

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
22 November 2007 (22.11.2007)

PCT

(10) International Publication Number
WO 2007/134167 A2

(51) International Patent Classification:
E21B 17/08 (2006.01)

(21) International Application Number:
PCT/US2007/068666

(22) International Filing Date: 10 May 2007 (10.05.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/799,271 10 May 2006 (10.05.2006) US
60/885,349 17 January 2007 (17.01.2007) US

(71) Applicant (for all designated States except US): **THE TRUSTEES OF COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK** [US/US]; 116th Street And Broadway, New York, NY 10027 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **ANASTASSIOU, Dimitris** [US/US]; 35 Depeyster Avenue, Tenafly, NJ 07670 (US). **VARADAN, Vinay** [IN/US]; 110W Morningside Drive, Apt.33, New York, NY 10027 (US).

(74) Agent: **RAGUSA, Paul, A.**; Baker Botts L.L.P., 30 Rockefeller Plaza, New York, NY 10112-4498 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

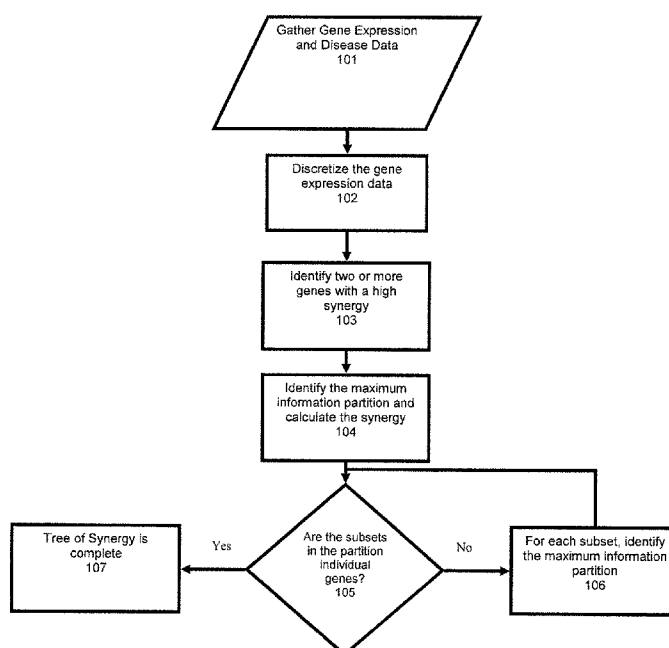
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: COMPUTATIONAL ANALYSIS OF THE SYNERGY AMONG MULTIPLE INTERACTING FACTORS



(57) Abstract: Systems and methods for selecting factors from a data set of measurements are provided. The measurements include values of factors and/or outcomes. Two or more factors that are jointly associated with one or more outcomes from the data set are identified. Each of the two or more factors are analyzed to determine at least one cooperative interaction among the factors with respect to an outcome. The two or more factors can be a module of factors.

WO 2007/134167 A2

COMPUTATIONAL ANALYSIS OF THE SYNERGY AMONG MULTIPLE INTERACTING FACTORS

PATENT APPLICATION SPECIFICATION

5 CROSS-REFERENCE TO RELATED APPLICATIONS

The present application claims priority to U.S. Provisional Application Serial No. 60/799,271 filed May 10, 2006 and U.S. Provisional Application Serial No. 60/885,349 filed January 17, 2007, the entire contents of which are incorporated by reference herein.

10 BACKGROUND

The disclosed subject matter relates generally to systems and methods for factor selection, including factors useful in gene expression analysis.

The expression levels of thousands of genes, measured simultaneously using DNA microarrays, can provide information useful for medical diagnosis and
15 prognosis. However, gene expression measurements have not provided significant insight into the development of therapeutic approaches. This can be partly attributed to the fact that while traditional gene selection techniques typically produce a “list of genes” that are correlated with disease, they do not reflect interrelationships.

