의 하위 단백질들의 Signal Transduction을 유발한다. 이는 핵 내의 다양한 유전자를 단백질로 발현시켜 세포에 특정한 영향을 주게 된다.

[0032] 모든 세포는 주위 환경의 자극을 감지하고, 그에 반응한다. 단순한 생물체인 박테리아도 포도당이나 아미노산이 있는 방향으로 이동하며, 많은 단세포 생물들은 다른 생명체가 분비하는 물질에 반응한다. 다세포 생물의 경우, 한 세포는 주위의 세포들로부터 받게 되는 다양한 신호들을 해석하고, 그에 따라 자신의 반응을 조절하고 다시 신호를 보내야만 한다.

[0033] 이러한 세포 간 신호전달이 이루어지면서, 세포의 대사, 이동, 증식, 생존 그리고 분화와 같은 세포의 활동이 조절된다. 세포 내 신호전달 과정은 외부 신호전달 물질을 세포 표면의 수용체 단백질이 인식하고 이를 세포 내에 전달함으로써 이루어진다.

[0034] 신호전달 경로는 세포 내 신호전달 과정을 말한다. 호르몬이나 환경변화 같은 외부의 자극이 올 경우, 세포는 이를 감지하여 세포 내로 전달하여 이에 대한 반응을 한다. 이 때 세포의 외부에서 오는 신호를 세포 내로 전달하는 과정을 신호전달 경로라고 하며, 대표적인 세포 내 신호전달 경로는 MAP Kinase Cascade (MAPKinase 신호전달 경로)등이 있다.

[0035] 신호전달 경로의 일반적인 특징은 첫째, 세포 막에 존재하는 수용체에 신호전달 물질이 결합을 하면서 복잡한 다단계 반응이 시작된다. 신호전달 물질로는 호르몬이 가장 대표적이고, 그 밖에 신경전달 물질이나 성장인자 심지어 포도당과 지방산도 신호전달 물질 역할을 할 수 있다.

[0036] 둘째, 신호전달 물질이 수용체에 결합한 후 이 신호는 여러 가지 방식으로 다단계 증폭된다. 일반적으로 G-단백질-cAMP, Ras-MAPK, JAK-STAT 등의 경로가 있다. 이런 일련의 다단계 반응은 주로 Kinase에 의한 인산화로 조절되며, 대부분의 단백질은 인산화가 되면 활성을 갖게 되고 인산기가 떨어져 나가면 활성을 잃게 된다.

[0037] 셋째, 많은 경우 부신호전달자(Second Messenger)의 도움으로 신호전달이 일어난다. 칼슘이온이 대표적인 경우이고, 지방관련 물질들도 있다.

[0038] 넷째, 이런 신호전달 경로는 생체 반응을 조절 한다. 포도당의 저장 및 합성 같은 대사반응의 방향을 조절하거나, 유전자의 발현의 억제 또는 발현을 조절하여 여러 생체 반응에 영향을 준다.

[0039] 한편 세포는 신호전달경로 이외에 대사경로(Metabolic Pathway)도 함께 가지고 있다. 신호전달경로는 외부의 자극을 받아서 세포내부의 수용체가 이를 받아서 이에 대한 반응을 보이는 일련의 단백질 상호작용이 일어나는 현상을 말한다. 그러나 대사경로란 세포 내에 필요한 화학물질(대사산물)을 생성하는 과정을 말한다. 세포는 외부의 변화에 반응하고 대처하기 위해, 신호를 감지하는 신호전달 경로와 더불어, 이에 필요한 산물을 얻기 위한 대사경로를 동시에 가지고 있다.

[0040] 다음은 단백질 상호작용의 활용에 대해 설명한다.

[0041] 단백질 상호작용 경로를 구현하는 것은 생물학적 현상을 이해하는 데에 매우중요하다. 세포의 생명활동은 세포 내의 다양한 종류의 단백질이 특정 상황에서 정확하게 기능을 발휘함으로써 가능한 것이다. 특히 세포 안에서는 한 종류의 단백질이 단독적으로 작용하기 보다는 기능적으로 연관되어 있는 단백질 상호작용이 일어남으로써 대사작용 및 신호전달이 체계적으로 이루어진다.

- [0042] 단백질 상호작용의 경로를 밝혀내는 것은 특정 생물학적 작용에 대한 기능적필연성에 대한 이해를 향상시킬 수 있으며, 아직 그 특성이 규명되지 않은 단백질의 기능을 예상해볼 수 있다. 더불어 역으로 특정 단백질 상호작용 단계에서 그 존재는 밝혀졌으나 생물학적 작용이 밝혀지지 않은 단백질의 기능을 파악하는 데에 중요하다.
- [0043] 단백질 상호작용은 단백질 신호전달 경로를 알아내는 데 중요한 정보로 활용되고 있다. 외부 자극이 세포에 가해지면 세포 내의 단백질들은 복잡한 단백질 상호작용을 통해 그 신호를 전달하여 유전자 발현과 같은 현상으로 연결한다.
- [0044] 세포 내 다양한 생화학적 작용을 일으키는 단백질은 세포 내에서 단독적으로 작용할 수도 있지만, 일반적으로는 다른 단백질 상호작용이 이루어지는 경우가 대다수이다. 따라서 특정 단백질과 이와 기능적으로 연관된 다른 단백질 상호작용을 이해하는 것은 특정 생화학적작용에 있어서 필수적인 부분이라 할 수 있다.
- [0045] 이와 관련하여 세포 내 단백질(Cellular Protein)의 Subcellular Localization 정보도 고려하여야 한다. 세포 안의 물질들은 Cytoskeleton(microtubule, actin filament, intermediate filament 등) 구조의 도움을 통해 세포 내에서 제 기능을 수행하는 데에 가장 적합한 곳에 위치하게 된다. 세포 내 다양한 세포 소기관 및 여러 단백질 등이 단순히 혼합되어 있는 상태로 보일지라도 실제로는 특정 생화학적 반응의 효율을 가장 높일 수 있는 최적 상태로 맞춰져 있는 것이다.
- [0046] 단백질 상호작용도 단백질간의 물리적인 거리가 대단히 짧은 상태에서 일어나는 것이므로 이러한 반응을 이해하려면 그 각각의 세포 내 단백질 위치정보를 이해하는 것이 바람직하다.
- [0047] 세포 내의 단백질들은 특정 위치에 존재하고 있다. 이는 단백질 상호작용이 어느 곳에서나 일어나는 것이 아니고, 특정한 위치에서만 일어날 수 있다는 것이 분자생물학 실험들을 통해 증명되고 있기 때문이다.
- [0048] 단백질 상호작용은 세포막으로 둘러 쌓여 있는 세포 소구역 안에 동시에 존재하여 가까운 거리에 존재하는 단백질 간에 일어난다.
- [0049] 도 1은 세포내 단백질 위치 정보를 나타낸 도면이다. 도 1과 같은 단백질 위치 정보에 대한 자료는 신호전달 경로를 연구하기 위한 단백질 위치정보에 활용 될 수 있다.
- [0050] 이하에서는 세포 내 단백질 위치정보를 이용하여 MAP Kinase 신호전달 경로를 추출하고 예측하는 방법에 대해 설명한다.
- [0051] 단백질 상호작용은 단백질 간 물리적으로 상당히 근접한 상태에서 일어난다. 예를 들어 막으로 구성된 위치에 존재하는 단백질은 그 위치안에서 우선적으로 단백질 상호작용이 일어난다. 따라서 본 실시예에서는 세포 내 단백질 위치가 알려진 Budding Yeast 의 단백질 위치정보를 이용한다.
- [0052] 도 2는 단백질 위치정보의 예를 도시한 도면이다. 도 2에 도시된 바와 같이, YAL001C 단백질은 핵(Nucleus)와 세포질(Cytoplasm)에 존재함을 알 수 있다.
- [0053] 한편, 단백질은 한곳에 머무르는 것이 아니라 이동하며 상호작용을 하기도 하기 때문에 하나의 단백질은 여러 위치정보를 가질 수 있다.
- [0054] 본 발명에서는 단백질 5,495 개를 가지고 다중 위치정보를 고려하여 다중 위치정보 각각을 21 개의 세포 내 단백질 위치에 분포 시킨 후 얻어진 7,786 개의 단백질 위치정보 데이터를 이용한다.
- [0055] 도 3은 세포 내 단백질 다중위치 분포도를 도시한 도면이다. 도 3에 도시된 바와 같이, 단백질은 하나 이상의 위치 정보를 갖고 있음을 확인할 수 있다.
- [0056] 도 4는 본 발명의 일 실시예에 따른 단백질의 위치 정보를 기반으로 단백질의 신호 경로를 추출하는 시스템에 대한 개략적인 블록도이고, 도 5는 본 발명의 일 실시예에 따른 단백질의 위치 정보를 기반으로 단백질의 신호 경로를 추출하는 방법에 대한 흐름도이다.
- [0057] 본 단백질의 신호 경로 추출 시스템을 LocSPF(Local Signaling Pathway Finder)라 명명할 수도 있고, 시스템을 크게 전처리부(410)와 신호전달 경로 탐색부(420)로 구분할 수 있다.
- [0058] 전처리부(410)는 단백질 코드를 단일화시키는 통합부(430)와 단백질 위치의 분포도 및 단백질 위치간 상호작용 분포도를 이용하여 단백질 위치에 대한 가중치와 단백질 상호작용에 대한 가중치를 구성요소로 하는 확률값 테이블을 작성하는 가중치부(440)로 구성된다.

- [0059] 전처리부(410)가 수행된 후 신호전달 경로 탐색부(420)는 단백질 상호작용의 신호전달 모든 가능 경로를 탐색하는 가능 경로 탐색부(450), 가중치 적용부(460) 및 경로 결정부(470)를 포함한다.
- [0060] 구체적으로, 가능 경로 탐색부(450)는 시작 단백질, 목표 단백질, 경로길이를 입력되면, DFS(Depth First Search) 방식으로 모든 가능경로를 탐색한다. 그리고, 가중치 적용부(460)는 모든 가능경로에 대해 단백질 위치에 대한 가중치와 단백질 상호작용에 대한 가중치를 이용하여 각 경로마다 평균값을 산출한다. 마지막으로, 경로 결정부(470)는 가중치 적용부(460)에서 얻은 각 경로마다의 평균값을 내림차순으로 정렬한 후 가장 큰 값을 신호전달 경로로 결정한다.
- [0061] 이하에서는 시스템의 각 구성요소를 보다 상세히 설명한다.
- [0062] 통합부(430)는 서로 다른 형태로 제공되는 단백질 코드를 특정 단백질 코드로 단일화한다(S510).
- [0063] 도 6은 본 발명의 일 실시예에 따른 데이터베이스 통합 방식을 도식적으로 도시한 도면이다.
- [0064] 본 실시예에서는 단백질의 신호 전달에 대한 설명의 편의를 도모하기 위해 단백질 중 효모(*Saccharomyces cerevisiae*) 단백질을 대상으로 신호 전달을 파악해 본다. 그리고, 상기한 효모 단백질 위치간 상호작용 정보와 효모 단백질 위치정보는 DIP 데이터 베이스(<http://dip.doe-mbi.ucla.edu/dip>)와 MIPS 데이터베이스(<http://www.helmholtz-muenchen.de/en/mips/>)에서 제공받을 수 있다.
- [0065] 이때 각각의 데이터베이스에서 제공하는 단백질 코드는 서로 다른 형태로 제공된다. 즉, DIP에서 사용되는 코드는 P 로 시작되는 코드이며 MIPS에서 사용되는 코드는 Y 로 시작되는 코드 형태이다. 코드 단일화를 위해 NCBI 데이터 베이스(<http://www.ncbi.nlm.nih.gov>)에서 제공되는 정보를 이용하여 MIPS 코드 형식으로 변환한다.
- [0066] 이와 같이 획득된 변환된 단백질 위치간 상호작용 정보는 약 46,121개이다. 도 7은 본 발명의 일 실시예에 따른 단백질 상호작용을 인접행렬로 표시한 도면이다. 도 7에 도시된 바와 같이, 인접행렬에서 두 단백질 사이에 단백질 상호작용이 있으면 '1' 로 표현하고 단백질 상호작용이 없으면 빈칸으로 낸다. 그리하여 행 A6 셀에 데이터 YAL008W 와 열 E1 셀에 데이터 YAL007C 와는 단백질 상호작용이 있음을 알 수 있다.
- [0067] 가중치부(440)는 단백질의 위치 분포도와 단백질의 위치간 상호작용 분포도를 이용하여 단백질 위치에 대한 가중치와 단백질 상호작용에 대한 가중치를 산출한다(S520).
- [0068] 예를 들어, 단백질 a 와 단백질 b 가 있다고 가정할 때, 단백질의 위치 정보와 단백질 상호작용 정보를 이용하여 단백질 위치에 대한 가중치와 단백질 상호작용에 대한 가중치를 구한다.
- [0069] 가중치부(440)는 단백질 a의 위치에 대한 가중치를 단백질 위치 정보를 이용해서 $\text{Log}(\frac{\text{단백질 a 위치에 속하는 단백질의 누적 경우의 수}}{\text{모든 위치의 단백질 누적 경우의 수의 총합}} * 100) * 100$ 로 구한다.
- [0070] 구체적으로, 가중치부(440)는 하기 수학적 식 1과 같은 단백질 a의 위치 정보의 확률값을 산출한다.

수학적 식 1

$$\left(\frac{|\{i | i \in \text{proteins}, a \in \text{단백질 } i \text{의 위치}\}|}{\sum_{j \in \text{proteins}} |\text{단백질 } j \text{의 위치}|} \right) * 100$$

[0071]

[0072] 여기서 '*' 기호는 곱하기 기호이다.

[0073] 그리고, 상기한 확률값이 작기 때문에 크기 보정을 위해 다시 100을 곱하고 로그를 취함으로써, 가중치부(440)는 하기 수학적 식 2와 같은 단백질 a의 위치에 대한 가중치를 산출한다.

