(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 July 2016 (28.07.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/118513 A1

(51) International Patent Classification:
G01N 33/50 (2006.01) *G06G 7/58* (2006.01)
G01N 33/84 (2006.01) *G06G 7/60* (2006.01)
G06F 19/12 (2011.01) *G06N 5/02* (2006.01)
G06F 19/18 (2011.01)

(21) International Application Number:
PCT/US2016/013935

(22) International Filing Date:
19 January 2016 (19.01.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/105,660 20 January 2015 (20.01.2015) US

(71) Applicants: **THE BROAD INSTITUTE, INC.** [US/US]; 415 Main Street, Cambridge, MA 02142 (US). **THE GENERAL HOSPITAL CORPORATION** [US/US]; 55 Fruit Street, Boston, MA 02114 (US).

(72) Inventors: **MERCER, Johnathan**; 18 Theurer Park, Watertown, MA 02472 (US). **HANSEN, Kasper, Lage**; 12 Walnut Street, Boston, MA 02118 (US).

(74) Agents: **SOLOMON, Mark, B.** et al.; Hamilton, Brook, Smith & Reynolds, P.C., 530 Virginia Rd, P.O.Box 9133, Concord, MA 01742-9133 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) **Title**: METHOD AND SYSTEM FOR ANALYZING BIOLOGICAL NETWORKS

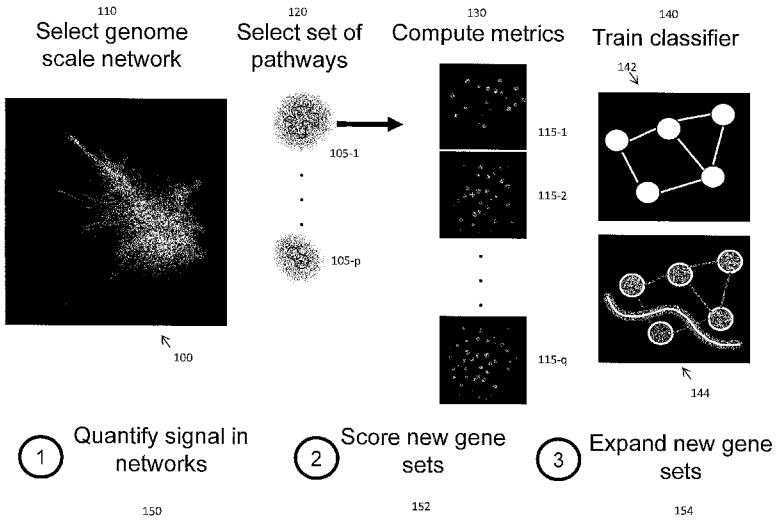


FIG. 1

(57) **Abstract**: A method and associated system are provided for analyzing biological networks. The method includes obtaining data representing biological networks from one or more data stores and obtaining data representing biological pathways, such as pathways defined for the biological networks. The biological networks are defined by respective nodes representing molecules and connections representing relationships between or among the molecules. Each pathway represents any set of molecules that work in a collaborative way to produce an outcome. The method generates, in one or more processors, a computational model (e.g., a classifier) based on the data representing the biological networks and the data representing the pathways. Further, a set of molecules within or related to a given biological network is classified into pathway molecules and non-pathway molecules using the generated model.

WO 2016/118513 A1

METHOD AND SYSTEM FOR ANALYZING BIOLOGICAL NETWORKS

RELATED APPLICATION

[0001] This application claims the benefit of U.S. Provisional Application No. 62/105,660, filed on January 20, 2015. The entire teaching of the above application is incorporated herein by reference.

BACKGROUND OF THE INVENTION

[0002] The recent explosion in genome-wide association studies, and exome-sequencing projects, have revealed many genes likely to be involved in disease processes. However, the organization and function of the circuits that contain the genes remain largely obscure. This limits progress towards biological understanding and therapeutic intervention.

SUMMARY OF THE INVENTION

[0003] Biological networks exhibit similar properties as social networks, such as scale-free, small-world, and modular architectures. Modeling pathway topologies using centrality measures, clustering, and community-based methods exploit these architectures and have driven much of the success of systems biology.