Gene selection techniques based on microarray analysis often involve
20 individual gene ranking depending on a numerical score measuring the correlation of each gene with particular disease types. The expression levels of the highest-ranked genes tend to be either consistently higher in the presence of disease and lower in the absence of disease, or vice versa. Such genes usually have the property that their joint expression levels corresponding to diseased tissues and the joint expression levels
25 corresponding to healthy tissues can be cleanly separated into two distinct clusters. These techniques are therefore convenient for classification purposes between disease and health, or between different disease types. However, they do not identify cooperative relationships or the synergy among multiple interacting genes.

There is therefore a need for the ability to analyze genes in terms of the
30 cooperative, as opposed to independent, nature of their contributions towards a phenotype.

SUMMARY

The disclosed subject matter provides techniques for the analysis of the cooperative interactions or synergy among multiple interacting factors. The factors can be features, elements or outcomes that are cooperatively associated with one or more factors or outcomes by their joint presence or absence.

In some embodiments of the disclosed subject matter, methods for selecting factors from a data set of measurements are provided. The measurements can include values of factors and/or outcomes. Two or more factors that are jointly associated with one or more outcomes from the data set are identified. Each of the two or more factors are analyzed to determine at least one cooperative interaction among the factors with respect to an outcome or factor. The factors can be a module of factors or a sub-module of factors. The interaction can be a structure of interactions.

The factors can include two or more genes. The data set can include gene expression data including expression levels for each of the genes. The outcomes includes presence or absence of a disease, and the genes can be a module of genes, preferably a smallest cooperative module of genes with joint expression levels that can be used for a prediction of the presence of a disease with high accuracy.

In other embodiments of the disclosed subject matter, methods for selecting two or more genes from gene expression data are provided. The gene expression data can include expression levels for each of the two or more genes. The method includes providing gene expression data for two or more genes, discretizing the gene expression data, identifying the two or more genes with a high synergy, and identifying a cooperative interaction that connects the expression levels in the two or more genes with presence or absence of a disease. The gene expression data can be derived from at least one microarray of gene expression data. The genes can be a module of genes. The cooperative interaction can be modeled using a Boolean function, a parsimonious Boolean function, or the most parsimonious Boolean function. The high synergy can be a maximum synergy.

In other embodiments of the disclosed subject matter, systems for selecting two or more genes from gene expression data are provided. The gene expression data includes expression levels for each of the two or more genes. The system includes at least one processor, a computer readable medium coupled to the processor including instructions which when executed cause the processor to provide

gene expression data for the genes. The instructions also cause the processor to discretize the gene expression data, choose a single threshold for each of the two or more genes, identify the two or more genes with a high synergy, and identify a cooperative interaction that connects the expression levels in the two or more genes with presence of a disease. The gene expression data can be derived from a microarray of gene expression data. The two or more genes can be a module of genes. The high synergy can be a maximum synergy.

In other embodiments of the disclosed subject matter, systems for selecting factors from a data set of measurements are provided. Each measurement can include values of the factors and outcomes. The system includes at least one processor and a computer readable medium coupled to the at least one processor. The computer readable medium includes instructions which when executed cause the processor to identify two or more factors that are jointly associated with one or more outcomes or factors from the data, and, analyze each of the two or more factors to determine at least one cooperative interaction among the factors with respect to an outcome or factor. The factors can be a module of factors. The module of factors can include at least one sub-module of factors. The cooperative interaction can include a structure of interactions, such as a logic function.

The two or more factors can be two or more genes, the data can include gene expression data including expression levels, and the outcomes can be the presence or absence of a disease. The genes can be a module of genes. The module of genes can include at least one sub-module of genes. The module of genes can be a smallest module of genes with joint expression levels that can be used for a prediction of the presence of disease.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a diagram according to some embodiments of the disclosed subject matter.

Fig. 2 is a diagram according to some embodiments of the disclosed subject matter.

Fig. 3 is a diagram according to some embodiments of the disclosed subject matter.

Fig. 4 is a schematic diagram according to some embodiments of the disclosed subject matter.

DETAILED DESCRIPTION

According to some aspects of the disclosed subject matter, a method for selecting factors from a data set of measurements is provided. The method includes identifying factors cooperatively associated with an outcome from a data set for two or more factors, and analyzing each of the factors or modules of factors to determine cooperative or synergistic interactions among the factors or outcomes with respect to the outcome. The data set can be a set of measurements that includes values of the factors and the outcomes.

