SYSTEM AND METHOD FOR CONSTRUCTING PATHWAY NETWORKS OF TWO OR MORE PATHWAYS FOR A SET OF GENES OR PROTEINS

Field of the Invention

[0001] The present invention relates generally to the field of bioinformatics, and more specifically to a system and method for constructing pathway networks of two or more pathways for a set of genes or proteins.

Background of the Invention

[0002] Gene expressions and protein function regulations in a cell or organism are events that the cell or organism orchestrates to correlate itself with its intrinsic states including developmental stages, physiological/pathological conditions and genetic backgrounds, and/or to cope with ever changing environment conditions. Measuring differential expressions of genes from organisms subjected to different stresses, from cells or organisms at varying developmental stages, or from different tissues, provides information instrumental in understanding the relationships between genes and their functions.

[0003] Recent technological advancement has made it possible to profile gene expressions on a genomic scale. Gene expression profiling can be carried out by one of a number of different technologies, such as microarray or GeneChip technologies, which typically measure the expression level of thousands of genes simultaneously using an array of cDNA clones, PCR products, or oligonucleotides bound to a solid support, such as a glass

04 MAY 2006

microscope slide or silicon surface. Since thousands of hybridisation reactions may occur in a single array, gene expression profiling assays generate huge data sets that are not amenable to simple analysis. Moreover, unlimited gene expression profiles can be procured through gene expression profiling. For example, the gene expression profiles could represent the differential expressions in many paired situations such as normal-disease, disease 1-disease 2, developmental stage 1-developmental stage 2, type 1 tissue-type 2 tissue, species 1-species 2, genetic background 1-genetic background 2, etc. In addition, proteonomics is on its way to generate even more data of protein profiles that cannot be handled by simple methods. Databases of metabolic, biochemical, signalling and the like have been constructed through vast amount of data from experiments done in wet laboratories. To maximize the use of such data, efforts are underway to develop algorithms for interpreting and interconnecting results from different genes under different conditions.

[0004] Organizing one or more genes or proteins into functional pathways has been proven a powerful tool in studying the relationships among different genes or proteins. Pathways have been used intensively in studies on how genes and/or proteins function in various organisms including bacteria, yeast, fungi, insects, animals and plants. In the literature, different bioinformatics approaches have been applied to construct pathways and pathway databases (Nakao et al., 1999, Genomescale gene expression analysis and pathway reconstruction in KEGG, Genome Informatics, 10:94—103; Hiroyuki et al., 2000, A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, Nucleic Acids Research, 28: 4021—4028; Kurhekar et al., 2002, Genome-wide pathway analysis and visualization using gene expression data, PSB 7; Rojas et al.,

2002, A database system for the analysis of biochemical pathways, In Silico Biol. 2: 75—86; Schreiber, 2002, High quality visualization of biochemical pathways in BioPath, In Silico Biology (2); Krishnamurthy et al., 2003, Pathways database system: an integrated system for biological pathways, Bioinformatics, 19: 930—7; Schreiber, 2003, Comparison of metabolic pathways using constraint graph drawing, In Proc. First Asia-Pacific Bioinformatics Conference (APBC2003)).

[0005] A biological system, even in a single cell such as a yeast or mammalian cell, is an immensely complicated set of interacting pathways involving environmental inputs, protein interactions, gene regulations, and so forth. Improved knowledge concerning these pathways would be of immense importance in many scientific and technological fields. For but one example, the pathway databases are useful in helping us to understand of how diseases are developed when a group of genes or proteins in one or more pathways are defective. Therefore, identification of the affected pathways in diseases would have helped the scientist to identify potential drug targets and gene therapy in diseases like cancers and neural disorders.

[0006] Construction of pathways and pathway databases is the subject matter of a few published inventions. For example, U.S. Patent Application 2002/0168664 discloses a method using clustering algorithms of analysing a group of genes identified through analysis of gene expression profiling experiments, wherein the groups of genes have been grouped together by a commonality in their gene expression patterns. U.S. Patent Application 2003/0100996 discloses of methods of construction and search of a protein pathway database by using FASTA format and dynamic programming method. And U.S. Patent Application 2003/0023388 discloses a method for

reconstructing pathways in large genetic networks from genetic perturbations by applying a recursive algorithm. However, while the disclosed inventions satisfy to certain extent the need of constructing individual pathways and pathway database containing the individual pathways, they fail to go beyond the construction and search of individual pathways.

[0007] Therefore, there is a need for systems and methods that are capable of analysing of interactions of genes or proteins at the pathway level. The present invention satisfies such a need in providing a system and methods of constructing a pathway network of a set of genes or proteins.

Summary of the Invention

[0008] The present invention provides systems and methods for constructing pathway networks of two or more pathways for a set of genes or proteins, wherein a single pathway contains one or more genes or proteins that are grouped based on one or more commonality of all the genes or proteins within the pathway, and wherein the set of genes or proteins are identified by differential gene expression profiling. More particularly the invention uses computer-implemented methods and systems to analyse gene expression profiling dataset by identifying a set of genes or proteins that are responsive in coordination to one or more effecters, constructing a secondary pathway database by mining primary databases, identifying the pathways that contains one or more genes or proteins of the set of genes or proteins, and constructing a pathway network of the identified pathways by using single source maximum pathways algorithm.

04 MAY 2006

[0009] According to one embodiment, the present invention provides a method for constructing a pathway network of two or more pathways for a set of genes or proteins by creating a secondary pathway database by extracting information from primary pathways database(s) and protein database(s); identifying the pathways that contain one or more genes or proteins of the set of genes or proteins by searching the secondary pathway database; and constructing the pathway network of the identified pathways related to the set of genes or proteins by applying a single source maximum pathways (SSMP) algorithm. In one aspect, the extracted information from the primary pathways database includes the type of the pathways, the description of the pathways, the EC numbers (EC_NO) of members of the pathways, the pathways which one pathway is linked to, and/or the pathways which links to the one pathway, while the extracted information from the protein databases includes Accession Number (AC), Description (DE), Gene Name (GN), Comments (CC), Reference Comment (RC), Reference Position (RP), Database Cross-Reference (DR), Keyword (KW) and Identification (ID). In other aspects, the constructed pathway networks will be visualized and compared to make common pathway networks.

[0010] In another embodiment, the present invention provides a secondary pathway database with two or more pathways, wherein said secondary pathway database contains the information including the type of the pathways, the description of the pathways, the EC numbers (EC_NO) of members of the pathways, the pathways which one pathway is linked to, the pathways which links to one pathway, and the description of members of the pathways; and wherein the description of members of the pathways includes Accession Number (AC), Description (DE), Gene Name (GN), Comments

04 MAY 2006

(CC), Reference Comment (RC), Reference Position (RP), Database Cross-Reference (DR), Keyword (KW) and Identification (ID).

[0011] In yet another embodiment, the present invention provides an algorithm for constructing an adjacent matrix for constructing a pathway network, and an algorithm for constructing a pathway network with single source maximum pathways (SSMP).