수학적 식 2

$$\text{EvL}_a = \log \left(\left(\frac{|\{i | i \in \text{proteins}, a \in \text{단백질 } i \text{의 위치}\}|}{\sum_{j \in \text{proteins}} |\text{단백질 } j \text{의 위치}|} \right) * 100 \right) * 100$$

[0074]

[0075] 또한, 가중치부(440)는 단백질 a-단백질 b 의 단백질의 상호작용에 대한 가중치를 단백질 상호작용의 정보와 각 단백질의 위치정보를 이용해서 $\text{Log}(\frac{\text{단백질 a 위치들과 단백질 b 위치들간 모든 상호작용 경우의 수}}{\text{모든 위치}})$

간 상호작용 경우의 수 총합* 100) * 100)으로 구한다.

[0076] 구체적으로, 가중치부(440)는 하기 수학적 식 3과 같은 단백질 a-단백질 b 의 단백질의 상호작용에 대한 가중치를 산출한다.

수학적 식 3

$$\begin{aligned} & EvL_{(a,b)} \\ &= \log \left(\frac{\left| \left\{ (i,j) \mid (i,j) \in \text{proteins}, a \in \text{단백질 } i \text{의 위치}, b \in \text{단백질 } j \text{의 위치} \right\} \right|}{\sum_{(i,j) \in \text{proteins}} \left| \left\{ \text{단백질 } i \text{의 위치} \times \text{단백질 } j \text{의 위치} \right\} \right|} \right) \\ & \quad * 100 \Big\} * 100 \end{aligned}$$

[0077]

[0078] 가능 경로 탐색부(450)는 시작 단백질, 종료 단백질 및 경로 길이가 입력되면, 시작 단백질, 종료 단백질 및 경로 길이를 기준으로 Depth First Search 방식을 통해 신호전달 가능 경로를 산출한다(540).

[0079] 단백질 신호전달 경로는 세포 내 세포막에 위치하는 단백질로부터 출발하여 핵에 존재하는 단백질까지 도달하는 과정이다. 본 실시예에서는 효모 단백질 MAPK 의 "High Osmolarity" 기능을 대표 모델로 하여 신호전달 경로를 예측한다.

[0080] 예를 들어, 시작 단백질(예:YIL147C)과 종료단백질(예:YMR037C) 그리고 경로길이(예:8)를 입력할 수 있다.

[0081] 본 발명에서는 통합부(410)에서 이미 추출해 놓은 단백질 상호작용 정보 46,121 개를 사용하여, 입력 받은 시작 단백질로부터 종료 단백질까지 경로길이를 제한하여 모든 신호전달 가능 경로를 찾을 수 있다. 이때, DFS(Depth First Search) 방식을 기반으로 시작 단백질로부터의 경로길이를 비교하여 경로길이 범위안에서 종료 단백질에 도달할 때까지 신호전달 경로를 탐색한다.

[0082] 도 8은 DFS 방식을 이용하여 탐색하는 예를 도식적으로 도시한 도면이고, 도 9는 본 발명의 일 실시예에 따른 단백질 상호작용의 모든 경로를 찾는 알고리즘에 관한 도면이며, 도 10은 도 9의 알고리즘의 결과를 도시한 도면이다.

[0083] 도면에 도시된 바와 같이, 시작 단백질(예:YIL147C)과 종료 단백질(예:YMR037C)이 입력되면, 도 9의 알고리즘에 의해 도 10의 결과가 도출된다.

[0084] 가중치 적용부(460)는 단백질 위치에 대한 가중치 및 단백질 상호작용에 대한 가중치 중 적어도 하나를 이용하여 평가 함수를 산출하고, 산출된 평가 함수로 신호 전달 가능 경로를 평가한다(S550).

[0085] 막으로 구성된 위치에 존재하는 단백질은 그 위치 안에 존재하는 단백질들끼리 우선적으로 단백질 상호작용이 일어날 수 있다. 따라서 단백질 위치정보를 신호전달 경로 찾기의 가중치로 반영하는 것이 바람직하다.

[0086] 가중치 적용부(460)에서 사용되는 평가함수는 하기 수학적 식 4로부터 산출된다.

수학적 식 4

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

[0087]

[0088] 그리고, 도 11은 본 발명의 일 실시예에 따른 수학적 식 4의 평가 함수를 적용한 알고리즘을 도시한 도면이다.

[0089] 구체적으로, 단백질 상호작용 신호전달 경로 찾기를 평가하기 위해서는 패스를 이루는 단백질 앞뒤 순서 간 모든 평가 값을 곱한다.

[0090] 예를 들어 YLR332W-YOL109W-YBR069C-YLR372W-YMR058W-YDL029W-YHR030C-YPL089C를 평가해 본다. YLR332W-YOL109W, YOL109W-YBR069C, YBR069C-YLR372W, YLR372W-YMR058W, YMR058W-YDL029W, YDL029W-YHR030C, YHR030C-YPL089C 단백질 간 모든 단백질 상호작용 평가 값을 구하고 그 값들을 곱한다.

[0091] 단백질 상호작용의 값을 평가하기 위해 단백질 위치정보와 단백질 상호작용 위치에 대한 상호작용 확률 값을 포

합하는 이차원 배열 정보를 사용한다. 이때 단백질 위치에 대한 가중치는 각 단백질 위치의 확률 값 중 가장 큰 값을 갖고 있는 단백질의 위치 값이다.

[0092] 단백질 상호작용의 확률 값은 배열(21위치X21위치)을 이용하여 두 단백질간 상호작용에서 갖는 모든 단백질 위치 상호작용 경우를 반영한 확률 값을 사용하며 단백질 위치의 상호작용에 대한 가중치는 모든 위치에 대한 상호작용의 확률값 중 가장 큰 값을 가지는 위치 값이다.

[0093] 예를 들어 단백질 위치정보가 YMR088C: vacuolar membrane, 세포질 YLL040C: mitochondrion, endosome, 세포질 이면, YMR088C와 YLL040C의 모든 단백질 위치의 상호작용 경우의 수는 위치 개수 2*3한 값인 6개가 발생한다. 도 12는 본 발명의 일 실시예에 따른 단백질 위치의 상호작용에 대한 모든 경우의 수를 도시한 도면이다.

[0094] 이하에서는 단백질 위치정보와 단백질 위치의 상호작용 값을 이용하여 2가지 형태의 평가 함수를 수학식으로 표현한다.

[0095] 첫째, 일 형태의 평가함수는 단백질I-단백질J 위치에 대한 상호작용 가중치 중 최고값으로 하기 수학식 5와 같다.

수학식 5

$$\max_{(a,b) \in \text{Protein i의 위치} \times \text{Protein j의 위치}} (\text{Local a와 Local b의 Interaction 확률})$$

[0096]

[0097] 둘째, 다른 형태의 평가함수는 단백질 I 와 단백질 J 의 단백질 상호작용 위치에 대한 상호작용 가중치 중 최고값으로 하기 수학식 6과 같다.

수학식 6

$$\max_{(a,b) \in \text{Protein i의 위치} \times \text{Protein j의 위치}} (\text{Local a와 Local b의 Interaction 확률} \times \text{Local a의 위치 가중치} \times \text{Local b의 위치 가중치})$$

[0098]

[0099] 즉, 단백질 I 와 단백질 J 의 단백질 상호작용 위치에 대한 상호작용 가중치를 평가하기 위해서 단백질 I 와 단백질 J 의 단백질 위치정보를 이용한다.

[0100] 도 13은 본 발명의 일 실시예에 따른 단백질 상호작용 위치에 대한 상호작용의 확률 값을 도시한 도면이다. 도면에 도시된 바와 같이, 단백질 I 의 위치가 없거나 단백질 J 의 위치가 없을 경우 1.0 의 값으로 평가한다. 그리고, 단백질 I 와 단백질 J 모두 하나 이상의 위치를 가질 경우 그 위치 간 쌍의 모든 경우의 수 중에서 평가함수를 통해 산출한 값 중 가장 큰 값을 선택하고, 선택된 단백질 상호작용 값들을 곱함으로써 한 개의 단백질 상호작용 신호전달 경로 찾기 단계의 최종값으로 한다.

[0101] 도 14는 A 위치 가중치 * B 위치 가중치 * A-B 위치 상호작용 가중치를 도시한 표이다.

[0102] 경로 결정부(470)는 가중치 적용부(460)에서 산출된 모든 단백질 상호작용 경로에 대해 단백질 위치에 대한 가중치와 단백질 상호작용 위치에 대한 가중치의 결과 중 어느 하나를 동순위를 허용하면서 내림차순으로 정렬하여 가장 큰 값을 신호전달 경로로 결정한다(S560).

[0103] 도 15는 본 발명의 일 실시예에 따른 경로길이 9를 제외한 전체 신호전달 경로에 대한 수행 결과표이다. 또한, 도 16은 본 발명에 따른 신호 전달 경로를 탐색하는데 소요된 시간을 도시한 표이다.

[0104] 이하에서는 단백질 정보와 단백질 상호작용 관계를 표현한 인접행렬을 바탕으로 시작 단백질에서 종료 단백질까지 그 경로길이가 6 인 신호전달 경로를 탐색한다.

[0105] 도 17은 본 발명의 일 실시예에 따른 경로길이가 6일 때의 결과표이다. 도 17에서 알 수 있듯이 KEGG 신호전달 경로 3 개의 경로에 대해 본 발명의 제1 방법(LocSPF_방법1)과 본 발명의 제2 방법(LocSPF_방법2)은 KEGG 신호전달 경로와 일치한다.

[0106] 결과를 분석해 본 결과 경로 1에 대한 본 발명의 제1 방법은 가중치 최종 값이 상위 3%에서 KEGG에서 제공하는 신호전달 경로와 일치하는 결과를 얻으며 나머지 모두에서는 상위 1%에서 KEGG에서 제공하는 신호전달 경로와

일치하는 결과를 얻어 본 발명이 유효함을 알 수 있다.

- [0107] 상기한 본 발명의 제1 방법과 본 발명의 제2 방법은 서로 다른 평가 함수를 적용한 것으로, 본 발명의 제1 방법은 단백질 A ??단백질 B 의 단백질 상호작용 위치에 대한 상호작용 가중치를 평가함수로 적용하였으며, 본 발명의 제2 방법은 단백질 A 위치의 가중치, 단백질 B 위치의 가중치, 단백질 A ??단백질 B 상호작용 위치에 대한 상호작용 가중치의 곱을 평가 함수로 적용한 것이다.
- [0108] "High Osmolarity" 기능에 대해 경로1의 본 발명의 제1 방법의 결과를 가지고 가중치 결과값을 기준으로 최고값 순으로 나열하였을 때, 가중치 점수의 상위 3%이내의 수준에서 KEGG에서 제공하는 신호전달 경로와 일치함은 도 17를 통해 알 수 있다. 또한 경로1의 제2 방법, 경로2와 경로3의 제1 방법과 제2 방법을 적용한 결과의 가중치를 큰 값순으로 나열하여 KEGG에서 제공하는 신호전달 경로를 비교해 본 결과 Top스코어에서 찾을 수 있다. 결국 목표 경로 찾기에서 찾고자 하는 경로를 제1 방법과 제2 방법에 의해서 가중치가 최고 값을 갖는 그룹에서 목표 경로를 찾고 있음을 보여 준다.
- [0109] 본 발명의 제1 방법 및 제2 방법에 의해 최고의 가중치 값을 갖는 신호전달 경로가 KEGG의 신호 전달 경로와 일치한다는 것은 본 발명의 제1 방법 및 제2 방법이 유효함을 증명한다.
- [0110] 하지만, 동일한 가중치값을 갖는 상당수의 신호전달경로가 존재 할 시에는 좀더 체계적인 방법이 요구되게 된다. 도 18은 LocSPF 시스템을 이용하여 구한 Top 스코어의 신호전달 경로의 수이다. 도 18에서 알 수 있는 것처럼 경로길이가 6 인 경우는 중복된 데이터의 수가 많지 않지만 경로의 길이가 7 이상되면 중복데이터의 수가 급격히 늘어난다.
- [0111] 따라서 조금 더 복잡한 신호전달 경로의 경우 더 세분화시킬 수 있는 방법이 필요하다. 즉, 다수의 Top스코어를 갖는 신호전달 경로 가중치 값에 대해 좀 더 세밀한 알고리즘 적용이 필요하다. 따라서 이를 위해 이하에서는 유전자 온톨로지(Gene Ontology: GO)를 이용하는 실시예에 관하여 설명하기로 한다.
- [0112] 도 19는 본 발명의 일 실시예에 따른 유전자 온톨로지를 이용하여 단백질의 신호 경로를 추출하는 시스템의 블록도이고, 도 20은 본 발명의 일 실시예에 따른 유전자 온톨로지를 이용하여 단백질의 신호 경로를 추출하는 방법의 흐름도이다. 본 실시예에 따른 단백질 신호 경로 추출 시스템 및 방법은 도시된 바와 같이, 도 4에 관하여 설명된 실시예에 GO 정보 데이터베이스(480)와 GO 적용부(490)가 추가되고, 도 5에 관하여 설명된 실시예에 S570단계가 추가된다.
- [0113] 유전자 온톨로지(Gene Ontology, GO)는 유전자 부산물인 단백질이나 RNA들에 대한 세포 구성체(Cellular Component, CC), 분자적 기능(Molecular Function, MF), 생물학적 작용(Biological Process, BP)의 3가지 독립된 영역에 대해 체계적이고 구조화된 분류 체계를 제공한다. 이러한 용어들의 정의는 기술적(Descriptive)인 정의가 아니라 용어들의 관계를 그래프화한 것으로써 이해하기 쉽고 체계화 된 형태로 제공되며, 이러한 그래프 구조는 컴퓨터를 이용하여 용어들의 관계를 유추하고 처리 하는 것을 용이하게 해준다.
- [0114] 이러한 GO 용어들은 GO 주석(GO Annotation, GOA)을 통하여 유전자 부산물(Gene Product)들과 연결 되므로, 원하는 정확한 분야에 관련이 있는 단백질들을 검색하거나, 또는 반대로 어떤 단백질이 어떤 위치에서 발견되며 어떠한 분자적 기능을 하고 어떠한 생물학적 프로세스와 연관되어 있는지 알 수 있게 해준다. 여기서 유전자 부산물은 단백질이나 RNA를 의미 한다. 이러한 GO 정보는 여러 데이터베이스와 종에 걸쳐서 제공되며, GO에는 많은 논문과 지식 그리고 실험 정보가 결집되어 있기 때문에 단순한 검색 이상의 용도로 활용도 검토될 수 있을 것이다.
- [0115] 이러한 GO 정보 활용의 가능성을 적용하고 테스트해 보고자 본 발명에서는, 상술한 바와 같이 효모 세포 내 21개 위치 정보를 활용한 MAP Kinase 신호전달 경로 추출 및 예측에서 LocSPF 알고리즘 적용 결과 우수한 정확도를 나타내었지만, 최종 결과 값에서 다수의 동점 값이 존재함에 따라 좀 더 세분된 알고리즘 적용 및 개선을 위하여 GO를 활용하는 방안을 고안하고 실험하였다.
- [0116] 도 19 및 도 20을 참조하면, GO 정보 데이터베이스(445)는 유전자 온톨로지 정보를 저장하고 있으며, GO 적용부(490)는 S560단계에서 얻어진 신호전달 경로 데이터 중 가장 상위의 동점 결과 값들을 대상으로 GO 정보를 이용하여 평가를 수행한다(S570).
- [0117] GO 정보 데이터베이스(442)는 구체적으로, 세포 구성체(CC) 정보, 분자적 기능(MF) 정보, 생물학적 작용(BP) 정보를 저장하고 있으며, GO 적용부(490)는 GO 정보 데이터베이스(442)에 저장된 세포 구성체(CC) 정보, 분자적 기능(MF) 정보, 생물학적 작용(BP) 정보 중 적어도 하나를 이용하여 신호전달 경로 데이터 중 상위의 동점 결과

값들을 대상으로 평가를 수행한다.