One application of the disclosed subject matter is where the factors are genes and the inference of the cooperative relationship or synergy among multiple interacting genes is desired. Although the following will describe disease data it can be more generally applicable to other data sets.

Other applicable data sets include other biological data such as how cells are influenced by stimuli jointly, financial data, internet traffic data, scheduling data for industries, marketing data, and manufacturing data, for example. Table 1, below, specifies other data sets relevant to various objectives, including factors and outcomes, to which the disclosed subject matter can also be applied:

TABLE 1

Objective	Factors	Outcomes
Disease pathway identification	Gene expression	Disease
Synaptic specificity factors	Gene expression	Neural synapses
Disease susceptibility of specific genotypes	Single Nucleotide Polymorphisms (SNPs)	Disease
Genotypic basis for gene expression profiles	SNPs	Gene expression
Gene regulation factors	Gene expression	Gene expression
Gene expression association with individual SNPs	Gene expression	SNP
Pharmacogenomics	SNPs	Drug resistance
Drug side-effect modeling	SNPs	Side effect of drug
Stocks/bonds/currency selling/buying price identification	Stocks/bonds/currency price time series data	Sell/Buy at given price
Macroeconomic models	Macroeconomic time series such as consumer index, housing market index, trade balance, etc	Federal interest rate increase/decrease

Referring to Figure 1, gene expression data for two or more genes is provided 101 in the form of, for example, a microarray of expression data. The gene expression data is then discretized 102. For example, the data can be binarized into
 5 two levels. Other levels of discretization, such as trinarization, can also be used.

Rather than independently binarizing each gene's expression level, which can be more appropriate for an individual gene ranking approach, single thresholds are used for the genes. This approach is consistent with the fact that finding global interrelationships among genes is desirable and that the microarray data
 10 have already been normalized across the tissues and genes. Therefore, a choice of high threshold will identify the genes that are "strongly" expressed, while a choice of a low threshold will identify the genes that are expressed even "weakly." Entropy Minimization and Boolean Parsimony ("EMBP") analyses can be performed across several thresholds and to determine the threshold levels that provide optimized
 15 performance, as disclosed in International Application No. PCT/US06/61749, the disclosure of which is incorporated herein by reference.

Following binarization, each gene can be assumed to be either expressed or not expressed in a particular tissue. It can also be assumed that there are two types of tissues, either healthy ones or tissues suffering from a particular disease.
 20 For example, Table 2, below shows results from hypothetical microarray measurements of three genes, G_1 , G_2 , and G_3 in both the presence and absence of a particular cancer C. N_0 and N_1 represent the presence and absence of cancer, respectively.

TABLE 2

G_1	G_2	G_3	N_0	N_1
0	0	0	2	6
0	0	1	0	24
0	1	0	2	7
0	1	1	0	26
1	0	0	1	7
1	0	1	0	25
1	1	0	76	1
1	1	1	19	4

The latter assumption can also be generalized to include more than two types of tissues, or modified to be used for classification among several types of cancer. According to one aspect of the disclosed subject matter, two or more genes with a high synergy are identified 103, as detailed below.

Cooperative or synergistic interactions between multiple genes can then be determined 104, as shown in Figure 1. Because systems biology is based on a holistic view of biological systems, synergy is important.

A gene set of n genes can have expression levels $G_1, G_2, \dots, G_n$ and a particular outcome C can be any phenotype, such as the presence of a particular disease or the differentiation of stem cells into a particular cell type when analyzing expression data of human tissues. The synergy $\text{Syn}(G_1, G_2, \dots, G_n; C)$ of the gene set with respect to the phenotype C is defined by equation (1):

$$I(G_1, G_2, \dots, G_n; C) - \max_{\substack{\text{all partitions} \\ \{S_i\} \text{ such that} \\ \bigcup_i S_i = \{G_1, \dots, G_n\} \\ \text{and } \bigcap_i S_i = \emptyset}} \sum_i I(S_i; C) \quad (1)$$

The partition of the gene set that is chosen in equation (1) above is the one that maximizes the sum of the amounts of mutual information connecting the subsets of that partition with the phenotype 104 as shown in Figure 1, and it is referred to as the “synergistic partition” of the gene set $\{G_1, G_2, \dots, G_n\}$ with respect to the phenotype C . The definition is consistent with the intuitive concept that synergy is the additional amount of contribution for a particular task provided by an integrated “whole” compared with what can best be achieved, after breaking the whole into “parts,” by the sum of the contributions of these parts. The above quantity can be divided by the entropy $H(C)$ from EMBP analysis, in which case the maximum possible thus normalized synergy will be +1.