[0012] In further embodiments, the present invention provides a computer system comprising a processor and a memory coupled to said processor, wherein the memory contains one or more programs for enabling the processor to perform the method, and wherein the memory stores the secondary pathway databases and the constructed pathway networks.

Brief Description of the Drawings

[0013] Preferred embodiments according to the invention will be described with reference to the Figures, in which like reference numerals denote like elements.

[0014] FIG 1A shows an exemplary pathway map from KEGG Pathways Database.

[0015] FIG 1B shows part of the Sterol, Vitamin K, Vitamin E, Carotenoids Biosynthesis Pathway from KEGG Pathways Database wherein legends are added to describe the information extracted from this database. The legends are: ⟹ indicating an example of one pathway that the current pathway is linked to; ⟸ indicating an example of one pathway

04 MAY 2006

that links to the current pathway; and ⬆ indicating the EC number for the enzyme involved in the current pathway.

[0016] FIG 1C shows an exemplary SWISS-PROT file record of a human protein from SWISS-PROT.

[0017] FIG 1D illustrates one configuration of the secondary pathway database.

[0018] FIG 2A is a simplified flowchart showing the search process into the secondary pathway database performed by one embodiment of the present invention for obtaining the affected pathways.

[0019] FIG 2B depicts a pathway network with five affected pathways with edges indicating the interconnections between two pathways and a corresponding adjacent matrix.

[0020] FIG 2C shows the resultant pathway networks with all possible starting points in the pathway network depicted in FIG 2B by performing the SSMP search algorithm for identifying all possible pathway network.

[0021] FIG 3 shows the layout of one pathway network after the SSMP algorithm shows the distance between pathways.

[0022] FIG 4A illustrates one pathway network generated from Alon's Colon Cancer Dataset by performing the present invention.

[0023] FIG 4B shows the adjacent matrix for the affected pathways generated from Alon's Colon Cancer Dataset by performing the present invention.

[0024] FIG 4C illustrates one pathway network generated from Febbo's Prostate Cancer Dataset by performing the present invention.

[0025] FIG 4D shows part of the adjacent matrix for the affected pathways generated from Febbo's Prostate Cancer Dataset by performing the present invention.

[0026] FIG 4E shows one common pathway network shared between the colon cancer pathway network of FIG 4A and the prostate cancer pathway network of FIG 4B.

Detailed Description of the Invention

[0027] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs.

[0028] A gene refers to a segment of a genome, wherein the segment has a physical location in the genome, and is composed of exons, introns, a promoter, and untranslated regulatory regions and expresses at least one functional protein. In some situations, the gene may not contain introns, promoter, or untranslated regulatory regions. A protein as used herein

includes any polypeptide that is encoded by a gene, and any biochemical compounds that may be produced by one or more proteins.

[0029] By gene expression profiling is meant a process by which molecular techniques are used to measure and compare expression levels of certain nucleic acid sequences (e.g., mRNAs, genes, expressed sequence tags (ESTs)), or levels of certain gene products, such as amino acid sequences (protein or fragments of proteins), in a cell-derived sample in relation to the levels of the same nucleic acid sequences or gene products from a different sample, or from the same sample measured at a different time point. As discussed above, the gene expression profiles could represent the differential expressions in many paired situations such as normal-disease, disease 1-disease 2, developmental stage 1-developmental stage 2, type 1 tissue-type 2 tissue, species 1-species 2, genetic background 1-genetic background 2, etc. An effecter means any chemical, composition, physical stimulus, manipulation or procedure that can cause a change of the expressions of a gene or protein.

[0030] As used herein, a pathway is defined as a group of two or more genes or proteins in which individual genes or proteins can affect the activity of other genes or proteins directly or indirectly. The activity of genes or proteins as used herein includes many different things, such as DNA methylation, gene expression—on the mRNA or protein level--, alternative splicing, post-transcriptional regulation, post-translational modification like phophorylation, or enzymatic activities. An affected pathway refers to any pathway that contains one or more genes or proteins whose expression profile has been affected by any effecter.

[0031] A pathway may be represented by a plurality of nodes connected by edges, and each pathway must contain at least two nodes and at least one edge. A "node" is a member of a pathway, wherein the member could be a protein or non-protein molecule including, but not limited to, inorganics such as atoms, ions, cofactors, prosthetic groups and side chains; organic molecules such as carbohydrates, lipids, fats, steroids, cyclic nucleotides, and compounds such as glycoproteins, glycolipids, NADH; and small drug molecules and the like that may have repetitious linear units and carry out a particular function in particular cellular location. A "edge" is any interaction that connects two nodes. It may indicate uni- or bi-directional interactions and be described using words such as activate, associate, carboxylate, coactivate, deactivate, decarbodylate, deglycosylate, dehydroxylate, dephosphorylate, downregulate, generate, glycosylate, hydroxylate, inhibit, phosphorylate, promote, stimulate, upregulate, and the like.

[0032] A pathway network contains two or more pathways for a set of genes or proteins, wherein a single pathway contains one or more genes or proteins that are grouped together based on one or more commonality of all the genes or proteins within the pathway, and wherein the set of genes or proteins are identified by differential gene expression profiling. The pathway network can be represented in a similar way as a pathway is represented with some modifications. For example, the node represents a pathway in a pathway network and the edge represents the interactions of pathways in the pathway network.

[0033] Because the available technology to measure gene expression is associated with great amounts of experimental noise, the description of present invention will be limited to qualitative information on the affected

genes. Qualitative information on gene interactions lends itself ideally to a graph representation of pathway networks. A directed graph is a mathematical object consisting of nodes and directed edges. In a graph representation of a pathway network, the nodes of the graph correspond to pathways, and two pathways, say pathway 1 and pathway 2, are connected by a directed edge if gene/genes or protein/proteins in pathway 1 influences the activity of gene/genes or protein/proteins in pathway 2 directly.

[0034] Genomic technology has made feasible large-scale experiments that affect the activity of many genes in a single cell or organism. In the meantime, many techniques enable scientists to do genome scale gene expression profiling. For example, the microarray that was developed in 1995 is capable of measuring thousands of genes in one single chip. Its significance is best viewed in relation to the main thrust of molecular genetics, which uses genetic perturbations to reconstruct biochemical pathways involving painstaking gene-by-gene analysis.

[0035] The explosion of gene expression profiling data undoubtedly enhances our understanding of gene regulation and interactions of genes. Gene regulation is useful for both assaying drugs and as a source of new molecular targets, assuming the regulatory network controlling a given gene is well-understood. As such, changes in gene expression patterns can be used to assay drug efficacy throughout the drug discovery process.

[0036] One function of constructing pathway networks is to discover the existence and determine the association of multiple dependencies among biological or transcriptional data that lead to the identification of possible associated pathways. There are a number of potential motivations for

constructing algorithms, and computer software for network construction. For instance, diagnostics and the diagnostic industry can use these techniques and tolls for application in disease identification. Other potential applications include treatment selection and outcome prediction. In addition, the derived information can be further applied to aid in drug development and areas of feature selection for classification. Moreover, the pathway network will help in understanding of development, organogenesis and evolution.