- [0118] 유전자 온톨로지(Gene Ontology, GO)는 세포 구성체(Cellular Component, CC), 분자적 기능(Molecular Function, MF), 생물학적 작용(Biological Process, BP)의 세 개의 영역으로 각각 세포 구성체, 분자적 기능, 생물학적 프로세스에 관한 용어들의 정의 관계 정보를 제공하여 준다. 이러한 정보는 GO 주석(GO Annotation, GOA)과 연계하여 단백질의 세포 구성체내 위치와 분자적 기능 그리고 연관된 생물학 작용을 알 수 있게 해준다. 비순환 방향 그래프(Direct Acyclic Graph, DAG)의 계층적 구조와 is-a, part-of 등의 관계로 체계화 되어 있어 컴퓨터를 활용한 처리를 용이하게 해준다. 그래서 이러한 정보를 어떻게 신호전달 경로 찾기 및 예측에 활용 할지에 대하여 CC, MF, BP의 세 가지 영역에 대해 설명하기로 한다.
- [0119] 우선, 세포 구성체(Cellular Component) 정보의 활용에 관하여 설명하면 다음과 같다.
- [0120] 유전자 온톨로지(Gene Ontology, GO)의 세포 구성체(Cellular Component, CC)는 세포의 구성체들을 세부적인 세포 구조나 분자구조체 범위에서 나타내게 된다. 이러한 세포 구성체 정보는 계층적인 구조로 세분화 된다. 본 발명의 실시예에서는 GO의 효모종에 대한 축약판인 Yeast GO Slim을 사용하였으며 계층화된 단백질 위치정보와 GO 상에서 그래프로 연결된 관계들을 사용하여 이전 실시예에서 사용한 효모 21개 단백질 위치정보와는 다른 방식으로 접근 하였다.
- [0121] 먼저 GO CC의 계층구조의 활용에 대해 설명한다.
- [0122] 도 21은 Yeast GO Slim의 CC 영역 전체(상)와 일부(중), 그리고 위치관계의 포함 관계 일부(하)를 나타내는 도면이다. 도 21(중)에서 볼 수 있듯이 세포질인 cytoplasm은 그를 구성하는 요소들인 vacuole, mitochondrion 등으로 나누어 지고 다시 mitochondrion은 그 세부 구성 요소인 mitochondrial envelope를 갖는다. 때문에 이러한 위치 정보는 단백질이 단순히 같은 위치에 있는지만을 알 수 있는 것이 아니라 한 단백질은 비순환 방향 그래프(Direct Acyclic Graph, DAG)상에서 부모 노드의 위치에 있고 다른 단백질은 자식 노드의 위치에 있다는 것의 정보로 고려 할 수 있게 해준다. 예를 들어 가상의 단백질A는 세포질인 cytoplasm에 존재하고 단백질B와 C는 vacuole에 단백질D는 mitochondrion에 존재 한다고 가정한다면 [그림4-1](하)와 같은 위치 관계를 예상할 수 있을 것이다.
- [0123] 그리고 도 21(하)에서 볼 수 있듯이 단백질 B와 C는 좀 더 세분화된 영역에 존재함으로 단백질 A와 B 보다는 최악의 경우 더 작은 영역에서 같이 있게 될 가능성이 보장된다. 이것은 단백질 A와 B가 꼭 단백질 B와 C 보다 멀리 있다는 것을 의미 하지는 않는다. 예를 들어 단백질 A는 세포질 범위에서 어느 곳에나 존재 할 수 있으므로 세포질 범위 안인 Vacuole의 단백질 B 바로 옆일 수도 있다.
- [0124] 하지만 단백질 C는 세포질 일 뿐 만 아니라 적어도 Vacuole 영역 안에 한정 되므로 제일 나쁜 경우에도 단백질 B와 Vacuole 영역의 지름 보다 더 멀리 있을 수는 없다.
- [0125] 단백질 D와 B와의 관계를 생각해 보면 D는 mitochondrion 영역에 한정되므로 A보다는 B와 만날 가능성이 적어 질 것이라고 생각 할 수 있다
- [0126] 여기에 추가로 microtubule organizing center에 속하는 단백질 E와 Cytoskeleton에 속하는 단백질 F를 가정하면 이러한 관계는 좀 더 명확해진다. B와 F는 공유 영역 microtubule organizing center가 있지만 직접 연결되는 영역이 없다고 생각되므로 이들의 직접적인 단백질 상호작용이 일어날 확률은 적다고 생각하는 것이 옳을 것이다. 이를 도 20(중)의 GO 그래프에서 보면 이들 용어의 거리(hop수)와 연관되어 있음을 알 수 있다. 예를 들어 거리가 0인 (같은 노드 안) B와 C는 거리 가 1인 A와 B 보다 최대거리가 더 짧고 거리가 2인 B와 D 또는 A와 F는 A와 B보다 최대 거리가 길다고 볼 수 있다. 이러한 논리를 종합해 보면 단백질 사이의 거리의 최대값을 제한하는 것은 즉 단백질 사이의 거리가 가장 멀 경우를 가정하여 보면 일반적으로 단백질들이 속한 GO용어들의 GO상의 거리에 비례함을 알 수 있다.
- [0127] 하지만 같은 거리에 존재하고 있다고 해도 세포질에 존재하는 두 단백질 사이의 거리보다는 이보다 하위 분류에 있는 mitochondrion에 존재하는 두 단백질 사이의 최대 거리가 더 짧을 것이라는 것을 유추 할 수 있다. 때문에 결론적으로 단백질들이 속해 있는 용어들 간의 GO상의 거리와 각 용어의 GO상에서 깊이(루트 노드로부터 거리)가 해당 단백질이 어느 범위 안에서 제한되는지를 유추하는데 이용 될 수 있음을 생각 할 수 있다.
- [0128] 여기에서 깊이를 정의하는데 문제점이 발생하게 되는데 GO는 DAG이므로 루트 노드로부터 자식 용어까지 도달하는 경로가 1개 이상 일 수 있으며 경로의 길이도 여러 가지 일 수 있다. 도 22는 is-a and part-of 관계의 예(좌)와, 다중 깊이 값의 예(우)를 나타내는 도면이다. 도 22(우)는 이러한 예를 보여주며 여기서 all을 루트 노

드로 봤을 때 membrane은 최단 깊이 3, 최장 깊이 4이다. 또 관계도 is-a 관계와 (좌측그림은 위쪽이 루트, 우측은 아래쪽이 루트임에 유의) part-of 관계로 된 경로가 있을 수 있으므로 각각의 경우에 대하여 특성을 고려하여 주어야 한다. 여기서는 단백질들이 얼마나 근접 할지에 대한 정보가 중요하므로 어떤 위치가 개념적으로 하위 개념인지(is-a) 보다는 어떤 요소의 일부인지(part-of)가 더 관련성이 있으므로 is-a 관계보다는 part-of 관계가 더 중요하다.

[0129] 일반적으로 part-of 관계가 한번 등장하면 다음에 is-a 관계가 등장하여도 지식 노드는 조상 노드와 part-of 관계가 성립 된다. 이러한 경우 중간에 오는 is-a 관계는 부분관계와는 상관없게 되며 이는 영역의 세분화에는 직접적인 영향이 없다고 볼 수 있을 것이다. 예를 들어 도 22(좌)에서 part-of로 부모 노드인 핵(nucleus)과 연결된 nucleolus는 핵의 더 세부적인 구성 요소인데 반면 membrane과 is-a관계로 연결된 plasma membrane이나 endomembrane system은 membrane의 한 종류이지만 membrane을 구성하는 더 세부적인 요소라고 볼 수는 없다. 때문에 본 발명의 실시예에서는 이러한 용어들과의 거리와 루트부터 해당 용어까지의 최대 part-of관계의 수를 이용하여 단백질간의 거리 및 위치의 인접성을 평가하여 신호전달 경로 추출에 사용하였다.

[0130] 다음으로, 유전자 온톨로지(Gene Ontology, GO)의 분자적 기능(Molecular Function, MF) 정보의 활용에 관하여 설명한다.

[0131] 온톨로지는 작은 범위의 분자단위의 기능을 다루게 된다. 본 발명의 실시예에서 이러한 MF를 상호작용하는 단백질끼리 공유하는 MF가 많을수록 그 두 단백질이 같은 기능과 관련이 있고 이 경우 이러한 기능을 바탕으로 경로를 구성할 확률이 크다고 가정하였다. 이 경우 세포 구성체(Cellular Component, CC)와 유사하게 정확하게 같은 GO MF 용어를 단백질들이 공유 할 때 같은 기능을 할 뿐만 아니라 유사한 기능을 하는 경우도 고려해야 할 것이다.

[0132] 도 23은 Yeast GO Slim Molecular Function의 일부를 나타내는 도면이다. 예를 들자면 도 23에서 어떤 단백질 A와 B가 모두 동일한 Protein Kinase activity에 관련되어 있다면 이 둘은 같은 기능을 하겠지만 한 단백질은 Protein Kinase activity에 관련되어 있고 다른 하나는 이의 is-a 부모인 transfer activity에 관련되어 있을 경우에도 비슷한 기능을 할 가능성을 고려 할 수 있을 것이다. 때문에 단백질 상호작용하는 두 단백질이 서로 동일한 또는 비슷한 기능을 하는지를 GO MF 용어들 사이의 거리로 평가하여 사용하였다.

[0133] 거리가 같이 3인 두 용어, 예를 들어 도 23의 molecular_function과 protein kinase activity 또는 nucleotidyltransferase activity와 protein kinase activity 같은 경우 깊이 값이 깊은 용어들의 관계가 밀접하다고 가정하여 CC와 비슷하게 루트 노드부터의 깊이 개념을 추가하였다. 이 경우는 is-a 관계 part-of 관계 모두 의미가 있을 수 있다고 고려되어 평가 값을 계산하는데 있어 두 가지 관계 모두를 사용하였다.

[0134] 다음으로, 유전자 온톨로지(Gene Ontology, GO)의 생물학적 작용(Biological Process) 정보의 활용에 관하여 설명한다.

[0135] 유전자 온톨로지(Gene Ontology, GO)의 생물학적 작용(Biological Process, BP)은 분자적 기능(Molecular Function, MF)와 달리 좀 더 큰 범위의 생물학적 기능들을 기술하며 기능 자체 보다는 기능의 목적에 초점을 맞춘 용어들을 정의하는 온톨로지이다. 본 발명의 실시예에서는 신호전달 경로를 추출하기 위해 신호전달 경로의 시작 단백질과 종료 단백질 그리고 신호전달 경로길이를 입력 값으로 사용하였다. 이러한 경우 이 신호전달 경로를 구성하는 단백질들은 하나의 신호전달 작용을 나타내는 생물학적 작용에 연관되어 있다고 생각 할 수 있을 것이다. 때문에 신호전달 경로의 시작 단백질과 종료 단백질이 포함되어 있는 생물학적 작용은 찾고자 하는 목표 작용일 가능성이 높다고 할 수 있을 것이다.

[0136] 그렇기 때문에 BP정보의 응용은 경로 구성 후보 단백질이 시작과 종료 단백질 그리고 BP 용어를 얼마나 공유하는지를 계산하여 적용할 수 있으며 이에 앞에서 설명한 CC, MF와 유사한 용어간의 거리와 깊이 정보를 사용할 수 있다.

[0137] 상기된 도 4 및 도 5에 관해 설명된 실시예에서, 21개 단백질 위치 정보를 이용하여 신호전달 경로를 추출하였다. 이의 추출 값들은 KEGG의 이미 밝혀진 신호전달 경로를 상위값에서 발견하였지만 여전히 동일한 평가 값을 갖는 다수의 경로들이 존재하는 문제점이 있다. 이는 결국 원하지 않는 경로를 구성하는 단백질의 위치 정보가 원하는 경로를 구성하는 단백질의 위치 정보와 비슷하여 모두 비슷한 평가 값을 얻게 되었다고 보여지며, 이러한 동일 평가 값을 갖는 경로 단백질에 대해 좀 더 세분화 하기 위하여 GO 데이터를 사용한다.