[0004] Embodiments of the present invention relate to a method and system for analyzing biological networks. The method and associated system are capable of learning systematic patterns that result both from evolution and experimental design to enable the identification of pathway members across diverse pathways and molecular relationships. In a particular embodiment, a web platform for the network-based analysis of genetic variants is provided. Embodiments of the present approach combine machine learning and cutting-edge web technologies to aid in the cognitive process of analyzing and interpreting large, experimentally noisy biological networks, thereby helping researchers to rapidly generate hypotheses that can be experimentally tested to elucidate the biology.

[0005] In one embodiment, the present invention provides a method for analyzing biological networks. The method includes obtaining data representing biological networks from one or more data stores and obtaining data representing biological pathways. The biological networks are defined by respective nodes representing molecules and connections representing relationships between or among the molecules. Each pathway represents any set

of molecules that work in a collaborative way to produce an outcome. The method generates, in one or more processors, a computational model (i.e., a classifier) based on the data representing the biological networks and the data representing the pathways. Further, a set of molecules within or related to a given biological network is classified into pathway molecules and non-pathway molecules using the generated computational model.

[0006] The biological pathways can be defined for the biological networks. The pathways, however, need not be derived based on the networks.

[0007] As used herein, a molecule means a whole molecule, such as a protein or nucleic acid molecule, or a unit thereof, such as a gene.

[0008] As used herein, a pathway gene means a gene influential in a given pathway.

[0009] As used herein, a context gene means a gene that is connected to a pathway gene of a given pathway but that is not known to be part of pathway with our current knowledge.

[0010] As used herein, a connectivity profile means the values of the set of topological metrics for a given pathway gene or member based on the pathway of interest. For example, there may be 18 topological metrics for a given pathway gene.

[0011] As used herein, a topological profile means a connectivity profile.

[0012] In an embodiment, the molecules are genes, the biological networks are genomic networks, and the set of molecules that the method classifies is a gene set. The method can further include assessing the classified genes to determine at least one of: significance of the gene set, structure of the gene set, components of the gene set, or relationship of the gene set to a known pathway. The classified genes can be assessed to determine if the gene set has topological characteristics of a pathway. The classified genes may be assessed to determine if the gene set is a variant of the known pathway or if the result of the classifying suggests a functional implication or association of the gene set to the known pathway. The method may further include predicting candidate genes for inclusion in the gene set.

[0013] Each pathway can include pathway genes and context genes. In an embodiment, each pathway gene is a known component of a molecular process and each context gene has a first order connection to a pathway gene. Generating the computational model can include systematically learning respective connectivity profiles across the pathways resulting, for example, in learned models ('trained models'). In an embodiment, each connectivity profile is a topological profile and learning the topological profile(s) can include evaluating connection characteristics of each pathway gene and each context gene, which can include

determining a node property of each pathway gene and each context gene. In an embodiment, a node property includes at least one of: number of connections, weighted number of connections (also referred to herein as "weighted degree"), eigenvector centrality, betweeness centrality, closeness centrality, and local clustering coefficient. The method may include determining these node properties and may also include determining other node properties. Each node property can be determined in a given pathway and in a respective network. Learning the topological profile(s) can further include calculating a ratio of each node property determined in the pathway to the respective node property determined in the network. In an embodiment, the method further includes grouping genes based on their connectivity. Grouping genes can include grouping the genes based on predicted probabilities from the generated model.

[0014] Generating the computational model can include building a model data set by stacking the learned connective profiles and, further, using machine learning techniques by performing a random forests analysis on the model data set.

[0015] The relationships between or among the genes can be based on at least one of physical interaction, similar global transcriptional response, co-dependencies in cancer cell lines, correlation in gene expression, or any other functional connection at the molecular or cellular level. A co-dependency in cancer or synthetic lethality relationship means that both genes of interest produce similar phenotypes across a set of cancer cell lines.

[0016] The method for analyzing biological networks can further include adjusting a respective pathway as a function of the classifying, in particular, as a function of the learning of connectivity profiles.

[0017] In an embodiment, obtaining the data representing biological networks includes allowing a user to select the biological network data. Further, obtaining the data representing the biological pathways can include allowing a user to input the pathway data.

[0018] In an embodiment, the generated computational model can be stored to a model store. Further, the set of molecules can be classified using the generated computational model or a computational model retrieved from the model store in response to a user selection.