For the special case of $n = 2$, the synergy is defined as shown in equation (2):

$$\text{Syn}(G_1, G_2; C) = I(G_1, G_2; C) - [I(G_1; C) + I(G_2; C)]. \quad (2)$$

Equation (2) is symmetric with respect to the three random variables and equal to the opposite of the mutual information $I(G_1; G_2; C)$ common to the three variables G_1, G_2, C . Contrary to the mutual information common to two variables, the

mutual information common to three variables is not necessarily a nonnegative quantity. This allows for a strictly positive synergy.

In one example, it can be assumed that each of the genes G_1 and G_2 is equally (50% of the time) expressed when $C = 1$ and $C = 0$. In that case, it would appear that the two genes are uncorrelated with the phenotype C , because $I(G_1; C) = I(G_2; C) = 0$, and the genes would not be found high up in any typical “gene ranking” computational method. However, C can be determined with absolute certainty from the joint state of the two genes, for example when $C = 1$ if $G_1 = G_2$, and $C = 0$ if $G_1 \neq G_2$, in which case $I(G_1, G_2; C) = 1$, and the synergy is positive and equal to +1. On the other hand, if $G_1 = G_2 = C$ then the synergy is negative and equal to -1. More generally, if $G_1 = G_2 = \dots = G_n = C$ (ultimate redundancy) then the multivariate synergy can become even more negative and equal to $-(n - 1)$.

Since $H(C | G_1, G_2, \dots, G_n) = H(C) - I(G_1, G_2, \dots, G_n; C)$, EMBP analysis naturally tends to find high-synergy results, although not necessarily the most synergistic ones.

In another example, as shown in Table 2, equation (3) can be used for $n = 3$ and is simplified by omitting the phenotypes.

$$\text{Syn}_{123} = I_{123} - \max(I_1 + I_{23}, I_2 + I_{13}, I_3 + I_{12}, I_1 + I_2 + I_3) \quad (3)$$

20

The values for mutual information between gene subsets and the phenotype can then be found in this example as follows: $I_{123} = 0.7963$; $I_{12} = 0.7152$; $I_{13} = 0.5094$; $I_{23} = 0.5163$; $I_1 = 0.3235$; $I_2 = 0.3335$; and $I_3 = 0.2782$.

Positive synergy implies some form of direct or indirect interaction of the genes, as a system. This definition of synergy also allows for insight into the structure of potential pathways by making iterative use of the “synergistic partition,” defined earlier, to generate a hierarchical decomposition of the gene set into smaller modules. In particular, consider a rooted and not necessarily binary tree with n leaves, each of which represents one of the genes. Each node of the tree represents a subset or sub-module of genes, which contains the genes represented by the leaves of the clade formed by the node. Therefore the root represents the whole gene set. The synergistic partition of the whole gene set, as defined above, can then be represented by the branching of the root, so that the nodes that are neighboring to the

root represent the gene subsets or sub-modules defined by the synergistic partition 105, 106 as shown in Figure 1. Some of these nodes can be leaves, representing a single gene. If they are not leaves, then they represent a subset of genes, which has its own synergistic partition, defined and evaluated as above, with respect to the

5 phenotype. This methodology can be repeated for all gene subsets, until the full tree is formed 107 as outlined in Figure 1. This is referred to as the tree of synergy of the gene set $\{G_1, G_2, \dots, G_n\}$ with respect to the phenotype C . Each intermediate node of the tree of synergy identifies a gene subset or sub-module with nonnegative synergy.