[0138] 본 발명의 일 실시예에서, GO를 적용한 신호전달 경로 찾기 세분화를 위한 알고리즘에 사용된 정보들은 다음과

같다.

- [0139] 1. Yeast GO Slim: OBO format-version 1.2, date: 26:03:2008 17:16
- [0140] 2. GO Annotation: Saccharomyces cerevisiae SGD(Filtered), 12/2/2008
- [0141] 3. 도 4 및 도 5 실시예의 결과인 상위 12개 경로 데이터 결과 중 6개: 상위 결과 경로 데이터를 나타낸 도 24에 박스로 표시(가능한 LocSPF_방법1 경로들을 대상으로 했으나 일부는 단백질 상호작용 미존재 문제로 LocSPF_방법2 경로들을 대상으로 적용함)
- [0142] 각각의 경로를 찾기 위해서는 도 4 및 도 5 실시예와 마찬가지로 경로의 시작 단백질과 종료 단백질 그리고 경로길이를 입력 값으로 준다.
- [0143] 계산에 사용된 GO와 GOA 정보 그리고 기초 계산 데이터의 형태는 도 25에 도시된 표와 같다. 도 25를 참조하면, GO 정보는 GO ID(id) 이름(NAME) 속한 온톨로지 도메인(NAMESPACE), 용어의 폐기 상태 여부(IS_ABSOLUTE), is_a 관계(IS_A), part-of 관계(RELATION), subset 정보(SUBSET), 바꿈 (REPLACED), 대체용어(ALT_ID), 관련된 단백질(ANNOTATIONID)을 보여준다. 참고로 이 정보는 GO 데이터뿐만 아니라 GO 용어에서 연관 단백질들을 찾기 위한 GOA 정보도 포함하고 있다.
- [0144] 이러한 GO 정보 중 is-a와 part-of 관계 정보를 활용해 그래프 형태로 나타내면 도 26과 같이 나타낼 수 있다. 도 26에서 파란색 엷지는 part-of 관계 빨간색 엷지는 is-a 관계를 나타내며, 왼쪽 자주색 노드들은 BP 영역, 가운데 흰색 노드들은 CC영역 오른쪽 노란색 노드들은 MF영역 GO 용어들을 각각 나타낸다.
- [0145] 도 27은 GOA 정보를 나타낸다. 도 27을 참조하면, GOA 정보는 단백질과 여기 연관된 GO 용어들을 참고 하기 위하여 사용되었으며 처음 3 필드가 단백질의 ID에 관한 정보이며 마지막 필드가 관련된 GO 용어들을 나타낸다. 이들 사이의 정보 필드들은 본 구현에서는 사용되지 않고 있다.
- [0146] 또한 계산에서는 용어간의 거리와 각각 용어의 is-a 관계 최단 깊이와 part-of 관계 최장 깊이 값을 사용하게 되므로 이 값들은 도 26으로부터 미리 계산될 수 있으며 그 형태는 도 28 내지 30에 도시된 바와 같다.
- [0147] 도 28은 CC의 전체 is-a, part-of 깊이 정보로서, 첫 A열은 GO ID, C열은 GO DAG상에서 최단 is-a 깊이, E열은 GO DAG상에서 최장 part-of 깊이를 나타낸다.
- [0148] 도 29는 BP의 전체 is-a, part-of 깊이 정보로서, A열은 GO ID, C열은 GO DAG상에서 최단 is-a 깊이, E열은 GO DAG상에서 최장 part-of 깊이를 나타낸다.
- [0149] 도 30은 MF의 전체 is-a, part-of 깊이 정보로서, A열은 GO ID, C열은 GO DAG상에서 최단 is-a 깊이, E열은 GO DAG상에서 최장 part-of 깊이를 나타낸다.
- [0150] 도 31 내지 33은 GO DAG상에서 용어들 간의 거리 값들의 일부를 예시로 보여주고 있다.
- [0151] 도 31은 GO CC용어간의 거리값을 보여주는 표이며 숫자는 A열의 용어들과 1행의 용어들간의 거리값이다. 두 용어가 같을 경우 거리값은 0으로 나타난다. 도 32는 BP GO 용어들 간의 거리값을 보여주는 표이고, 도 33은 MF GO 용어들 간의 거리값을 보여주는 표이다.
- [0152] 이하에서는 본 발명의 일 실시예에 따라 GO 적용부(490)에서 유전자 온톨로지를 적용하여 평가를 수행하는 방법을 상세히 설명한다.
- [0153] 먼저, 세포 구성체(Cellular Component) 정보가 적용되는 경우에 관하여 설명한다.
- [0154] GO CC 정보를 이용하여 시작 단백질부터 시작하여 이 단백질과 이웃하는 다음 단백질들의 위치 정보를 다음 수학적 식 7에 따라 계산한다.
- [0155] 이러한 공식의 적용은 계속해서 인접한 단백질 간에 적용된다.

수학식 7

EvCC_(a,b)

$$= \{1 - \log(\text{distance} + 1)\} * \left(\frac{\text{partof1} + \text{partof2} + 1}{2} \right)$$

* CC_{factor}

(단, $0 \leq \text{distance} \leq 3$, $1 \leq \text{partof1}, \text{partof2}$)

[0156]

[0157]

상기 수학식 7에서 EvCC_(a,b)는 GO CC용어 a와 b간의 평가 값을 나타낸다. 비교하는 GO 용어는 2개 각각 a, b라고 할 때 distance는 GO DAG상에서 a와 b 사이의 거리이다. 이 distance가 0인 경우 a와 b는 동일한 GO CC 용어임을 뜻 하고 distance가 클 수록 서로 다른 범위의 세포 내 위치임을 의미하게 된다.

[0158]

distance에 1값을 더해 주는데 그 이유는 distance가 0인 경우 log 0이 되어 계산이 불가능 하기 때문이다. '1-log(distance+1)'을 하는 이유는 distance가 가까울수록 높은 평가값을 산출하기 위함이다. GO CC용어의 DAG상의 깊이가 깊을수록 더 작은 최악의 거리를 가지며 루트 노드에서 부터 대상 용어까지 part_of 관계가 많을수록 더 세분화된 위치임을 나타낸다. 때문에 이를 반영하여 주기 위해 공식에서 GO용어 a와 b의 part_of 관계 깊이의 평균값을 이전 결과에 곱해주게 된다. 공식에서 partof1은 루트부터 a 용어까지 도달 할 때 지날 수 있는 최대 part-of 엣지의 수이며 partof2는 루트부터 b 용어까지의 최대 part-of 엣지 수이다.

[0159]

여기에 part_of 관계의 개수가 모두 0이어서 최종값이 0이 되는 것을 막기 위해 1을 더해 준다(part_of의 범위가 1보다 크게 설정되어 있지만 이 후에 다른 방식으로 적용 시 문제를 방지하기 위해 이 부분이 존재하고 있다). 만약 어떤 단백질에 대하여 GO CC 정보가 없는 경우, distance 값은 distance의 최대 값인 3이 되며 partof값은 최소 값인 1이 된다. distance의 최대값을 3으로 정하는 이유는 distance가 커질수록 위치 관계는 모호해지고 불확실해 지기 때문이다.

[0160]

예를 들어 distance가 0이면 동일한 노드, 1이면 부모와 자식관계, 2이면 조부모와 자식 또는 형제관계가 되지만 3이상에서는 가능한 관계가 크게 늘어나게 된다.

[0161]

마지막으로 CC factor는 GO CC 정보의 중요도로써 여기서는 1을 사용하고 있다. 이러한 factor를 곱해 주는 것은 CC와 MF, BP 정보에 대해 각각 정보의 중요도를 조절함으로써 좀 좋은 결과를 얻기 위함이며, 이들의 중요도는 아직 실험적으로 밝히지 못했으므로 임의적으로 할당하였다. 추가로 곱하기와 카티전 곱(Cartesian product) 모두 '×' 기호로 표기 가능해서 둘 사이의 표기가 모호 할 수 있기 때문에 곱하기는 '*' 기호로 카티전 곱은 '×' 기호로 표기하였다.

[0162]

또 이러한 공식을 적용 할 때 한 단백질에 1개 이상의 연관된 GO CC용어가 존재 할 수 있기 때문에 이러한 가능한 모든 조합들 중 가장 큰 값을 취한다.

수학식 8

$$\text{EvPCC}_{(i,j)} = \max_{(a,b) \in \text{protein } i \text{의 CC} \times \text{protein } j \text{의 CC}} (\text{EvCC}_{(a,b)})$$

[0163]

[0164]

상기 수학식 8은 두 단백질 i와 j 사이의 평가값 공식으로써 서로 단백질 상호작용하는 단백질 i와 j 사이의 평가 값 중 최대값을 취한다. 값이 높을수록 경로 구성 단백질일 가능성이 큰 것으로 평가하며 그 이유로는 i 단백질이 여러 곳의 위치에 있고 j도 여러 곳의 위치에 존재 할 경우 가장 가까운 곳에서 서로 단백질 상호작용을 할 확률이 크고 그 경우의 평가값을 적용하는 것이 타당하다고 보이기 때문이다.

[0165]

다시 상기 수학식 8에 근거한 예를 들어 보자. 가상의 단백질 i에 GO CC 용어 a1, a2가 관련되어 있고(단백질 i가 세포 내의 a1과 a2위치에서 발견됨), 단백질 j에 GO CC 용어 b1, b2, b3가 있다면 [공식4-2]에 의해 (a1,b1), (a1,b2), (a1,b3), (a2,b1), (a2,b2), (a2,b3)의 6쌍 조합이 생기며 이들 사이의 [공식4-1]에 의한 EvCC 평가값들 중 가장 큰 값을 취하게 된다.

[0166]

다음으로, 분자적 기능(Molecular Function) 정보가 적용되는 경우에 관하여 설명한다.

[0167]

GO MF 정보의 경우도 GO CC의 경우와 비슷한 방식을 취하나 MF와 CC정보의 성격 차이를 고려하여 약간의 변화를

준다. 예를 들어 GO CC는 위치간의 구성 관계가 중요하므로 GO DAG 상에서 part-of 엣지만 평가 값에 영향을 미치게 하였으나 GO MF 정보의 경우 비슷한 종류의 작용에 관련이 있을 경우도 서로 관계될 수 있다고 판단하여 is-a 관계도 고려하여 두 GO MF 용어들 사이의 평가 값은 다음과 같은 평가 값 수학적식 9를 적용한다.

수학적식 9

$$EvMF_{(a,b)} = \{1 - \log(\text{distance} + 1)\} * \left\{ \frac{\text{partof1} + \text{partof2} + \text{isa1} + \text{isa2} + 1}{4} \right\} * MF_{\text{factor}}$$

(단, $0 \leq \text{distance} \leq 3$, $1 \leq \text{partof1}$, partof2 , isa1 , isa2)

[0168]

[0169] 상기 수학적식 9에서는 GO CC의 상기 수학적식 7과 같이 해당 단백질에 GO MF 용어가 없는 경우 distance는 3, partof와 isa는 1값을 할당한다. 추가로 MF용어는 CC용어 보다 중요도를 낮게 부가할 수 있다. 그래서 MF factor 값을 0.5로 설정할 수 있다.

[0170] 최종적으로 주어진 단백질 i와 j간의 평가 값을 구하는데 있어서 GO CC와 같은 위치 정보의 경우 해당 단백질들이 여러 군데에서 발견될 경우 둘 사이에 가장 가까운 위치를 취하는 방식 (조합의 max 값을 취함)이 타당하겠지만 MF의 경우 이러한 최대 값을 갖는 GO 용어들의 조합이 실제로 원하는 경로 탐색 과정에서 사용되는 분자적 작용이 아니라 다른 생물학적 작용에 구성요소가 되는 작용일 수도 있다. 때문에 MF에 대하여 두 단백질 i와 j사이의 평가 값은 단백질에 관련된 GO 조합들의 최대값을 취하는 대신에 조합들의 평균값을 취하는 다음 수학적식 10을 사용한다.

수학적식 10

$$EvPMF_{(i,j)}$$

$$= \sum_{(a,b) \in \text{protein i의 MF} \times \text{protein j의 MF}} \frac{EvMF_{(a,b)}}{|\text{protein i의 MF} \times \text{protein j의 MF}|}$$

[0171]

[0172] 상기 수학적식 10에서 |protein i의 MF × protein j의 MF| 는 단백질 i와 단백질j의 GO MF조합의 개수를 나타낸다. 예를 들어 가상의 단백질 i에 GO MF 용어 a1, a2가 관련되어 있고(단백질i가 a1과 a2 분자 기능을 수행함), 단백질 j에 GO MF 용어 b1, b2가 있다면 (a1,b1), (a1,b2), (a2,b1), (a2,b2)의 4쌍 조합이 생기며 |protein i의 MF × protein j의 MF| 값은 4가 된다.

[0173] 다음으로, 생물학적 작용(Biological Process) 정보가 적용되는 경우에 관하여 설명한다.

[0174] GO BP는 큰 규모의 생물학적 작용을 다루기 때문에 이웃한 단백질 간에 평가값을 적용하기 보다는 신호 전달 가능 경로를 구성하는 전체 단백질들이 어떠한 생물학적 작용에 연관되어 있는지를 평가하는 것이 의미 있다고 볼 수 있다. 그리고 이러한 생물학적 작용은 경로 탐색의 입력 값인 시작과 종료 단백질이 속하는 작용이어야 할 것임을 추측할 수 있다. 따라서 GO BP 용어 a, b간의 평가값을 구하는 공식은 MF와 거의 동일한 다음 수학적식 11을 적용하나, 경로 구성 후보 단백질 i에 대한 평가 값은 이전 방식과는 상이한 다음 수학적식 12를 적용한다.