[0037] An initial construction of a pathway network of two or more pathways for a set of genes or proteins in the present application depends in large part upon the availability of gene expression profiles of the set of genes or proteins and of primary pathway databases and protein databases. Gene expression profiles can be obtained from conventional techniques in biochemistry, molecular biology and molecular genetics, and mass techniques such as microarrys and genechips. All the techniques of gene expression profiling are known to one skilled in the art. This application will not spend more time on describing any of the techniques.

[0038] Currently, there are available of many pathways databases and vast amount of genes and proteins experimental data. It is predictable that there will be more pathways databases and more comprehensive pathways databases with more experimental data in the coming days. Selected below are exemplary pathways databases.

[0039] BRENDA is the main collection of enzyme functional data available to the scientific community. It is available free of charge for academic, non-profit users via the internet (www.brenda.uni-koeln.de).

[0040] The Main Metabolic Pathways for Internet (MMPI), by Pieter van Santen, is available at http://home.planet.nl/~pvsanten/mmp/main.htm. However, there are no libraries for searching on names of pathways, molecules or enzymes.

[0041] The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways is another useful database storing pathways information. The pathways database are constructed to store higher order functional information, which contains graphical representation of cellular processes, such as metabolism, membrane transport, signal transduction and cell cycle. FIG 1A shows a typical pathway map available from the KEGG web interface, which is daily updated and freely available at http://www.genome.ad.jp/kegg. It does provide search capabilities through its bioinformatics solution DBGET. DBGET is a simple database retrieval system for a diverse range of molecular biology databases available at the KEGG.

[0042] Other pathways databases are: International Union of Biochemistry & Molecular Biology (2001) Nicholson minimaps. http://www.tcd.ie/Biochemistry/IUBMB-Nicholson/; Michal, G. (Ed.) (2002) Roche Applied Science: Biochemical pathways chart. Boehringer Mannheim, http://biochem.boehringer-mannheim.com/prodinfo_fst.htm?/techserv/metmap.htm; Miller, K. J. (1998) Metabolic Pathways of Biochemistry. http://www.gwu.edu/~mpb/; and BioCarta - Charting Pathways of Life at www.biocarta.com/

[0043] While applicants are aware that gene expression differences may be quantitative and/or qualitative, for the sake of simplicity of illustration of

the present invention, the description of the present invention will be limited to qualitative information on pathways and genes or proteins. All the principles applied to qualitative features of pathways and genes or proteins are applicable to quantitative freatures.

[0044] Now, the creation of a secondary pathway database from one primary pathways database and one protein database will be described. It is evident that the secondary pathway database can be constructed from two or more pathways databases and protein databases depending upon the need and affordability. Here, KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolic pathways database is used to illustrate the construction of a secondary pathway database. The KEGG database was chosen because it provides the needed information online that facilitates the collection of necessary data and the construction of the secondary pathway database. An exemplary pathway map from the KEGG Pathway Database is shown in FIG 1A.

[0045] In order to construct the secondary pathway database, the following information is extracted from the KEGG Pathways Database: type of metabolic pathways, the enzymes' EC numbers (EC_NO) involved in the pathways, the pathways which the current pathways are linked to, and the pathways which link to the current pathways. The extracted information is stored in the secondary pathway database. Now refer to FIG 1D for the design of the secondary pathway database. As is illustrated by the Sterol, Vitamin K, Vitamin E, Carotenoids Biosynthesis Pathway that is one involved in the Lipid Metabolism under the KEGG Metabolic Pathways database. The table titled "metabolism" will store the information "Lipid Metabolism" in the field "Type". Similarly, the table entitled "pathway" will

store in the field "Type" the "Lipid Metabolism", in the field "Name" the "Sterol, Vitamin K, Vitamin E, Carotenoids Biosynthesis", in the field "Link" the http://www.genome.ad.jp/kegg/pathway/map/map00100.html which is the actual URL link to the MAP, and in the field "map_id" the "map00100". FIG 1B shows part of the Sterol, Vitamin K, Vitamin E, Carotenoids Biosynthesis Pathway from KEGG Pathways Database wherein legends are added to describe the information extracted from this database. The legends are: ⇒ indicating an example of one pathway that the current pathway is linked to; ⇐ indicating an example of one pathway that links to the current pathway; and ⇑ indicating the EC number for the enzyme involved in the current pathway. Still referring to FIG 1D, the table entitled "Con_Path_out" will store in the field "Name" the "Sterol, Vitamin K, Vitamin E, Carotenoids Biosynthesis" and in the field "con_db" the map_id of the pathways that the current pathway is linked to. The table entitled "Con_Path_in" will store in the field "Name" the "Sterol, Vitamin K, Vitamin E, Carotenoids Biosynthesis" and the field "con_db" as the map_id of the pathways that links to the current pathway. The table entitled "EC_Pro" will store in the field "Name" the "Sterol, Vitamin K, Vitamin E, Carotenoids Biosynthesis" and in the field "EC_NO" the EC number of the enzyme involved in the current pathway. The fields "ID", "ID1" and "ID3" in the tables "EC_Pro", "Con_Path_out" and "Con_Path_in" respectively, are just a unique ID assigned to each entry as running number to the entries.

[0046] To complete the secondary pathway database, information related to individual members of each pathway is to be integrated. The sources for obtaining such information can be any protein database and/or experimental data. Here, the information related to enzymes involved in all the pathways is extracted from the SWISS-PROT database as an example.

The SWISS-PROT file of the Homo sapiens non-redundant proteome set is available from http://www.ebi.ac.uk/proteome/HUMAN/download.html. An example of the fields and record in this file is shown in FIG 1C. In this file, each line begins with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order in which they appear in an entry, are shown in Table 1:

[0047] Table 1: Fields available in the SWISS-PROT File

ID - Identification.	RC - Reference comments.
AC - Accession number(s).	RX - Reference cross-references.
DT - Date.	RA - Reference authors.
DE - Description.	RL - Reference location.
GN - Gene name(s).	CC - Comments or notes.
OS - Organism species.	DR - Database cross-references.
OG - Organelle.	KW - Keywords.
OC - Organism classification.	FT - Feature table data.
RN - Reference number.	SQ - Sequence header.
RP - Reference position.	- (blanks) sequence data.
// - Termination line	

[0048] The information extracted from the SWISS-PROT database includes the Accession Number (AC), Description (DE), Gene Name (GN), Comments (CC), Reference Comment (RC) , Reference Position (RP), Database Cross-Reference (DR), Keywords (KW) and Identification (ID). The information is stored into the secondary pathway database that has already contained the information of pathways. FIG 1D as mentioned above shows one configuration of the relational database built in the secondary

pathway database. The relationship between the information of pathways and the information of each member of pathways is not immediately available from the diagram. In the present invention, the pathways are linked by the EC numbers provided in the DE fields and the one from the pathways. This link is only established during the search process as illustrated in FIG 2A. It is noted that a secondary pathway network can be updated accordingly with the update of primary databases.