As shown in Figure 2, G_1 and G_2 can be a subset of genes or a module

10 or sub-module of genes that together have a nonnegative synergy 202 and a third independent gene G_3 with a negative synergy 201 between itself and the subset of G_1 and G_2 . In the example of Table 2, the synergy between G_1 and G_2 is +0.0582 and the synergy between the subset of G_1 and G_2 and G_3 is -0.1971.

The synergy, as defined above, refers to the combined cooperative

15 participation of *all* n genes. If, for example, the expression of one of these genes is independent of all the other genes including the phenotype, then the synergy of the n -gene set will be zero, even if the set contains synergistic subsets or sub-modules. Therefore, for a thorough synergistic analysis of a gene set, it can be desirable to also identify the most synergistic subsets of size $n - 1$, $n - 2$, ..., 2, which may not

20 necessarily appear in the tree of synergy. For $n = 3$, however, it can be proved that the most synergistic subset of size 2, if it has positive synergy, is defined by an existing clade of the full tree of synergy. In this way, the disclosed subject matter identifies a cooperative or synergistic interaction that connects the expression levels in two or more genes with the presence or absence of disease. The interaction can be

25 modeled using, for example, using a most parsimonious Boolean function as disclosed in International Application No. PCT/US06/61749.

For small gene set sizes n , such as 3 or 2, it is possible to identify all the synergistic sets of genes by doing a search of possible gene sets of size n . However, as the size increases, searches of all gene sets become computationally

30 expensive and therefore heuristic methods can be employed to find a collection of gene sets with positive synergy, as shown in Figure 3. Starting with the binarized gene expression data 301 an initial random gene set is first chosen 302 and the synergy of the gene set is calculated 304. If the synergy is greater than a predetermined threshold 305 the gene set is added to the list of gene sets with

synergy. The predetermined threshold could be zero, for example. The gene set is then modified using a heuristic algorithm such as simulated annealing 303 to identify a new gene set whose synergy is evaluated 303. If the synergy of the new gene set is positive, it is added to the list of existing gene sets 306. This process is repeated until
5 the desired number of gene sets is reached. The gene sets are then selected based on the statistical significance of their synergy values 307.

For small sizes of gene sets, synergistic analysis can be done with algorithms that list all the partitions of a particular set of genes. The total number of partitions of a set with n elements is given by the "Bell number." As the values of
10 n increase, however, the increased computational complexity makes the problems more complex and heuristic solutions can be used to do the analysis.

The techniques of the disclosed subject matter can be implemented by way of off-the-shelf software such as MATLAB, JAVA, C++, or other software. Machine language or other low level languages can also be utilized. Multiple
15 processors working in parallel can also be utilized.

As illustrated in the embodiment depicted in Figure 4, a system in accordance with the disclosed subject matter can include a processor or multiple processors 404 and a computer readable medium 401 coupled to the processor or processors 404. The computer readable medium includes data such as factors and
20 outcomes 402 and can also include programs for synergy analysis and/or EMBP analysis 403. The system leads to the identification of the synergy, synergies or cooperative interaction(s) among multiple interacting factors 405. Multiple processors 404 working in parallel can also be utilized.

The foregoing merely illustrates the principles of the disclosed subject matter. Various modifications and alterations to the described embodiments will be
25 apparent to those skilled in the art in view of the teachings herein. It will thus be appreciated that those skilled in the art will be able to devise numerous techniques which, although not explicitly described herein, embody the principles of the disclosed subject matter and are thus within the spirit and scope of the disclosed
30 subject matter.