수학적식 11

$$EvBP_{(a,b)} = \{1 - \log(\text{distance} + 1)\} * \left\{ \frac{\text{partof1} + \text{partof2} + \text{isa1} + \text{isa2} + 1}{4} \right\} * BP_{\text{factor}}$$

[0175] (단, $0 \leq \text{distance} \leq 3$, $1 \leq \text{partof1}$, partof2 , isa1 , isa2)

[0176] GO BP용어가 존재 하지 않을 때의 distance와 partof, isa 값 처리는 MF와 동일하게 distance 3, partof 1, isa 1을 적용해 준다. BP 용어 정보의 중요도인 BP factor는 MF와 동일한 0.5를 적용하여 준다.

수학식 12

EvPBP_(s,t,i)

$$= \sum_{(a,b) \in \text{protein s의 BP} \times \text{protein i의 BP}} \frac{\text{EvBP}_{(a,b)}}{|\text{protein s의 BP} \times \text{protein i의 BP}|}$$

$$+ \sum_{(a,b) \in \text{protein t의 BP} \times \text{protein j의 BP}} \frac{\text{EvBP}_{(a,b)}}{|\text{protein t의 BP} \times \text{protein j의 BP}|}$$

[0177]

[0178]

다시 정리하면 상기 수학식 12는 시작 단백질 s와 경로 구성 후보 단백질 i 사이의 GO BP 평가 값의 평균과 종료 단백질 t와 경로 평가 대상 단백질 j간의 EvBP평가 값의 평균을 합한 것이다. 여기서 주목 할 점은 단백질 s와 t에 모두 속해있는 GO BP 용어들에 대해서는 2번 값이 누적된다는 점이다.

[0179]

이는 시작 단백질과 종료 단백질 모두와 연관되어 있는 생물학 작용은 좀 더 찾고자 하는 작용과 밀접 할 가능성이 있으므로 중복 용어를 제거하지 않고 누적하도록 하였다. 예를 들면, 시작 단백질 s에 GO BP 용어 a1, b1가 있고 종료 단백질 t에 a1, b2 평가 대상 단백질 i에 a1, b3가 있다고 하면 이에 대해 (a1, a1), (a1, b3), (b1, a1), (b1, b3) 사이의 EvBP 평가값 합 나누기 4 한 값과 (a1, a1), (a1, b3), (b2, a1) (b2, b3)사이의 EvBP 평가값 합 나누기 4 한 값을 더한 값이 EvPBP값이 된다.

[0180]

여기서 (a1, a1)은 동일한 용어이므로 좋은 평가값이 나올 것이 예상된다.

[0181]

또한, a1은 시작 단백질과 종료 단백질 둘 모두 가지고 있으므로 대상 단백질이 a1을 가짐으로써 (a1, a1) EvBP 평가값은 시작과 대상 단백질 사이 종료와 대상 단백질 사이에 2번 더해지게 되므로 시작, 종료 단백질 모두가 가지고 있는 BP를 평가 대상 단백질이 갖는 것은 하나의 GO BP 용어로 2배의 평가값을 얻게 되는 것이다.

[0182]

GO 데이터는 CC, MF, BP를 가지고 도 4 및 도 5에 관해 설명된 실시예에서 단백질 위치정보를 이용하여 신호전달 경로를 추출한 결과값 중 상위의 동점결과 값들을 대상으로 적용하며, GO의 CC정보를 활용하여 평가를 수행하거나, CC 정보와 MF 정보를 적용하여 평가를 수행하거나, CC정보와 MF정보 BP정보 모두를 적용하여 평가를 수행할 수 있다.

[0183]

이상에서는 본 발명의 바람직한 실시예에 대하여 도시하고 설명하였지만, 본 발명은 상술한 특징의 실시예에 한정되지 아니하며, 청구범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발명이 속하는 기술분야에서 통상의 지식을 가진자에 의해 다양한 변형실시가 가능한 것은 물론이고, 이러한 변형실시들은 본 발명의 기술적 사상이나 전망으로부터 개별적으로 이해되어져서는 안될 것이다.

도면의 간단한 설명

[0184]

도 1은 세포내 단백질 위치 정보를 나타낸 도면,

[0185]

도 2는 단백질 위치정보의 예를 도시한 도면,

[0186]

도 3은 세포 내 단백질 다중위치 분포도를 도시한 도면,

[0187]

도 4는 본 발명의 일 실시예에 따른 단백질의 위치 정보를 기반으로 단백질의 신호 경로를 추출하는 시스템에 대한 개략적인 블록도,

[0188]

도 5는 본 발명의 일 실시예에 따른 단백질의 위치 정보를 기반으로 단백질의 신호 경로를 추출하는 방법에 대한 흐름도,

[0189]

도 6은 본 발명의 일 실시예에 따른 데이터베이스 통합 방식을 도식적으로 도시한 도면,

[0190]

도 7은 본 발명의 일 실시예에 따른 단백질 상호작용을 인접행렬로 표시한 도면,

[0191]

도 8은 DFS 방식을 이용하여 탐색하는 예를 도식적으로 도시한 도면,

[0192]

도 9는 본 발명의 일 실시예에 따른 단백질 상호작용의 모든 경로를 찾는 알고리즘에 관한 도면,

[0193]

도 10은 도 9의 알고리즘의 결과를 도시한 도면,

- [0194] 도 11은 본 발명의 일 실시예에 따른 수학적 식 4의 평가 함수를 적용한 알고리즘을 도시한 도면,

[0195] 도 12는 본 발명의 일 실시예에 따른 단백질 위치의 상호작용에 대한 모든 경우의 수를 도시한 도면,

[0196] 도 13은 본 발명의 일 실시예에 따른 단백질 상호작용 위치에 대한 상호작용의 확률 값을 도시한 도면,

[0197] 도 14는 A 위치 가중치 * B 위치 가중치 * A-B 위치 상호작용 가중치를 도시한 표,

[0198] 도 15는 본 발명의 일 실시예에 따른 경로길이 9를 제외한 전체 신호전달 경로에 대한 수행 결과표,

[0199] 도 16은 본 발명에 따른 신호 전달 경로를 탐색하는데 소요된 시간을 도시한 표,

[0200] 도 17은 본 발명의 일 실시예에 따른 경로길이가 6일 때의 결과표,

[0201] 도 18은 LocSPF 시스템을 이용하여 구한 Top 스코어의 신호전달 경로의 수,

[0202] 도 19는 본 발명의 일 실시예에 따른 유전자 온톨로지를 이용하여 단백질의 신호 경로를 추출하는 시스템의 블록도이다.

[0203] 도 20은 본 발명의 일 실시예에 따른 유전자 온톨로지를 이용하여 단백질의 신호 경로를 추출하는 방법의 흐름도이다.

[0204] 도 21은 Yeast GO Slim의 CC 영역 전체(상)와 일부(중), 그리고 위치관계의 포함 관계 일부(하)를 나타내는 도면이다.

[0205] 도 22는 is-a and part-of 관계의 예(좌)와, 다중 깊이 값의 예(우)를 나타내는 도면이다.

[0206] 도 23은 Yeast GO Slim Molecular Function의 일부를 나타내는 도면이다.

[0207] 도 24는 도 4 및 도 5 실시예의 결과인 상위 12개 경로 데이터 결과 중 6개를 나타낸다.

[0208] 도 25는 계산에 사용된 GO와 GOA 정보 그리고 기초 계산 데이터의 형태를 나타낸다.

[0209] 도 26은 GO 정보 중 is-a와 part-of 관계 정보를 활용해 그래프 형태로 나타낸 도면이다.

[0210] 도 27은 GOA 정보를 나타낸다.

[0211] 도 28은 CC의 전체 is-a, part-of 깊이 정보를 나타낸다.

[0212] 도 29는 BP의 전체 is-a, part-of 깊이 정보를 나타낸다.

[0213] 도 30은 MF의 전체 is-a, part-of 깊이 정보를 나타낸다.

[0214] 도 31 내지 33은 GO DAG상에서 용어들 간의 거리 값들의 일부를 예시를 나타낸다.

[0215] * 도면의 주요 부분에 대한 부호의 설명 *

[0216] 410: 전처리부 420: 신호전달 경로 탐색부

[0217] 430: 통합부 440: 가중치부

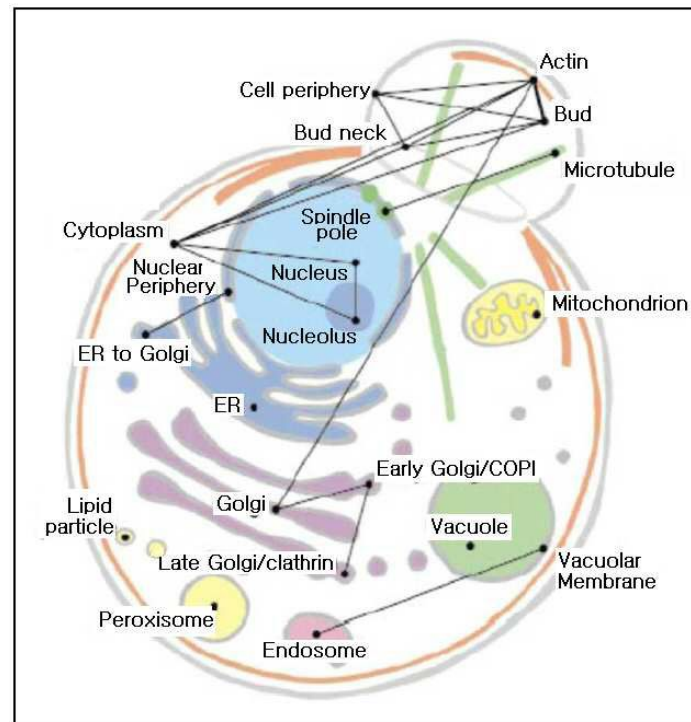
[0218] 450: 가능 경로 탐색부 460: 가중치 적용부

[0219] 470: 경로 결정부 480: GO 데이터베이스

[0220] 490: GO 적용부

도면

도면1



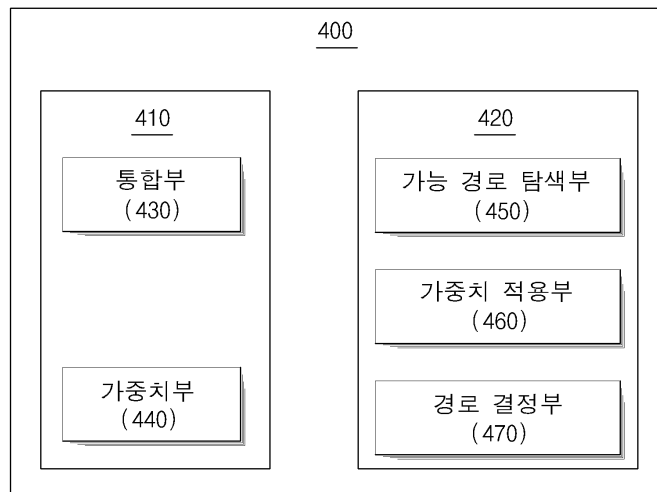
도면2

단백질이름	위치 1	위치 2	위치 3	위치 4	위치 5
YMR089C	미토콘드리아	세포질			
YIL101C	핵	세포질			
YER139C	세포질				
YPL240C	세포질				
YDL060W	핵	세포질			
YDR024W	세포질	핵			

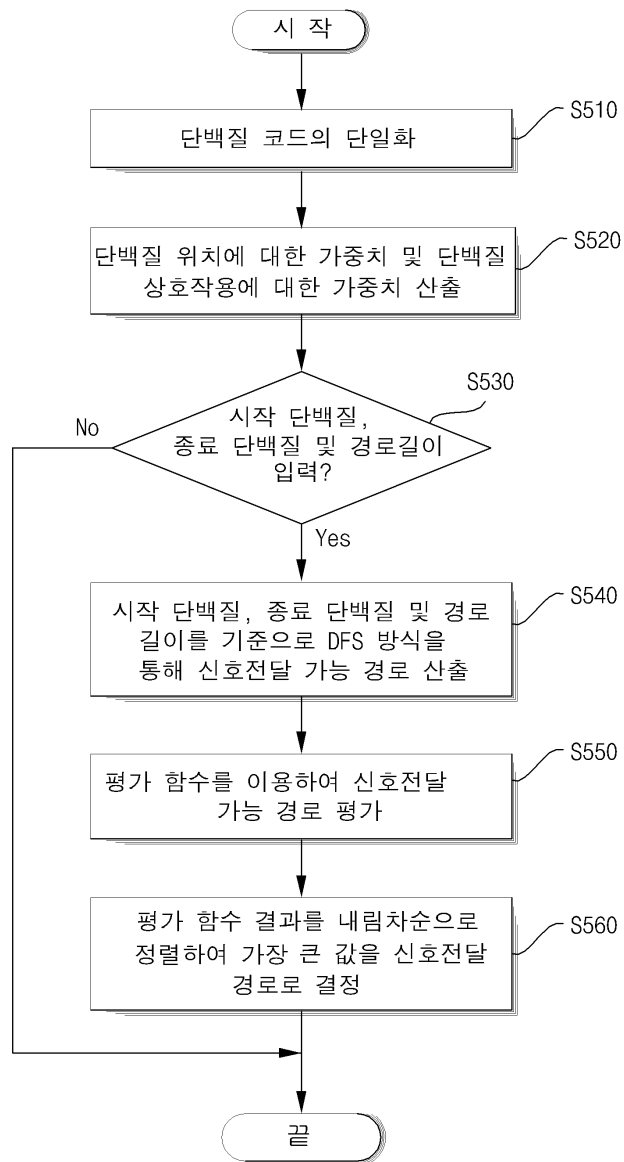
도면3

단백질 위치	단백질 개수	단백질 위치	단백질 개수
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	