[0049] Upon the completion of constructing a secondary pathway database, a pathway network of a set of genes or proteins can be constructed. The set of genes or proteins may be any combinations of genes or proteins. Preferably, the set of genes or proteins is identified by their commonalities. For example, the set of genes or proteins may be from an organism that is at a certain developmental stage, a type of tissue, a diseased tissue or organism or tissues or organisms that undergo stresses or other environmental stimuli. Their expression profiles can be procured by any technique known to one skilled in the art.

[0050] The chosen set of genes or proteins is then used to obtain the affected pathways. FIG 2A is a flowchart according to the present invention for searching the affected pathways in the secondary pathway database. First, at block 210 the system receives and queues the IDs (n) of the set of genes or proteins in order to search the IDs in the secondary pathway database. Table 2 shows the possible IDs to be used for the search. At block 220, the queued IDs are searched one by one in the DR field of the secondary pathway database to verify whether an imputed ID exists in the searched database. If the imputed ID is not found in the searched database, the system stops forward search and records the un-matched ID and moves to the next ID. If

the imputed ID is found in the DR field, the search continues to block 230. At block 230, the system searches the EC number for each ID in the DE field of the searched database. If the EC number for any ID is not available in the DE field of the searched database, the system stops forward search and records the un-matched ID and moves to the next ID. If the EC number for an imputed ID is found in the searched database, the search continues to block 240. At block 240, the system searches the EC number to verify whether this EC number for an imputed ID is available in the EC_NO field. If the EC number for the imputed ID is not available in the EC_NO field, the system stops forward search and records the un-matched ID and moves to the next ID. If the EC number for the imputed ID is available in the EC_NO field, the system records the ID and its corresponding pathway that is designated as affected pathway. If more than one pathways are affected by the same ID, they are all recorded for further construction of the pathway network. If the system finishes all the queued IDs, the search ends and all the affected pathways are stored. If one or more genes of a pathway are affected, the pathway will be regarded as one affected pathway. In certain cases, if too many pathways are identified as affected ones by requiring only one gene affected in one pathway, the number of the affected pathways can be reduced by requiring that one affected pathway have two or more affected genes.

[0051] Table 2 lists exemplary IDs useable for search in the Database Cross-Reference field provided by SWISS-PROT file.

DR AARHUS/GHENT-2DPAGE; 8006; IEF.	DR PDB; 3ADK; 16-APR-88.
DR DICTYDB; DD01047; MYOA.	DR PIR; A02768; R5EC7.
DR EMBL; X01704; GMNOD23.	DR PROSITE; PS00021; KRINGLE.

DR ECO2DBASE; G052.0; 6TH EDITION.	DR REBASE; BSURI; RELEASE 9410.
DR ECOGENE; EG10054; ARAC.	DR SGD; L0000008; AAR2.
DR FLYBASE; FBGN0000055; ADH.	DR STYGENE; SG10312; PROV.
DR GCRDB; GCR_0087; -.	DR SUBTISET; BG10774; OPPD.
DR HIV; K02013; NEF$BRU.	DR SWISS-2DPAGE; P10599; HUMAN.
DR HSSP; P00438; 1DOB.	DR TRANSFAC; T00141; -.
DR SETA; SC00018; ACT1.	DR WORMPEP; ZK637.7; CE00437.
DR MAIZEDB; 25342; -.	DR YEPD; 4270; -.
DR MIM; 249900; 11TH EDITION.	

[0052] While the pathway databases are currently used in helping the understanding of genes and their interactions, their uses have been confined to study the interactions between the members of individual pathways within the pathways. Applicants of the present invention recognized that many pathways might be affected by one effecter. It cannot be more true in situations when organisms go through different developmental stages or suffer from different diseases such as cancers. Applicants discovered that pathway networks are valuable tools with many applications and for many purposes.

[0053] The affected pathways obtained from the search of the secondary pathway database can be used to construct an adjacent matrix and/or pathway network. The order is not important even though the construction of an adjacent matrix will be described first.

[0054] An adjacent matrix is constructed from the nodes and edges representing the affected pathways and their links to other affected pathways. For illustration, in one experiment to study a disease A, 10 genes are identified to be differentially over-expressed in diseased tissue samples. When the genes are submitted into the search procedure, 5 pathways are hit. The 5 pathways are designated as affected pathways. For the affected pathways, their links are extracted from the Con_Path_out and Con_Path_in tables of the secondary pathway database. With these information, a sample possible graph for this result is shown in FIG 2B.

[0055] The adjacent matrix is created using the following procedure:

N = Total number of affected pathways;

For i = 0 to N –1

For j = 0 to N-1

w[i][j] = 0;

For i = 0 to N-1 {

w[i][j] = 1 if P_i links to P_j

w[j][i] = 1 if P_i linked from P_j

}

[0056] The adjacent matrix created implies that the graph is directed with the "weight" equals 1. The "weight" indicates distant in term of number of pathways from the original pathways. This "weight" can be scored based on the other characteristics between the pathways by means and methods known to one skilled in the art.

[0057] Graph theories use nodes and edges from graph theory to represent affected pathways and links between the affected pathways, and hence forming the pathway network. The pathway network is modeled as directed graph as each node represent an affected pathway, which is linked to another affected pathway (stored in the Con_Path_out table field), or linked from another affected pathways (Con_Path_in table field).

[0058] Single Source Shortest Path Algorithm

[0059] There are two popular algorithm used in graph theory to find the single source shortest path solution, namely the Dijkstra's Algorithm and Bellman-Ford Algorithm. For a better understanding of the present invention, the Djikstra's Algorithm will be briefly described as an example of similar algorithms that can be used or modified to apply to the present invention.

[0060] Djikstra's algorithm (named after its discover, E.W. Dijkstra) solves the problem of finding the shortest path from a point in a graph (the source) to a destination. One can find the shortest paths from a given source to all points in a graph at the same time, hence this problem is sometimes called the single-source shortest paths problem.

[0061] For a graph,
G = (V, E) where **V** is a set of vertices and **E** is a set of edges.

Dijkstra's algorithm keeps two sets of vertices:

S the set of vertices whose shortest paths from the source have already been determined, and

V-S the remaining vertices.

The other data structures needed are:

d array of best estimates of shortest path to each vertex

pi array of predecessors for each vertex

The basic mode of operation is:

- Initialize **d** and **pi**,
- Set **S** to empty,
- Set starting/source node,
- While there are still vertices in **V-S**,
    - Sort the vertices in **V-S** according to the current best estimate of their distance from the source,
    - Add **u**, the closest vertex in **V-S**, to **S**,
    - Relax all the vertices still in **V-S** connected to **u**

[0062] Relaxation

[0063] The relaxation process updates the costs of all the vertices, **v**, connected to a vertex, **u**, if we could improve the best estimate of the shortest path to **v** by including (**u, v**) in the path to **v**.

The relaxation procedure proceeds as follows:

Initialize (Graph g, Node s)

- for each vertex v in Vertices(g)
    - d[v] = infinity
    - pi[v] = nil
    - d[s] = 0;

This sets up the graph so that each node has no predecessor (**pi[v]** = nil) and the estimates of the cost (distance) of each node from the source (**d[v]**) are infinite, except for the source node itself (**d[s]** = 0).