WHAT IS CLAIMED IS:

1. A method for selecting factors from a data set of measurements, the measurements including values of the factors and outcomes, comprising:
 - identifying two or more factors that are jointly associated with one or more
 - 5 outcomes from the data set; and
 - analyzing each of the two or more factors to determine at least one cooperative interaction among the factors with respect to an outcome.
2. The method of claim 1, wherein the two or more factors comprise a module of factors.
- 10 3. The method of claim 1, wherein the two or more factors comprise a sub-module of factors.
4. The method of claim 1, wherein the at least one interaction comprises a structure of interactions.
5. The method of claim 1, wherein the two or more factors comprise two or more
- 15 genes, the data includes gene expression data comprising expression levels for each of the two or more genes, and the one or more outcomes includes presence or absence of a disease.
6. The method of claim 5, wherein the two or more genes comprise a module of genes.
- 20 7. The method of claim 6, wherein the module of genes comprise a smallest cooperative module of genes with joint expression levels that can be used for a prediction of the presence of disease.
8. A method for selecting two or more genes from gene expression data, comprising:
 - 25 providing gene expression data for two or more genes, the gene expression data comprising expression levels for each of the two or more genes;
 - discretizing the gene expression data;
 - identifying the two or more genes with a high synergy; and
 - identifying a cooperative interaction that connects the expression levels in the
 - 30 two or more genes with presence or absence of a disease.
9. The method of claim 8 wherein the gene expression data is derived from at least one microarray of gene expression data.
10. The method of claim 8 wherein the two or more genes comprise a module of genes.

11. The method of claim 8, wherein the cooperative interaction is modeled using a most parsimonious Boolean function.
12. The method of claim 8, wherein the high synergy comprises a maximum synergy.
- 5 13. A system for selecting two or more genes from gene expression data, comprising:
- at least one processor, and
 - a computer readable medium, coupled to the at least one processor, having stored thereon instructions which when executed cause the processor to:
- 10 provide gene expression data for the two or more genes, the gene expression data includes expression levels for each of the two or more genes;
- discretize the gene expression data;
 - choose a single threshold for each of the two or more genes;
 - identify the two or more genes with a high synergy; and
- 15 identify a cooperative interaction that connects the expression levels in the two or more genes with presence of a disease.
14. The system of claim 13 wherein the gene expression data is derived from a microarray of gene expression data.
15. The method of claim 13 wherein the two or more genes comprise a module of genes.
- 20 16. The method of claim 13, wherein the high synergy comprises a maximum synergy.
17. A system for selecting factors from a data set of measurements, each measurement comprising values of the factors and outcomes comprising:
- 25 at least one processor, and
- a computer readable medium coupled to the at least one processor, having stored thereon instructions which when executed cause the at least one processor to:
- 30 identify two or more factors that are jointly associated with one or more outcomes or factors from the data; and
- analyze each of the two or more factors to determine at least one cooperative interaction among the factors with respect to an outcome or factor.
18. The system of claim 17, wherein the two or more factors comprise a module of factors.

19. The system of claim 18, wherein the module of factors comprises at least one sub-module of factors.
20. The system of claim 18, wherein the at least one cooperative interaction comprises a structure of interactions.
- 5 21. The system of claim 20, wherein the at least one cooperative interaction comprises a logic function.
22. The system of claim 21, wherein the two or more factors comprise two or more genes, the data comprises gene expression data comprising expression levels for each of the two or more, and the one or more outcomes comprise presence or
- 10 absence of a disease.
23. The system of claim 22, wherein the two or more genes comprise a module of genes.
24. The system of claim 23, wherein the module of genes comprises at least one sub-module of genes.
- 15 25. The system of claim 23, wherein the module of genes comprises a smallest module of genes with joint expression levels that can be used for a prediction of the presence of disease with high accuracy.