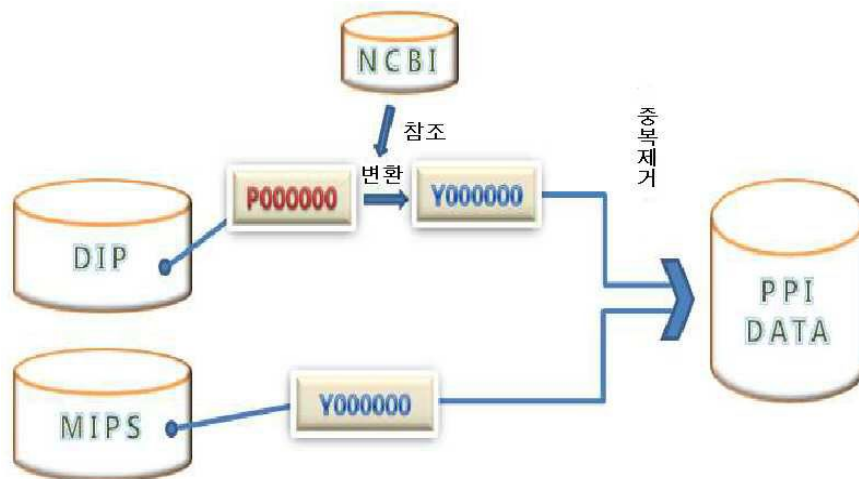
도면4



도면5



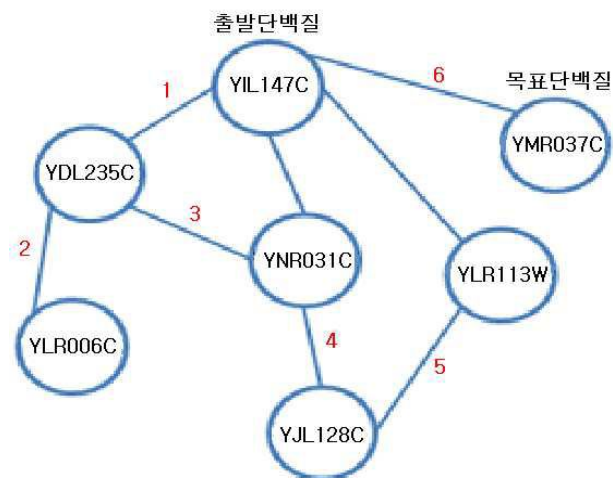
도면6



도면7

	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		

도면8



도면9

```

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: }

```

도면10

	A	B	C	D	E	F	G	H
1	start : YLR332W, target : YPL089C, maximum length : 8							
2	YLR332W	YOL109W	YBR069C	YPR19W	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)							

도면11

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.	
1:	Evaluate(node maximum length)
2:	{
3:	
4:	for all path $\in$ paths do
5:	for (protein1, protein2) $\in$ path until maximum length do
6:	Multiply score with value of EvaluationEach(protein1, protein2)
7:	&
8:	Save into score variable
9:	end for
10:	end for
11:	}
//Evaluate the score between two proteins	
1:	EvaluateEach(protein1, protein2)
2:	{
3:	
4:	if protein1.locations = no location or protein2.locations = no location
5:	return 1.0f
6:	else if protein1.locations $\neq$ no location and protein2.locations $\neq$ no location
7:	choose max value from criterion each protein1.locations and protein2.locations
8:	return max value
9:	}

도면12

	A	B	C	D	E	F	G	H	I	J
1	Y56A3A2-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-YLR148W									
8	endosome(13)-golgi(10) endosome(13)-vacuole(19) endosome(13)-vacuolar membrane(16) 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)									
17	YAL003W-YML064C									
18	cytoplasm(5)-er(15) cytoplasm(5)-spindle pole(18)									

도면13

	golgi	microtubule	bud neck	endosome	nuclear periphery	er	vacuolar membrane	cell periphery	spindle pole	vacuole	mitochondrion
bud	0.0401	0.1126	0.1168	0.0035	0.0084	0.1259	0.0141	0.1140	0.0232	0.0408	0.0957
nucleus	0.4981	0.7161	0.3306	0.0957	0.4601	1.3542	0.1716	0.4101	0.5121	0.4284	1.8178
lipid particle	0.0077	0.0021	0.0021	0.0007	0.0028	0.0127	0.0000	0.0021	0.0014	0.0035	0.0056
early golgi	0.1435	0.0183	0.0091	0.0190	0.0063	0.1597	0.0218	0.0134	0.0070	0.0443	0.0478
er to golgi	0.0091	0.0014	0.0007	0.0000	0.0028	0.0141	0.0007	0.0014	0.0000	0.0007	0.0021
cytoplasm	0.8688	1.1185	0.6423	0.1611	0.4017	2.0134	0.3095	0.7471	0.5762	0.6725	2.0345
nucleolus	0.0345	0.0324	0.0183	0.0056	0.0218	0.0689	0.0113	0.0148	0.0176	0.0239	0.1646
actin	0.0380	0.1674	0.0457	0.0091	0.0028	0.0661	0.0056	0.0457	0.0077	0.0225	0.0739
peroxisome	0.0063	0.0035	0.0063	0.0021	0.0098	0.0302	0.0035	0.0049	0.0021	0.0070	0.0232
late golgi	0.1112	0.0317	0.0162	0.0091	0.0056	0.0689	0.0162	0.0246	0.0056	0.0310	0.0577
golgi	0.4854	0.0844	0.0514	0.0654	0.0246	0.4896	0.0696	0.0809	0.0246	0.1569	0.1949
microtubule	0.0844	0.5375	0.1850	0.0098	0.0197	0.1316	0.0120	0.1541	0.2026	0.0478	0.1899
bud neck	0.0514	0.1850	0.1787	0.0056	0.0127	0.0907	0.0120	0.1505	0.0394	0.0464	0.1273
endosome	0.0654	0.0098	0.0056	0.0422	0.0042	0.0626	0.0246	0.0120	0.0070	0.0457	0.0280
nuclear periphery	0.0246	0.0197	0.0127	0.0042	0.1548	0.1104	0.0063	0.0239	0.0204	0.0218	0.1013
er	0.4896	0.1316	0.0907	0.0626	0.1104	1.9571	0.1632	0.2884	0.0556	0.4242	0.5656
vacuolar membrane	0.0696	0.0120	0.0120	0.0246	0.0063	0.1632	0.0816	0.0253	0.0070	0.1112	0.0780
cell periphery	0.0809	0.1541	0.1505	0.0120	0.0239	0.2884	0.0253	0.1815	0.0274	0.0893	0.1864
spindle pole	0.0246	0.2026	0.0394	0.0070	0.0204	0.0556	0.0070	0.0274	0.2490	0.0204	0.1189
vacuole	0.1569	0.0478	0.0464	0.0457	0.0218	0.4242	0.1112	0.0893	0.0204	0.1928	0.1773
mitochondrion	0.1949	0.1899	0.1273	0.0280	0.1013	0.5656	0.0760	0.1864	0.1189	0.1773	0.9356

도면14

	bud	nucleus	lipid particle	early golgi	er to golgi	cytoplasm	nucleolus	actin	peroxisome	late golgi	golgi
bud	763	43,356	6	27	1	114,705	231	77	23	71	635
nucleus	43,356	38,690,790	1,208	12,465	160	44,778,920	485,743	10,811	6,219	10,360	179,393
lipid particle	6	1,208	0	2	0	2,584	23	2	1	2	36
early golgi	27	12,465	2	260	0	33,009	80	17	2	37	1,421
er to golgi	1	160	0	0	0	473	2	0	0	0	10
cytoplasm	114,705	44,778,920	2,584	33,009	473	70,130,553	403,366	29,468	16,219	25,633	416,920
nucleolus	231	485,743	23	80	2	403,366	33,218	51	14	161	1,284
actin	77	10,811	2	17	0	29,468	51	134	1	22	219
peroxisome	23	6,219	1	2	0	16,219	14	1	174	3	54
late golgi	71	10,360	2	37	0	25,623	161	22	3	98	920
golgi	635	179,393	36	1,421	10	416,920	1,284	219	54	920	15,727
microtubule	1,159	167,662	6	113	1	348,896	784	627	19	170	1,778
bud neck	1,017	65,499	5	50	0	169,519	375	145	29	74	915
endosome	18	10,912	1	60	0	24,481	66	17	6	24	671
nuclear periphery	46	57,078	5	22	1	66,396	280	6	29	16	275
er	6,128	1,498,498	182	4,857	47	2,968,302	7,892	1,170	786	1,754	48,737
vacuolar membrane	118	32,630	0	114	0	78,395	221	17	16	71	1,191
cell periphery	2,096	171,520	11	154	2	416,279	639	306	48	237	3,043
spindle pole	165	83,007	3	31	0	124,418	295	20	8	21	359
vacuole	790	188,599	20	538	1	394,452	1,089	158	73	313	6,212
mitochondrion	7,906	3,415,532	137	2,471	12	5,093,041	31,997	2,220	1,025	2,492	32,936

	microtubule	bud neck	endosome	nuclear periphery	er	vacuolar membrane	cell periphery	spindle pole	vacuole	mitochondrion
bud	1,159	1,017	18	46	6,128	118	2,096	165	790	7,906
nucleus	167,662	65,499	10,912	57,078	1,498,498	32,630	171,520	83,007	188,599	3,415,532
lipid particle	6	5	1	5	182	0	11	3	20	137
early golgi	118	50	60	22	4,857	114	154	31	536	2,471
er to golgi	1	0	0	1	47	0	2	0	1	12
cytoplasm	348,896	169,519	24,481	66,396	2,968,302	78,395	416,279	124,418	394,452	5,093,041
nucleolus	784	375	66	280	7,892	221	639	295	1,089	31,997
actin	627	145	17	6	1,170	17	306	20	158	2,220
peroxisome	19	29	6	29	786	16	48	8	73	1,025
late golgi	170	74	24	16	1,754	71	237	21	313	2,492
golgi	1,778	915	671	275	48,737	1,191	3,043	359	6,212	32,936
microtubule	7,357	2,143	66	143	8,511	133	3,767	1,920	1,231	20,867
bud neck	2,143	1,751	32	78	4,968	112	3,115	316	1,011	11,837
endosome	66	32	137	15	1,974	133	142	32	573	1,393
nuclear periphery	143	78	15	595	3,787	37	310	102	297	5,898
er	8,511	4,968	1,974	3,787	598,495	8,574	33,336	2,489	51,608	293,697
vacuolar membrane	133	112	133	37	8,574	736	503	54	2,323	6,777
cell periphery	3,767	3,115	142	310	33,336	503	7,928	464	4,108	36,586
spindle pole	1,920	316	32	102	2,489	54	464	1,634	364	9,043
vacuole	1,231	1,011	573	297	51,608	2,323	4,108	364	9,329	36,622
mitochondrion	20,867	11,837	1,393	5,898	293,697	6,777	36,586	9,043	36,622	824,966

도면15

KEGG_단백질수	실험_단백질수	찾은_단백질수	방법1_Pathway수	방법1(%)	방법2_Pathway수	방법2(%)
6	6	6	3	75%	3	75%
		5	1	25%	1	25%
		합계	4	100%	4	100%
7	8	7	3	33%	3	33%
		5	3	33%	4	44%
		4	2	22%	1	11%
		2	1	11%	1	11%
		합계	9	100%	9	100%

도면16

시작단백질	종료단백질	KEGG-Depth	본_System	Program_Start	Program_End	OutputFile_Size	소요시간
YER118C	YMR037C	6	6	2008-11-11 5:42 PM	2008-11-11 6:33 PM	2.12k	0시간 51분
YER118C	YKL062W	6	6	2008-11-11 5:45 PM	2008-11-11 6:43 PM	10.6k	0시간 58분
YER118C	YMR043W	6	6	2008-11-11 5:47 PM	2008-11-11 6:38 PM	34.3k	0시간 51분
YER118C	YGR040W	6	6	2008-11-11 5:49 PM	2008-11-11 6:41 PM	175k	0시간 52분
YLR332W	YPL089C	7	8	2008-11-04 4:23 AM	2008-11-04 4:38 PM	639k	12시간 15분
YLR332W	YER111C	7	8	2008-11-04 4:31 AM	2008-11-04 6:26 PM	1.15m	13시간 55분
YLR332W	YLR182W	7	8	2008-11-04 4:27 AM	2008-11-04 4:39 PM	2.13m	12시간 12분
YIL147C	YMR037C	7	8	2008-11-04 4:15 AM	2008-11-06 10:06 PM	215k	17시간 51분
YIL147C	YKL062W	7	8	2008-11-04 4:16 AM	2008-11-05 9:55 AM	506k	5시간 39분
YIL147C	YMR043W	7	8	2008-11-04 4:21 AM	2008-11-05 10:54 AM	2.32m	6시간 33분

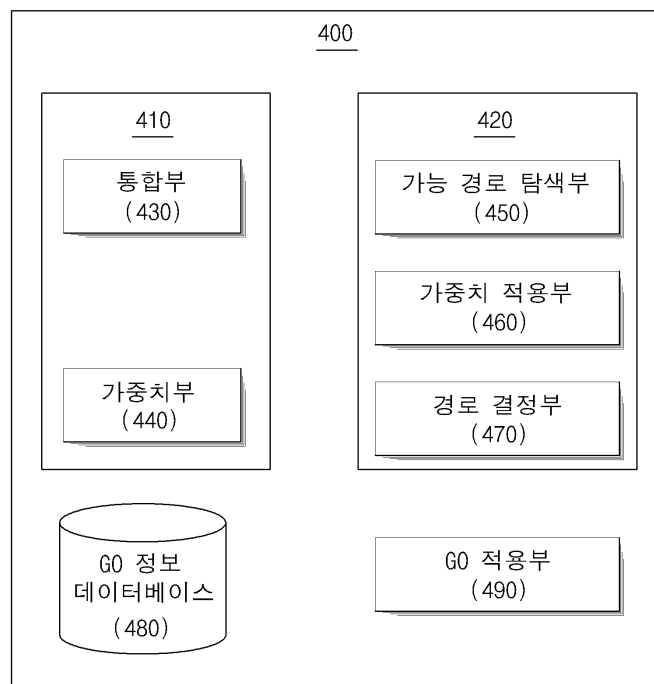
도면17

Function	경로	구분	Pathway						방법1_찾은노드수	방법2_찾은노드수
			1	2	3	4	5	6		
High osmolarity	경로1	MAPK	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YMR043W	6	6
		LocSPF_방법1	YIL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	6	6
		LocSPF_방법2	YIL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	6	6
	경로2	MAPK	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YMR037C	6	6
		LocSPF_방법1	YIL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	6	6
		LocSPF_방법2	YIL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	6	6
	경로3	MAPK	YER118C	YHL007C	YLR362W	YJL128C	YLR113W	YKL062W	6	6
		LocSPF_방법1	YIL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	6	6
		LocSPF_방법2	YIL147C	YDL235C	YLR006C	YNR031C	YJL128C	YLR113W	6	6