[0064] The relaxation procedure checks whether the current best estimate of the shortest distance to **v** (**d[v]**) can be improved by going through **u** (i.e. by making **u** the predecessor of **v**):

relax(Node u, Node v, double w[][])

- if d[v] > d[u] + w[u,v] then
    - d[v] = d[u] + w[u,v]
    - pi[v] = u

[0065] The pseudo-code of the entire Dijkstra's algorithm is shown below:

Dijkstra(Graph g, Node s)

- Initialize(g, s)
- S = { 0 } /* initialize S to empty set */
- Q = Vertices(g) /* Put the vertices in a Queue */
- while not Empty(Q) /* while Q is not an empty set */
    - u = ExtractClosest(Q); /* Choose node u in Q with minimum distance */
    - AddNode(S, u); /* Add u to S */
    - for each vertex v in Adjacent(u) /* for each adjacent node */
        - relax(u, v, w), /* calculate the shortest distance to v */

04 MAY 2006

[0066] Single Source Maximum Pathways Algorithm

[0067] However, the problem to be solved in the present invention differs from the single source shortest path problem. What is needed is to identify the single source that lead to the maximum nodes being connected in order to form a network of pathways. An algorithm termed as Single Source Maximum Pathways is developed in the present invention. The Dijkstra's Algorithm is as the core of the search as it helps to link all connected pathways by looking at the "shortest" pathway distance from the single source. As mentioned earlier, the pathways distance is always equal to 1 for a neighboring pathway.

[0068] The Dijkstra algorithm is selected as it travels through all possible nodes from a single source and find the shortest path from the single source. The idea here is to imagine a "pathway distance" to represent the "physical" distance or cost that is used in the original form. It is chosen by heuristic that a pathway is granted a link as an affected pathway if there is one or more than one genes/proteins differentially over expressed in that pathway. There is also a need to know the total of nodes visited after the completion of the search; this requirement is important as looking for a single source that could lead to a maximum number of nodes being visited. The new algorithm will perform a search through all the nodes to identify a set of potential single source that could produce a maximum connected nodes in the entire graph generated from the affected pathways. When the algorithm is applied to the current context, it is used to find the single source pathway that could lead to maximum number of pathways and that forms the pathway network.

[0069] The procedure for creating the single source maximum pathways (SSMP) is shown below. The inner loop of this algorithm deploys Dijkstra's algorithm with modifications mentioned earlier, to identify maximum visited pathways.

```
GetSSMP (pathway network P)
{
   Visited[N]={};
   For each pathway p_i in the pathway network P
   {
      s = p_i              /* initialize starting/source pathways */
      Initialize( g, s )   /* initialize graph */
      S = { 0 }            /* initialize S to empty set */
      Q = Vertices( g )    /* Put the vertices in a Queue */
      while not Empty(Q)   /* while Q is not an empty set */
      {
         u = ExtractClosest( Q );   /* Choose node u in Q with
   minimum dist.*/
         AddNode( S, u );           /* Add u to S */
         for each vertex v in Adjacent( u )  /* for each adjacent node
   */
         If (total genes/proteins > 1 )
         {
            relax( u, v, w ),    /* calculate the shortest distance to v */
            update pathway visit status /* record visited pathway */
         }
      }
```

```
        Visited[i] = total of visited pathways in S /* keep total pathways
    visited */
        }
        return Max(Visited[N])  /* return source that links to maximum
nodes */
    }
```

P - connected pathways that form a network, presented as a directed weighted graph in graph theory

p_i - is the source pathway, represent the starting node in graph theory

S - is the set of all affected pathways in diseased tissue samples

N - number of pathways in P

D - set of distances to p_0

w - set of links/edges in P

V – visited pathways with a different starting pathway point

[0070] Once the single source pathway is identified, the pathway network is presented for visualization and analysis. FIG 2C shows the results of this procedure applied on graph shown in FIG 2B. The network which starts from P3 has the most nodes included, hence will be selected and returned by the algorithm.

[0071] In the case where more than one solution is possible, further analysis on each case is required. The further analysis includes checking the genes or proteins submitted, eliminating links between affected pathways that are not important, and reconstructing the network using the above algorithm.

The algorithm can be repeated several times until a stable and complete pathways network is generated.

[0072] Visualization & Layout Algorithm

[0073] The present invention visualizes the affected pathways depending on the number of pathways away from the source pathway. This idea shows how the other pathways are being affected, for example, when the genes or proteins in the source pathway are expressed higher than in normal ones. The layout algorithm takes into consideration of the "pathway distance" that the SSMP algorithm has generated from the pathway network. It starts with the single source node that has been identified from the SSMP algorithm. It will then layout the pathways that has one "pathway distance" from the first node. Once all the pathways with one "pathway distance" is displayed, the algorithm will proceed to the next increment of "pathway distance, i.e. two; the process will continue till all the affected pathways are displayed. FIG 3A shows the output layout after the layout algorithm. From FIG 3A, it is noticed that P1, P4 & P5 form a cyclic graph within the ensemble of affected pathways. P3 could be the triggering pathway on the entire system. P2, terms as the sink of the network, is the last affected pathway in the diseased tissue samples, with the assumption that P3 is the starting point, or the source.

[0074] Pathway Network Comparison

[0075] Two pathway networks can be compared to generate a common pathway network, and at the same time to identify the not common, or unique, pathways from the two pathway networks. This comparison has many

applications. For example, in the case of the same disease, it allows comparison at the pathway level of genes or proteins expression profile from different stages of the disease. It is therefore helping in understanding the development of the disease. In the case of two different diseases, the comparison allows extraction of a common pathway network between the diseases, which could show that the onset of the diseases could come from the same pathway group. Therefore, proteins or genes involved in the diseases can be identified through this method. At the same time, the pathways not common to the two diseases can be used as possible markers to differentiate the two diseases.

[0076] The comparison of the pathway network has been made easier with the formation of the first graph using the layout algorithm. It compares the existence of each pathway that is found in the networks provided, then a new layout is being presented.

P = { pathways from network 1 }
Q = { pathways from network 2 }
R = {}

For each node p_i in P {
    Compare with each node q_j in Q
        If element in p_i equals element in q_j, add p_i to R
}

[0077] Once the common pathways are identified, the same single source maximum pathways algorithm is applied onto the pathways set R. Subsequently pathway layout algorithm is applied onto the result obtained.

04 MAY 2006

[0078] Examples

[0079] The following examples are presented by way of illustration of the previously described invention and is not limiting to that description.

[0080] 1. Construction of a secondary pathway database

[0081] The secondary pathway database was constructed as described above by using the KEGG metabolic pathways database and SWISS-PROT protein database.

[0082] 2. Selecting of differentially expressed genes from colon cancer and prostate cancer and finding corresponding pathways for the selected genes

[0083] Two sets of genes from colon cancer and prostate cancer were used to implement and test the system and method of the present invention as above discussed.