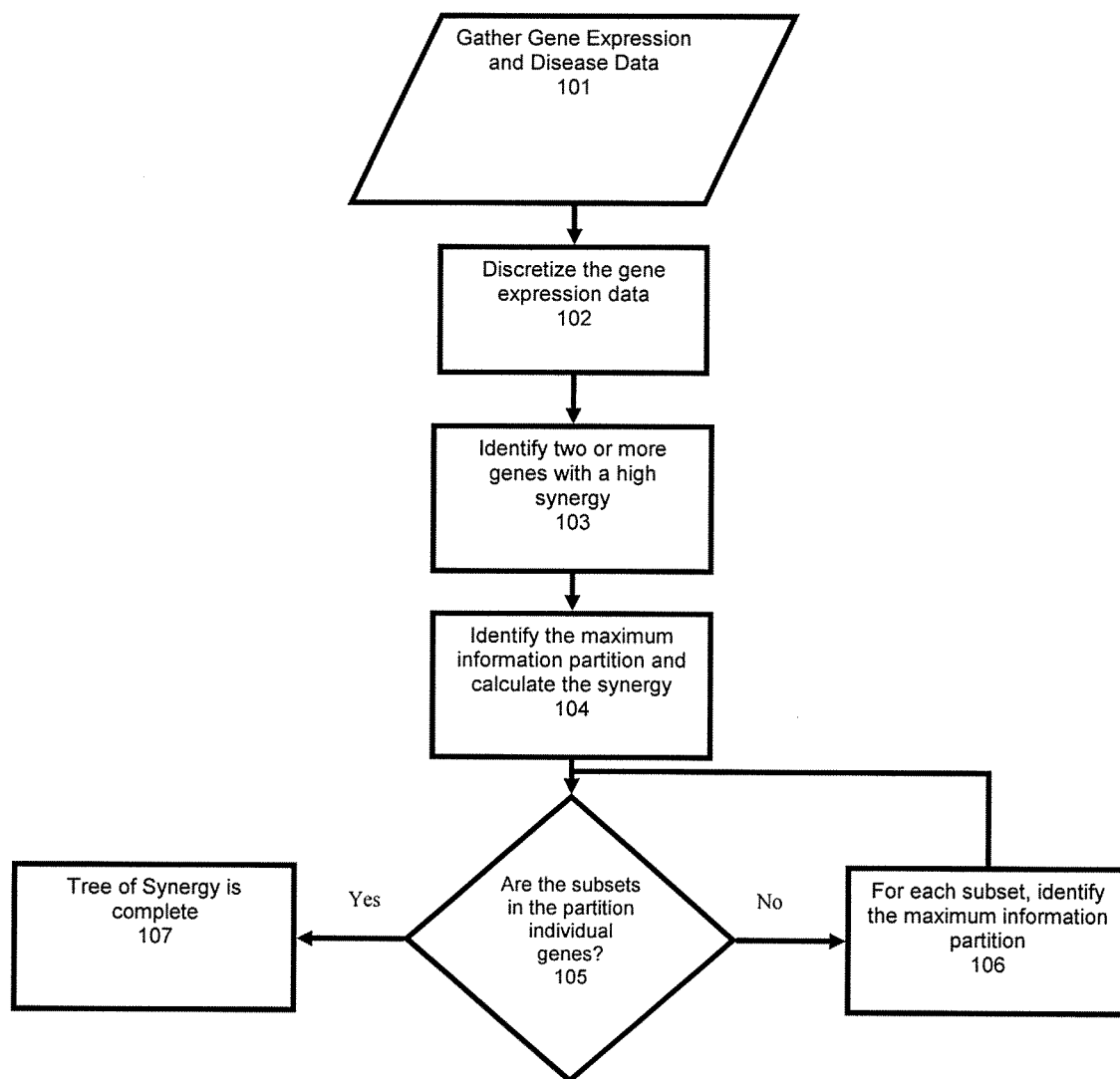
Figure 1

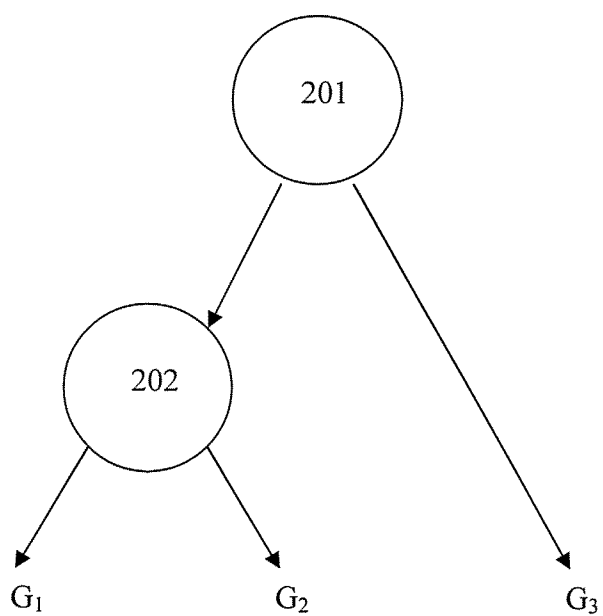