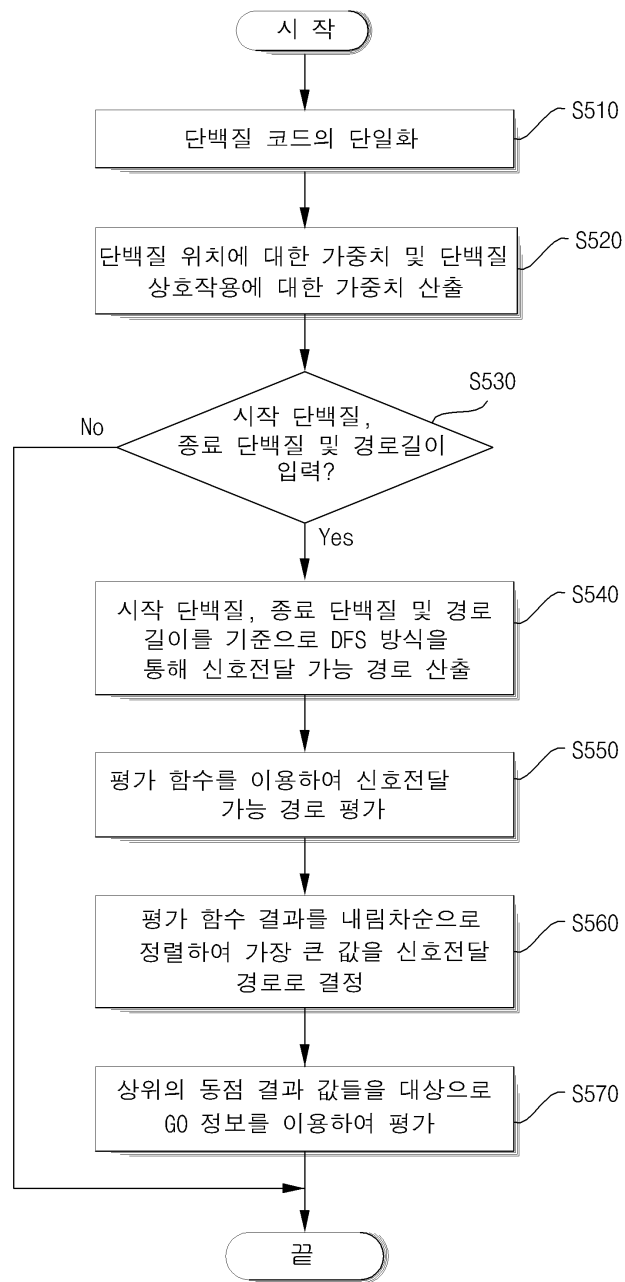
도면18

Function	KEGG_경로길이	LocSPF_경로길이	경로	구분	Start 단백질	Target 단백질	Top스코어 중복데이터_수
High Osmolarity	7	8	경로1	LocSPF_방법1	YIL147C	YKL062W	285
				LocSPF_방법2	YIL147C	YKL062W	285
			경로2	LocSPF_방법1	YIL147C	YMR037C	181
				LocSPF_방법2	YIL147C	YMR037C	181
			경로3	LocSPF_방법1	YIL147C	YMR043W	8
				LocSPF_방법2	YIL147C	YMR043W	1218
	6	6	경로1	LocSPF_방법1	YER118C	YMR043W	9
				LocSPF_방법2	YER118C	YMR043W	27
			경로2	LocSPF_방법1	YER118C	YMR037C	14
				LocSPF_방법2	YER118C	YMR037C	15
			경로3	LocSPF_방법1	YER118C	YKL062W	26
				LocSPF_방법2	YER118C	YKL062W	26

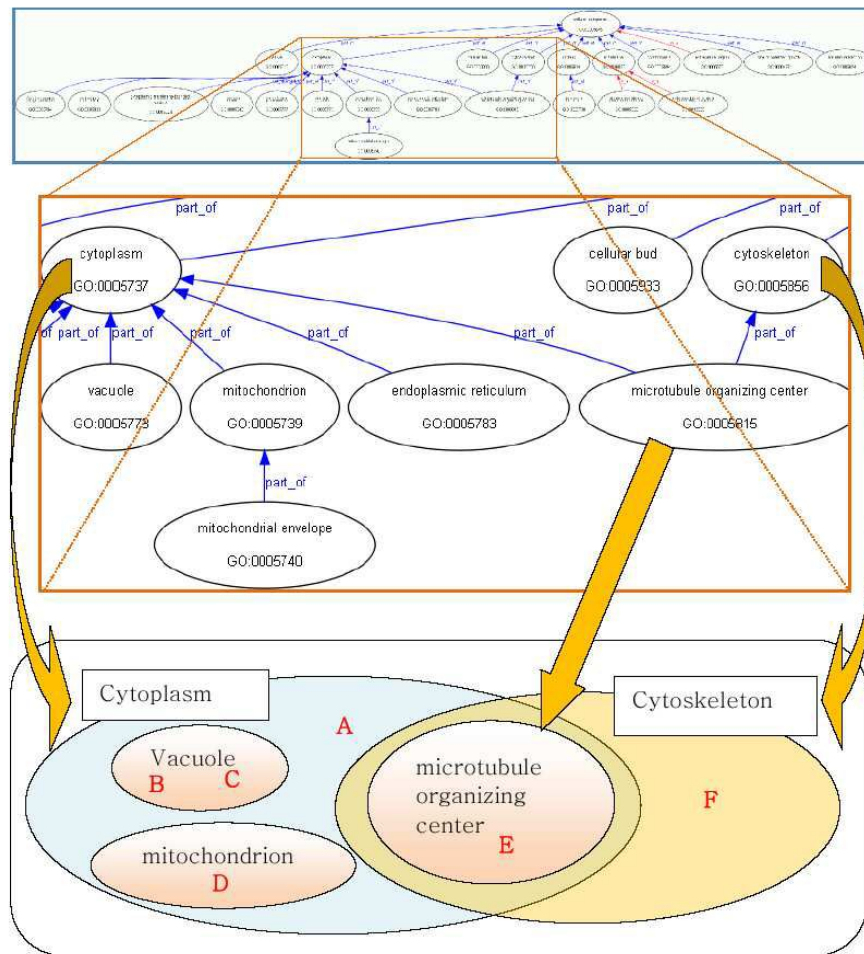
도면19



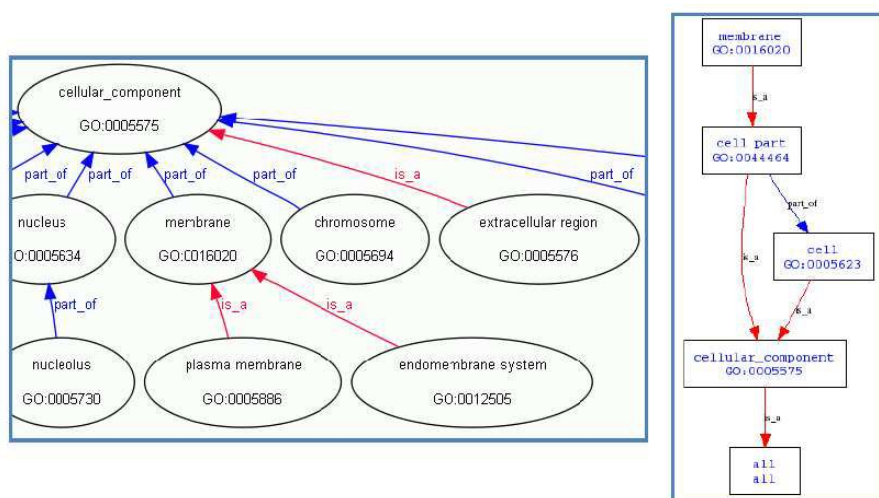
도면20



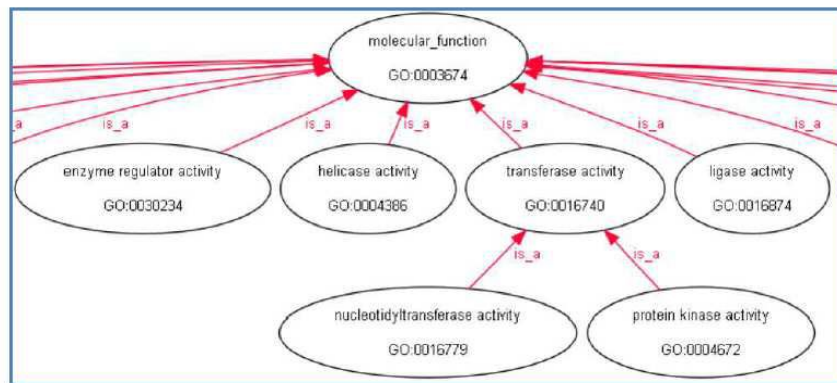
도면21



도면22



도면23



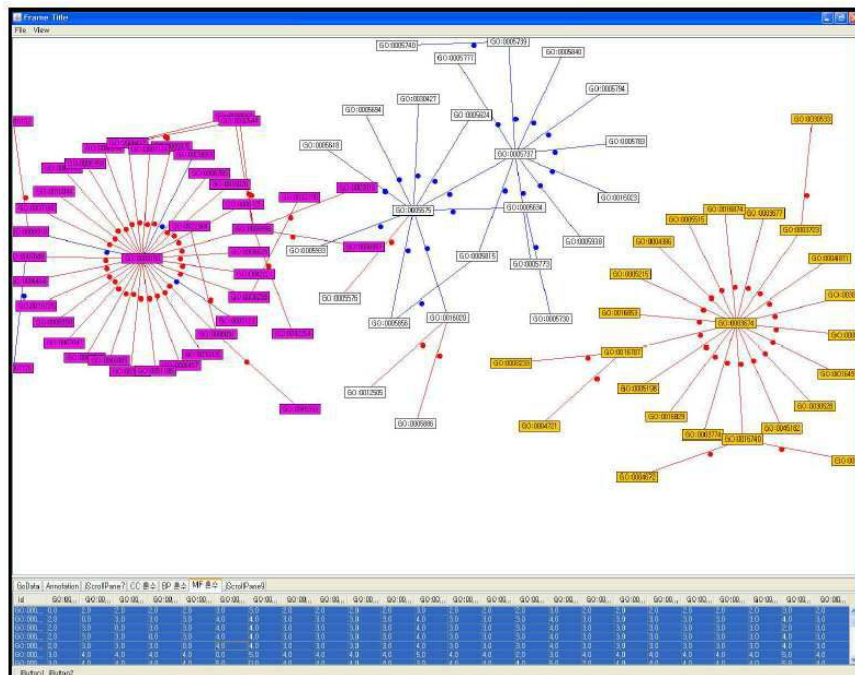
도면24

Function	KEGG_경로길이	LocSPF_경로길이	경로	구분	Start 단백질		Target 단백질		Top스코어 중복데이터 수
High Osmolarity	7	8	경로1	LocSPF_방법1	VIL147C	SLN1	YKL062W	MSN4	285
				LocSPF_방법2	VIL147C	SLN1	YKL062W	MSN4	285
			경로2	LocSPF_방법1	VIL147C	SLN1	VMR037C	MSN2	181
				LocSPF_방법2	VIL147C	SLN1	VMR037C	MSN2	181
			경로3	LocSPF_방법1	VIL147C	SLN1	VMR043W	MCM1	8
				LocSPF_방법2	VIL147C	SLN1	VMR043W	MCM1	1218
	6	6	경로1	LocSPF_방법1	YER118C	SHO1	VMR043W	MCM1	9
				LocSPF_방법2	YER118C	SHO1	VMR043W	MCM1	27
			경로2	LocSPF_방법1	YER118C	SHO1	YMR037C	MSN2	14
				LocSPF_방법2	YER118C	SHO1	YMR037C	MSN2	15
			경로3	LocSPF_방법1	YER118C	SHO1	YKL062W	MSN4	26
				LocSPF_방법2	YER118C	SHO1	YKL062W	MSN4	26

도면25

	A	B	C	D	E	F	G	H	I	J
1	id	NAME	NAMESPA	ISABSOLU	IS_A	RELATION	SUBSET	REPLACED	ALT_ID	ANNOTATION_ID
2	GO:0000746	conjugation	biological_process	FALSE	GO:0008150,		goslim_candida.goslim_plr.goslim_yeast.gosubset_prok,			S000005327.S000003000.S000003706.S000001277.S000004306.S000003693.S000002044.S000000685.S000000532.S000004845.S000000112.S000002717.S000002382.S000000560.S000002869.S000002285.S000005223.S000004510.S000002384.S000004305.S000004444.S000001126.S000001868.
	GO:0000910	cytokinesis	biological_process	FALSE	GO:0008150,		goslim_candida.goslim_yeast.gosubset_prok,	GO:0007104,GO:0016288,		S000001855.S000004965.S000005110.S000000520.S000004311.S000000595.S000003837.S000001149.S000000072.S000004306.S000000242.S000000227.S000005242.S000002275.S000005271.S000001531.S000005028.S000003621.S00000306.S000004635.S0000006163.S000005707.S000003074.S000001065.S000001445.S000001357.S000000653.S000004156.S000003202.S000000810.S000001889.S000002573.S000001330.S000006259.S000002384.S000002736.S000000157.S000000347.S000001177.S000005098.
3	GO:0003674	molecular_function	molecular_function	FALSE			goslim_candida.goslim_gonerl.goslim_gosli_m_plr.goslim_plant.goslim_yeast.gosubset_prok,		GO:0005554,	S000000170.S000003087.S000003869.S000003843.S000002362.S000002919.S000004796.S000001689.S000004668.S000004895.S000005926.S000000778.S000005655.S000000678.S0000007266.S000000044.S000002960.S000001907.S000001013.S000001241.S000000047.S000001420.S000001442.S000003582.S000003667.S000003841.S000003861.S000001520.S000001724.S000001782.S000000312.S000004492.S000004514.S000004552.S000004605.S000004766.S000005044.S000005301.S000005414.S000000398.S000005741.S000005812.S000006020.S000006079.S000006208.S000001242.S000000466.S000002470.S000002901.S000000882.S000003658.