[0084] The set of genes from colon cancer was obtained from a published paper. (Alon et al., 1999, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc Natl Acad Sci USA, 96: 6745-50) The paper reported that the gene expressions in 40 tumor and 22 normal colon tissue samples were analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. The expression data were normalized and clustered using the Cluster and TreeView software, both were

an integrated pair of programs for analyzing and visualizing the results of complex microarray experiments, written by Michael Eisen. (http://rana.lbl.gov/EisenSoftware.htm) (Eisen et al., 1998, Cluster Analysis and Display of Genome-Wide Expression Patterns, Proc Natl Acad Sci U S A 95: 14863-8)

[0085] After clustering, 1168 genes/ESTs were extracted that were differentially over-expressed in colon tumor tissues. 142 genes were identified with their IDs from the secondary pathway database. Then the IDs were used to extract the affected pathways from the above-constructed secondary pathway database by following the steps of the flowchart in FIG 2A as discussed above. With 142 genes with corresponding accession numbers, a total of 32 pathways were identified from the secondary pathway database, as shown in Table 3.

[0086] Table 3 shows the pathways that each of them contains more than one gene with over-expression in colon cancer tissues.

1 = Pyrimidine metabolism	17 = Alanine and aspartate metabolism
2 = Folate biosynthesis	18 = Cysteine metabolism
3 = One carbon pool by folate	19 = N-Glycans biosynthesis
4 = Pentose phosphate pathway	20 = Fructose and mannose metabolism
5 = Purine metabolism	21 = Butanoate metabolism
6 = Pyruvate metabolism	22 = Glutamate metabolism
7 = Nitrogen metabolism	23 = Oxidative phosphorylation
8 = Glycine, serine and threonine metabolism	24 = Inositol phosphate metabolism
9 = Citrate cycle (TCA cycle)	25 = Propanoate metabolism

10 = Lysine degradation	26 = Arginine and proline metabolism
11 = Aminosugars metabolism	27 = Fatty acid metabolism
12 = Sterol biosynthesis	28 = Tryptophan metabolism
13 = Methionine metabolism	29 = gamma-Hexachlorocyclohexane degradation
14 = Selenoamino acid metabolism	30 = Valine, leucine and isoleucine biosynthesis
15 = Glycolysis / Gluconeogenesis	31 = Glutathione metabolism
16 = Porphyrin and chlorophyll metabolism	32 = Tyrosine metabolism

[0087] The set of genes from prostate cancer was obtained from a published paper. (Singh et al., 2002, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell, 1: 203-9) The paper reported that 52 prostate tumors and 50 non-tumor prostate samples (referred to as normal) were tested using oligonucleotide microarrays containing probes similar to the one used in the above colon cancer case. The expression data were also normalized and clustered using the Cluster and TreeView software mentioned above. There were identified 850 genes/ESTs which were differentially over-expressed in the prostate tumors. 153 of the 850 genes were identified with their IDs following the same procedures as in the colon cancer case. Then the affected pathways were identified by following the same procedures as in the colon cancer case. With 153 genes with corresponding accession numbers, a total of 50 pathways were identified from the secondary pathway database, as shown in Table 4.

04 MAY 2006

[0088] Table 4 shows the pathways that each of them contains more than one gene with over-expression in colon cancer tissues.

1 = Purine metabolism	26 = Sphingoglycolipid metabolism
2 = Pyrimidine metabolism	27 = Ganglioside biosynthesis
3 = Sterol biosynthesis	28 = N-Glycans biosynthesis
4 = Arginine and proline metabolism	29 = Pantothenate and CoA biosynthesis
5 = Urea cycle and metabolism of amino groups	30 = Methionine metabolism
6 = beta-Alanine metabolism	31 = Selenoamino acid metabolism
7 = Glycine, serine and threonine metabolism	32 = Pyruvate metabolism
8 = Pentose phosphate pathway	33 = Fatty acid biosynthesis (path 2)
9 = Fructose and mannose metabolism	34 = Fatty acid metabolism
10 = Glycerolipid metabolism	35 = Bile acid biosynthesis
11 = Phospholipid degradation	36 = Valine, leucine and isoleucine degradation
12 = gamma-Hexachlorocyclohexane degradation	37 = Benzoate degradation via hydroxylation
13 = Oxidative phosphorylation	38 = Aminosugars metabolism
14 = Citrate cycle (TCA cycle)	39 = Alanine and aspartate metabolism
15 = Reductive carboxylate cycle (CO2 fixation)	40 = Lysine biosynthesis
16 = Glutathione metabolism	41 = Prostaglandin and leukotriene metabolism
17 = Pentose and glucuronate interconversions	42 = Glycolysis / Gluconeogenesis
18 = Starch and sucrose metabolism	43 = Lysine degradation

19 = Nucleotide sugars metabolism	44 = Inositol phosphate metabolism
20 = Tyrosine metabolism	45 = Propanoate metabolism
21 = Phenylalanine metabolism	46 = N-Glycan degradation
22 = Folate biosynthesis	47 = Fatty acid biosynthesis (path 1)
23 = Porphyrin and chlorophyll metabolism	48 = Cysteine metabolism
24 = C21-Steroid hormone metabolism	49 = Butanoate metabolism
25 = Androgen and estrogen metabolism	50 = Synthesis and degradation of ketone bodies

[0089] 3. Construction of the adjacent matrix of the affected pathways

[0090] Based on the procedure described above, an adjacent matrix that shows the links between affected pathways was created. FIG 4B shows the adjacent matrix for the affected pathways generated from Alon's Colon Cancer Dataset by performing the present invention. FIG 4D shows part of the adjacent matrix for the affected pathways generated from Febbo's Prostate Cancer Dataset by performing the present invention.

[0091] 4. Construction and visualization of the pathway networks

[0092] With the ensemble of pathways and the adjacent matrix, a pathway network was then formed using SSMP algorithm described above. The pathway network for colon cancer genes/ESTs is shown in FIG 4A. And the pathway network for prostate cancer genes/ESTs is shown in FIG 4B.

The visualization is done by visualization & layout algorithm described earlier.

[0093] 5. Identification of common pathways between the colon cancer and prostate cancer

[0094] When the comparison algorithm as disclosed above was applied to the two pathway networks, the common pathway network was identified and shown in FIG 4C. The 16 common affected pathways found in these two diseases are listed in Table 5.