Figure 2

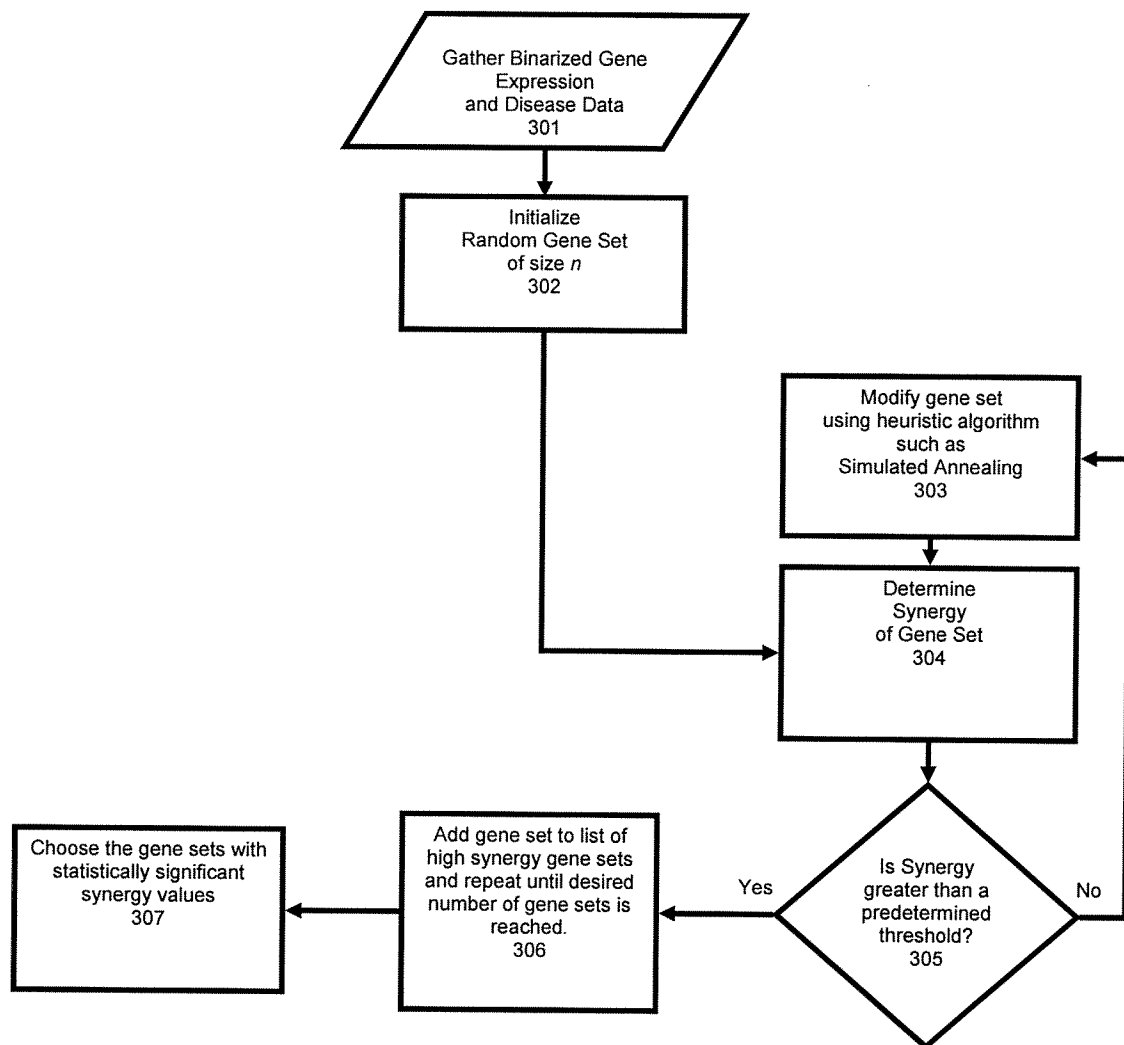
Figure 3

Figure 4