도면26



도면27

[illegible]

도면28

	A	B	C	D	E
1	CC				
2	GO:0005575	degreeOfDepth	0	degreeOfPartOf	0
3	GO:0005576	degreeOfDepth	1	degreeOfPartOf	0
4	GO:0005618	degreeOfDepth	1	degreeOfPartOf	1
5	GO:0005624	degreeOfDepth	1	degreeOfPartOf	1
6	GO:0005634	degreeOfDepth	1	degreeOfPartOf	1
7	GO:0005694	degreeOfDepth	1	degreeOfPartOf	1
8	GO:0005730	degreeOfDepth	2	degreeOfPartOf	2
9	GO:0005737	degreeOfDepth	1	degreeOfPartOf	1
10	GO:0005739	degreeOfDepth	2	degreeOfPartOf	2
11	GO:0005740	degreeOfDepth	3	degreeOfPartOf	3
12	GO:0005773	degreeOfDepth	2	degreeOfPartOf	2
13	GO:0005777	degreeOfDepth	2	degreeOfPartOf	2
14	GO:0005783	degreeOfDepth	2	degreeOfPartOf	2
15	GO:0005794	degreeOfDepth	2	degreeOfPartOf	2
16	GO:0005815	degreeOfDepth	2	degreeOfPartOf	2
17	GO:0005840	degreeOfDepth	2	degreeOfPartOf	2
18	GO:0005856	degreeOfDepth	1	degreeOfPartOf	1
19	GO:0005886	degreeOfDepth	2	degreeOfPartOf	1
20	GO:0005933	degreeOfDepth	1	degreeOfPartOf	1
21	GO:0005938	degreeOfDepth	2	degreeOfPartOf	2
22	GO:0012505	degreeOfDepth	2	degreeOfPartOf	1
23	GO:0016020	degreeOfDepth	1	degreeOfPartOf	1
24	GO:0016023	degreeOfDepth	2	degreeOfPartOf	2
25	GO:0030427	degreeOfDepth	1	degreeOfPartOf	1

도면29

	A	B	C	D	E
26	BP				
27	GO:0000746	degreeOfDepth	1 degreeOfPartOf	0	
28	GO:0000910	degreeOfDepth	1 degreeOfPartOf	0	
29	GO:0005975	degreeOfDepth	1 degreeOfPartOf	0	
30	GO:0006091	degreeOfDepth	1 degreeOfPartOf	0	
31	GO:0006259	degreeOfDepth	1 degreeOfPartOf	0	
32	GO:0006350	degreeOfDepth	1 degreeOfPartOf	0	
33	GO:0006412	degreeOfDepth	1 degreeOfPartOf	0	
34	GO:0006457	degreeOfDepth	1 degreeOfPartOf	0	
35	GO:0006464	degreeOfDepth	1 degreeOfPartOf	0	
36	GO:0006519	degreeOfDepth	1 degreeOfPartOf	0	
37	GO:0006629	degreeOfDepth	1 degreeOfPartOf	0	
38	GO:0006725	degreeOfDepth	1 degreeOfPartOf	0	
39	GO:0006766	degreeOfDepth	1 degreeOfPartOf	0	
40	GO:0006810	degreeOfDepth	1 degreeOfPartOf	1	
41	GO:0006950	degreeOfDepth	1 degreeOfPartOf	0	
42	GO:0006996	degreeOfDepth	1 degreeOfPartOf	0	
43	GO:0006997	degreeOfDepth	2 degreeOfPartOf	0	
44	GO:0007010	degreeOfDepth	2 degreeOfPartOf	0	
45	GO:0007047	degreeOfDepth	1 degreeOfPartOf	0	
46	GO:0007049	degreeOfDepth	1 degreeOfPartOf	0	
47	GO:0007114	degreeOfDepth	1 degreeOfPartOf	0	
48	GO:0007124	degreeOfDepth	1 degreeOfPartOf	1	
49	GO:0007126	degreeOfDepth	2 degreeOfPartOf	1	
50	GO:0007165	degreeOfDepth	1 degreeOfPartOf	0	
51	GO:0008150	degreeOfDepth	0 degreeOfPartOf	0	
52	GO:0009653	degreeOfDepth	1 degreeOfPartOf	1	
53	GO:0016044	degreeOfDepth	1 degreeOfPartOf	0	
54	GO:0016070	degreeOfDepth	1 degreeOfPartOf	0	
55	GO:0016192	degreeOfDepth	2 degreeOfPartOf	1	
56	GO:0019725	degreeOfDepth	1 degreeOfPartOf	0	
57	GO:0022904	degreeOfDepth	2 degreeOfPartOf	0	
58	GO:0030163	degreeOfDepth	1 degreeOfPartOf	0	
59	GO:0030435	degreeOfDepth	1 degreeOfPartOf	0	
60	GO:0032196	degreeOfDepth	2 degreeOfPartOf	0	
61	GO:0032543	degreeOfDepth	2 degreeOfPartOf	0	
62	GO:0032544	degreeOfDepth	2 degreeOfPartOf	0	
63	GO:0042221	degreeOfDepth	1 degreeOfPartOf	0	
64	GO:0042254	degreeOfDepth	2 degreeOfPartOf	0	
65	GO:0045333	degreeOfDepth	2 degreeOfPartOf	0	
66	GO:0046483	degreeOfDepth	1 degreeOfPartOf	0	
67	GO:0051186	degreeOfDepth	1 degreeOfPartOf	0	

도면30

	A	B	C	D	E
68	MF				
69	GO:0003674	degreeOfDepth	0	degreeOfPartOf	0
70	GO:0003677	degreeOfDepth	1	degreeOfPartOf	0
71	GO:0003723	degreeOfDepth	1	degreeOfPartOf	0
72	GO:0003774	degreeOfDepth	1	degreeOfPartOf	0
73	GO:0004386	degreeOfDepth	1	degreeOfPartOf	0
74	GO:0004672	degreeOfDepth	2	degreeOfPartOf	0
75	GO:0004721	degreeOfDepth	2	degreeOfPartOf	0
76	GO:0004871	degreeOfDepth	1	degreeOfPartOf	0
77	GO:0005198	degreeOfDepth	1	degreeOfPartOf	0
78	GO:0005215	degreeOfDepth	1	degreeOfPartOf	0
79	GO:0005515	degreeOfDepth	1	degreeOfPartOf	0
80	GO:0008233	degreeOfDepth	2	degreeOfPartOf	0
81	GO:0008289	degreeOfDepth	1	degreeOfPartOf	0
82	GO:0016491	degreeOfDepth	1	degreeOfPartOf	0
83	GO:0016740	degreeOfDepth	1	degreeOfPartOf	0
84	GO:0016779	degreeOfDepth	2	degreeOfPartOf	0
85	GO:0016787	degreeOfDepth	1	degreeOfPartOf	0
86	GO:0016829	degreeOfDepth	1	degreeOfPartOf	0
87	GO:0016853	degreeOfDepth	1	degreeOfPartOf	0
88	GO:0016874	degreeOfDepth	1	degreeOfPartOf	0
89	GO:0030234	degreeOfDepth	1	degreeOfPartOf	0
90	GO:0030528	degreeOfDepth	1	degreeOfPartOf	0
91	GO:0030533	degreeOfDepth	2	degreeOfPartOf	0
92	GO:0045182	degreeOfDepth	1	degreeOfPartOf	0

도면31

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		GO:0005575	GO:0005576	GO:0005618	GO:0005624	GO:0005634	GO:0005694	GO:0005730	GO:0005737	GO:0005739	GO:0005740	GO:0005773	GO:0005777	GO:0005783	GO:0005794	GO:0005815
2	GO:0005575	0	1	1	1	1	1	2	1	2	3	2	2	2	2	2
3	GO:0005576	1	0	2	2	2	2	3	2	3	4	3	3	3	3	3
4	GO:0005618	1	2	0	2	2	2	3	2	3	4	3	3	3	3	3
5	GO:0005624	1	2	2	0	2	2	3	2	3	4	3	3	3	3	3
6	GO:0005634	1	2	2	2	0	2	1	2	3	4	3	3	3	3	3
7	GO:0005694	1	2	2	2	2	0	3	2	3	4	3	3	3	3	3
8	GO:0005730	2	3	3	3	1	3	0	3	4	5	4	4	4	4	4
9	GO:0005737	1	2	2	2	2	2	3	0	1	2	1	1	1	1	1
10	GO:0005739	2	3	3	3	3	3	4	1	0	1	2	2	2	2	2
11	GO:0005740	3	4	4	4	4	4	5	2	1	0	3	3	3	3	3
12	GO:0005773	2	3	3	3	3	3	4	1	2	3	0	2	2	2	2
13	GO:0005777	2	3	3	3	3	3	4	1	2	3	2	0	2	2	2
14	GO:0005783	2	3	3	3	3	3	4	1	2	3	2	2	0	2	2
15	GO:0005794	2	3	3	3	3	3	4	1	2	3	2	2	2	0	2
16	GO:0005815	2	3	3	3	3	3	4	1	2	3	2	2	2	2	0
17	GO:0005840	2	3	3	3	3	3	4	1	2	3	2	2	2	2	2
18	GO:0005856	1	2	2	2	2	2	3	2	3	4	3	3	3	3	1
19	GO:0005886	2	3	3	3	3	3	4	3	4	5	4	4	4	4	4
20	GO:0005933	1	2	2	2	2	2	3	2	3	4	3	3	3	3	3
21	GO:0005938	2	3	3	3	3	3	4	1	2	3	2	2	2	2	2
22	GO:0012505	2	3	3	3	3	3	4	3	4	5	4	4	4	4	4

도면32

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		GO:0000746	GO:0000910	GO:0005975	GO:0006091	GO:0006259	GO:0006350	GO:0006412	GO:0006457	GO:0006464	GO:0006519	GO:0006629	GO:0006725	GO:0006766	GO:0006810	GO:0006950	GO:0006996
2	GO:0000746	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	GO:0000910	2	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2
4	GO:0005975	2	2	0	2	2	2	2	2	2	2	2	2	2	2	2	2
5	GO:0006091	2	2	2	0	2	2	2	2	2	2	2	2	2	2	2	2
6	GO:0006259	2	2	2	2	0	2	2	2	2	2	2	2	2	2	2	2
7	GO:0006350	2	2	2	2	2	0	2	2	2	2	2	2	2	2	2	2
8	GO:0006412	2	2	2	2	2	2	0	2	2	2	2	2	2	2	2	2
9	GO:0006457	2	2	2	2	2	2	2	0	2	2	2	2	2	2	2	2
10	GO:0006464	2	2	2	2	2	2	2	2	0	2	2	2	2	2	2	2
11	GO:0006519	2	2	2	2	2	2	2	2	2	0	2	2	2	2	2	2
12	GO:0006629	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2	2
13	GO:0006725	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2
14	GO:0006766	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2
15	GO:0006810	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2
16	GO:0006950	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2
17	GO:0006996	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0
18	GO:0006997	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1
19	GO:0007010	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	1
20	GO:0007047	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
21	GO:0007049	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
22	GO:0007114	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

도면33

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		GO:00036 74	GO:00036 77	GO:00037 23	GO:00037 74	GO:00043 86	GO:00046 72	GO:00047 21	GO:00048 71	GO:00051 98	GO:00052 15	GO:00055 15	GO:00082 33	GO:00082 89	GO:00164 91	GO:00167 40	GO:00167 79
2	GO:00036 74	0	1	1	1	1	2	2	1	1	1	1	2	1	1	1	2
3	GO:00036 77	1	0	2	2	2	3	3	2	2	2	2	3	2	2	2	3
4	GO:00037 23	1	2	0	2	2	3	3	2	2	2	2	3	2	2	2	3
5	GO:00037 74	1	2	2	0	2	3	3	2	2	2	2	3	2	2	2	3
6	GO:00043 86	1	2	2	2	0	3	3	2	2	2	2	3	2	2	2	3
7	GO:00046 72	2	3	3	3	3	0	4	3	3	3	3	4	3	3	1	2
8	GO:00047 21	2	3	3	3	3	4	0	3	3	3	3	2	3	3	3	4
9	GO:00048 71	1	2	2	2	2	3	3	0	2	2	2	3	2	2	2	3
10	GO:00051 98	1	2	2	2	2	3	3	2	0	2	2	3	2	2	2	3
11	GO:00052 15	1	2	2	2	2	3	3	2	2	0	2	3	2	2	2	3
12	GO:00055 15	1	2	2	2	2	3	3	2	2	2	0	3	2	2	2	3
13	GO:00082 33	2	3	3	3	3	4	2	3	3	3	3	0	3	3	3	4
14	GO:00082 89	1	2	2	2	2	3	3	2	2	2	2	3	0	2	2	3
15	GO:00164 91	1	2	2	2	2	3	3	2	2	2	2	3	2	0	2	3
16	GO:00167 40	1	2	2	2	2	1	3	2	2	2	2	3	2	2	0	1
17	GO:00167 79	2	3	3	3	3	2	4	3	3	3	3	4	3	3	1	0
18	GO:00167 87	1	2	2	2	2	3	1	2	2	2	2	1	2	2	2	3
19	GO:00168 29	1	2	2	2	2	3	3	2	2	2	2	3	2	2	2	3
20	GO:00168 53	1	2	2	2	2	3	3	2	2	2	2	3	2	2	2	3
21	GO:00168 74	1	2	2	2	2	3	3	2	2	2	2	3	2	2	2	3
22	GO:00302 34	1	2	2	2	2	3	3	2	2	2	2	3	2	2	2	3