[0095] Table 5 shows the common pathways between the colon cancer and prostate cancer

1. Oxidative phosphorylation	9. Alanine and aspartate metabolism
2. Purine metabolism	10. Citrate cycle (TCA cycle)
3. Pentose phosphate pathway	11. Sterol biosynthesis
4. Glycine, serine and threonine metabolism	12. Fructose and mannose metabolism
5. Pyrimidine metabolism	13. Propanoate metabolism
6. Pyruvate metabolism	14. Arginine and proline metabolism
7. Glycolysis / Gluconeogenesis	15. Lysine degradation
8. Porphyrin and chlorophyll metabolism	16. N-Glycans biosynthesis

[0096] 6. Experimental data substantiated the value of the pathway network

04 MAY 2006

[0097] As shown in FIG 4A and 4B, the Oxidative phosphorylation pathway was presented as the single source for the two pathway networks of colon cancer and prostate cancer. A further inquiry into the members of the Oxidative phosphorylation pathway revealed that COXJ_HUMAN (P14406) EC 1.9.3.1 and VAB2_HUMAN (P21281 Q14544 Q15859 Q96IR0) EC 3.6.3.14 were from colon tumor sample dataset, and VAG1_HUMAN(O75348) EC 3.6.3.14, CX41_HUMAN(P13073) EC 1.9.3.1 and DHSA_HUMAN(P31040 Q16395 Q9UMY5) EC 1.3.5.1 were from the prostate tumor sample dataset. The common genes/proteins in tumor tissues are two COX genes with E.C. 1.9.3.1: cyclooxygenase-2 or COX-2 and Cytochrome c oxidase polypeptide IV or the COX IV-1.

[0098] The importance of COX-2 and COX IV-1 in tumor development is highlighted by the following exemplary published papers. Aznar et al., 2003, ROCK and Nuclear Factor-{kappa}B-dependent Activation of Cyclooxygenase-2 by Rho GTPases: Effects on Tumor Growth and Therapeutic Consequences, Mol Biol Cell. 14:3041-3054; Davies et al., 2003, Correlation between Cyclooxygenase-2 Expression and Angiogenesis in Human Breast Cancer, Clin Cancer Res. :2651-6; Shun et al., 2003, Cyclooxygenase-2 expression correlates with nuclear p53 accumulation in gastric carcinoma, Hepatogastroenterology 50:988-92; and Chun et al, 2003, Curcumin inhibits phorbol ester-induced expression of cyclooxygenase-2 in mouse skin through suppression of extracellular signal-regulated kinase activity and NF-{kappa}B activation, Carcinogenesis 2003 Jul 4 (ePub).

[0099] This example shows that the constructed pathway networks have many applications and provide practical benefits. One example is to identify

a common pathway network between two pathway networks as shown herein. The common pathway network can also be constructed for more than two pathway networks. The common pathway networks may provide us insight into the common regulator(s)/initiator(s) for the occurrence and progression of multiple cancers. Another is to identify the potential single source pathway of the entire network, which in this case is the Oxidative phosphorylation. With the known source pathway, it is possible to subsequently trace back the potential genes or proteins that may initiate tumor development. The initiator(s) therefore presents a potential target to be used to detect, cure and prevent a particular disease. Other not common pathways enable scientists to distinguish different diseases by exploiting the different genes or proteins unique to particular diseases.

[0100] While the foregoing has presented descriptions of certain preferred embodiments of the present invention, it is to be understood that these descriptions are presented by way of example only and are not intended to limit the scope of the present invention. It is expected that others skilled in the art will perceive variations which, while differing from the foregoing, do not depart from the spirit and scope of the invention as herein described and claimed.

CLAIMS

What is claimed is:

1. A method for constructing a pathway network of a set of genes or proteins, wherein the pathway network contains two or more pathways with nodes representing the individual pathways and edges representing the connections between two pathways, said method comprising:

creating a secondary pathway database by extracting information from primary pathways database(s) and protein database(s);

identifying the pathways related to the set of genes or proteins by searching the secondary pathway database; and

constructing the pathway network of the identified pathways related to the set of genes or proteins.

2. The method of claim 1, wherein the extracted information from the primary pathways database includes the type of the pathways, the description of the pathways, the EC numbers (EC_NO) of members of the pathways, the pathways which one pathway is linked to, and/or the pathways which links to one pathway.

3. The method of any one of claims 1 and 2, wherein the extracted information from the protein databases includes Accession Number (AC), Description (DE), Gene Name (GN), Comments (CC), Reference Comment (RC), Reference Position (RP), Database Cross-Reference (DR), Keyword (KW) and Identification (ID).

4. The method of any one of claims 1-3, wherein the identification of the pathways comprises: submitting of the set of genes or proteins by their IDs; verifying the existence of the submitted IDs through the DR field of the secondary pathway database; extracting the EC number of the submitted genes or proteins primary in the secondary pathway database by their IDs through the DE field of the secondary pathway database; verifying the existence of the extracted EC numbers in the EC_NO field; and recording the pathways identified in the EC_NO field.

5. The method of any one of claims 1-4, wherein the construction of the pathway network comprises: constructing an adjacent matrix of the identified pathways related to the set of genes or proteins by executing an algorithm represented by the following pseudo-code:

N = Total number of identified pathways;
For i = 0 to N –1
    For j = 0 to N-1
        w[i][j] = 0;

For i = 0 to N-1 {
    w[i][j] = 1 if P_i links to P_j
    w[j][i] = 1 if P_i linked from P_j
}

wherein the i and j represent the nodes of the pathway network and P represents the edges of the pathway network.

6. The method of any one of claims 1-5, wherein the construction of the pathway network further comprises executing one or more times of a single

source maximum pathways (SSMP) algorithm represented by the following pseudo-code:

```
GetSSMP (pathway network P)
{
   Visited[N]={};
   For each pathway p_i in the pathway network P
   {
      s = p_i              /* initialize starting/source pathways */
      Initialize( g, s )   /* initialize graph */
      S = { 0 }            /* initialize S to empty set */
      Q = Vertices( g )    /* Put the vertices in a Queue */
      while not Empty(Q)   /* while Q is not an empty set */
      {
         u = ExtractClosest( Q );   /* Choose node u in Q with
   minimum dist.*/
         AddNode( S, u );           /* Add u to S */
         for each vertex v in Adjacent( u )  /* for each adjacent node
   */
         If (total genes/proteins > 1 )
         {
            relax( u, v, w ),    /* calculate the shortest distance to v */
            update pathway visit status /* record visited pathway */
         }
      }
      Visited[i] = total of visited pathways in S /* keep total pathways
visited */
   }
```

return Max(Visited[N]) /* return source that links to maximum nodes */

}

wherein P represents the connected pathways that form the network; p_i as the source pathway represents the starting node; S is the set of all identified pathways; N represents the number of the pathways in P; D represents set of distances to p_0; w represents set of links/edges in P; and V represents visited pathways with a different starting pathway point.

7. A method for constructing a pathway network of a set of genes or proteins, wherein the pathway network contains two or more pathways with nodes representing the individual pathways and edges representing the connections between two pathways, said method comprising:

creating a secondary pathway database by extracting information from primary pathways database(s) and protein database(s); wherein the extracted information including the type of the pathways, the description of the pathways, the EC numbers (EC_NO) of members of the pathways, the pathways which one pathway is linked to, the pathways which links to one pathway, and the description of members of the pathways; and wherein the description of members of the pathways includes Accession Number (AC), Description (DE), Gene Name (GN), Comments (CC), Reference Comment (RC), Reference Position (RP), Database Cross-Reference (DR), Keyword (KW) and Identification (ID);

identifying the pathways related to the set of genes or proteins by searching the secondary pathway database; wherein the identification comprises: submitting of the set of genes or proteins by their IDs; verifying the existence of the submitted IDs through the DR field of the secondary pathway database; extracting the EC number of the submitted genes or

proteins primary in the secondary pathway database by their IDs through the DE field of the secondary pathway database; verifying the existence of the extracted EC numbers in the EC_NO field; and recording the pathways identified in the EC_NO field;

constructing the pathway network of the set of genes or proteins; wherein the construction comprises: constructing an adjacent matrix of the identified pathways related to the set of genes or proteins by executing an algorithm represented by the following pseudo-code:

N = Total number of identified pathways;

For i = 0 to N –1

For j = 0 to N-1

w[i][j] = 0;

For i = 0 to N-1 {

w[i][j] = 1 if P_i links to P_j

w[j][i] = 1 if P_i linked from P_j

}

wherein the i and j represent the nodes of the pathway network and P represents the edges of the pathway network; and creating the pathway network with executing a single source maximum pathways (SSMP) algorithm represented by the following pseudo-code:

```
GetSSMP (pathway network P)
{
   Visited[N]={};
   For each pathway p_i in the pathway network P
   {
      s = p_i                 /* initialize starting/source pathways */
      Initialize( g, s )      /* initialize graph */
```

```
S = { 0 }              /* initialize S to empty set */
Q = Vertices( g )      /* Put the vertices in a Queue */
while not Empty(Q)     /* while Q is not an empty set */
{
   u = ExtractClosest( Q );      /* Choose node u in Q with
minimum dist.*/
   AddNode( S, u );              /* Add u to S */
   for each vertex v in Adjacent( u )   /* for each adjacent node
*/
   If (total genes/proteins > 1 )
   {
      relax( u, v, w ),      /* calculate the shortest distance to v */
      update pathway visit status /* record visited pathway */
   }
}
Visited[i] = total of visited pathways in S /* keep total pathways
visited */
}
return Max(Visited[N])  /* return source that links to maximum
nodes */
}
```

wherein P represents the connected pathways that form the network; p_i as the source pathway represents the starting node; S is the set of all identified pathways; N represents the number of the pathways in P; D represents set of distances to p_0; w represents set of links/edges in P; and V represents visited pathways with a different starting pathway point.

8. The method of any one of claims 1-7, wherein the set of genes or proteins are selected by their expression profiles.

9. The method of any one of claims 1-8, wherein the expression profiles represent the responses of the set of genes or proteins to one effecter.

10. The method of any one of claims 1-9, wherein the effecter includes developmental stages, physiological conditions, pathological conditions and genetic diversity.

11. The method of any one of claims 1-10, wherein the primary pathways databases and protein databases include the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the SWISS-PROT database.

12. The method of any one of claims 1-11, said method further comprising:

visualizing the constructed pathway network by selecting one pathway from the constructed pathway network and using the selected pathway as a starting point to layout the constructed pathway network.

13. The method of any one of claims 1-11, said method further comprising:

generating a common pathway network from two or more constructed pathway networks; wherein the common pathway network from two constructed pathway networks can be found by executing a comparison algorithm represented by the following pseudo-code:

P = { pathways from network 1 }

Q = { pathways from network 2 }

R ={}

For each node p_i in P {

Compare with each node q_j in Q

If element in p_i equals element in q_j, add p_i to R

}; and

wherein, optionally, the resultant common pathway network can be repeatedly compared with another pathway network to obtain further common pathway networks by repeatedly executing the comparison algorithm; and wherein unique pathways for each pathway network is also identified.

14. A secondary pathway database with two or more pathways, wherein said secondary pathway database contains the information including the type of the pathways, the description of the pathways, the EC numbers (EC_NO) of members of the pathways, the pathways which one pathway is linked to, the pathways which links to one pathway, and the description of members of the pathways; and wherein the description of members of the pathways includes Accession Number (AC), Description (DE), Gene Name (GN), Comments (CC), Reference Comment (RC), Reference Position (RP), Database Cross-Reference (DR), Keyword (KW) and Identification (ID).

15. The secondary pathway database of claim 14, wherein the information is extracted from primary pathways database(s) and protein database(s).

16. The secondary pathway database of claim 15, wherein the primary pathways databases and protein databases include the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the SWISS-PROT database.

04 MAY 2006

17. An article of manufacture, comprising:

at least one computer usable medium having computer readable program code means embodied therein for constructing an adjacent matrix for constructing a pathway network, the computer readable program code means in said article of manufacture comprising:

computer readable program code means for causing a computer to construct an adjacent matrix for constructing a pathway network, wherein the computer readable program code means is represented by the following pseudo-code:

N = Total number of identified pathways;

For i = 0 to N –1

For j = 0 to N-1

w[i][j] = 0;

For i = 0 to N-1 {

w[i][j] = 1 if P_i links to P_j

w[j][i] = 1 if P_i linked from P_j

}

wherein the i and j represent the nodes of the pathway network and P represents the edges of the pathway network.

18. An article of manufacture, comprising:

at least one computer usable medium having computer readable program code means embodied therein for constructing a pathway network with single source maximum pathways (SSMP), the computer readable program code means in said article of manufacture comprising:

computer readable program code means for causing a computer to construct a pathway network with single source maximum pathways (SSMP), wherein the computer readable program code means is represented by the following pseudo-code:

```
GetSSMP (pathway network P)
{
   Visited[N]={};
   For each pathway p_i in the pathway network P
   {
      s = p_i              /* initialize starting/source pathways */
      Initialize( g, s )   /* initialize graph */
      S = { 0 }            /* initialize S to empty set */
      Q = Vertices( g )    /* Put the vertices in a Queue */
      while not Empty(Q)   /* while Q is not an empty set */
      {
         u = ExtractClosest( Q );   /* Choose node u in Q with
   minimum dist.*/
         AddNode( S, u );           /* Add u to S */
         for each vertex v in Adjacent( u )  /* for each adjacent node
   */
         If (total genes/proteins > 1 )
         {
            relax( u, v, w ),    /* calculate the shortest distance to v */
            update pathway visit status /* record visited pathway */
         }
      }
      Visited[i] = total of visited pathways in S /* keep total pathways
   visited */
```

04 MAY 2006

```
        }
        return Max(Visited[N]) /* return source that links to maximum
nodes */
    }
```

wherein P represents the connected pathways that form the network; p_i as the source pathway represents the starting node; S is the set of all identified pathways; N represents the number of the pathways in P; D represents set of distances to p_0; w represents set of links/edges in P; and V represents visited pathways with a different starting pathway point.

19. A computer system comprising:

a processor; and

a memory coupled to said processor, wherein the memory contains one or more programs for enabling the processor to perform the method disclosed in any one of claims 1-13, and wherein the memory stores the secondary pathway databases of any one of claims 14-16 and the constructed pathway networks resulting from the performance of any one of claims 1-13.

20. The computer system of claim 19, wherein the programs include the adjacent matrix algorithm of claim 17 and the SSMP algorithm of claim 18.