[0019] A system for analyzing biological networks includes elements, such as a processor, one or more data stores, and a model store, configured to perform the method as described above.

[0020] A system for analyzing biological networks includes a network module configured to obtain data representing biological networks from one or more data stores, the biological networks being defined by respective nodes representing molecules and connections representing relationships between or among the molecules. The system further includes a pathway module configured to obtain data representing biological pathways (e.g., pathways defined for the biological networks), each pathway representing any set of molecules that work in a collaborative way to produce an outcome. One or more processors are configured to generate a computational model based on the data representing the biological networks and the data representing the pathways. A classifier module is configured to classify a set of molecules within or related to a given biological network into pathway molecules and non-pathway molecules using the generated model.

BRIEF DESCRIPTION OF THE DRAWINGS

[0021] The foregoing will be apparent from the following more particular description of example embodiments of the invention, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating embodiments of the present invention.

[0022] FIG. 1 is a schematic diagram illustrating a method and system for analyzing biological networks according to an example embodiment of the invention;

[0023] FIGs. 2A and 2B illustrate building a general classifier to predict pathway membership from networks according to an example embodiment of the invention;

[0024] FIG. 3 illustrates a known pathway including pathway genes, context genes and connections between and among pathway and context genes;

[0025] FIG. 4 is a schematic flow diagram illustrating a process of analyzing biological networks in accordance with example embodiments of the invention;

[0026] FIG. 5 is a schematic illustration of gene-by-gene learning of a topological profile;

[0027] FIG. 6 illustrates determining a node property relating to a degree, e.g., the number of connections, of a node;

[0028] FIG. 7 illustrates determining a node property relating to a weighted degree, e.g., the weighted number of connections, of a node;

[0029] FIG. 8A illustrates determining a node property relating to an eigenvector centrality;

[0030] FIG. 8B illustrates modeling of signaling networks;

[0031] FIG. 9 illustrates determining a node property relating to a betweeness centrality;

[0032] FIG. 10 illustrates determining a node property relating to a closeness centrality;

[0033] FIG. 11 illustrates determining a node property relating to a local clustering coefficient;

[0034] FIG. 12 illustrates grouping nodes representing genes based on connectivity;

[0035] FIG. 13 illustrates determining node properties in a given pathway, node properties in a respective network (overall), and calculating ratios of properties;

[0036] FIG. 14 illustrates an example gene set including pathway genes, context genes and connections (edges) between or among genes;

[0037] FIG. 15 shows an example table of genes and node properties for the gene set of FIG. 14;

[0038] FIG. 16 illustrates stacking of results (connectivity profiles) from many pathways to build a modeling data set;

[0039] FIG. 17A illustrates a machine learning technique that can be employed to generate a computational model;

[0040] FIG. 17B illustrates an example tree of a random forest analysis that can be used to generate a computational model;

[0041] FIG. 18 shows another example table of genes and node properties along with results of the classifying of the gene set of FIG. 14;

[0042] FIG.19 is a schematic view of a computer network environment in which embodiments of the present invention may be deployed;

[0043] FIG. 20 is a block diagram of computer nodes or devices in the computer network of FIG. 19;

[0044] FIG. 21 is a schematic diagram illustrating interaction among modules in an example system for analyzing biological networks;

[0045] FIG. 22A is a screenshot of an example user interface of the GENETS pathway analysis platform;

[0046] FIG. 22B illustrates example analysis and visualization tools available to a user of the GENETS pathway analysis platform;

[0047] FIG. 23 illustrates Quack probability distributions of pathway genes and context genes for the analysis of data from a set of 647 pathways;

[0048] FIG. 24 is a diagram illustrating pathway classes, including metabolic, signaling, regulatory and immune pathways;

[0049] FIG. 25A illustrates results of the analysis of average path length by network and pathway class;

[0050] FIG. 25B illustrates results of the analysis of clustering coefficient by network and pathway class;

[0051] FIG. 25C illustrates results of the analysis of modularity by network and pathway class;

[0052] FIG. 25D illustrates results of the analysis of density by network and pathway class;

[0053] FIG. 25E illustrates results of the analysis of context size by network and pathway class;

[0054] FIG. 26 illustrates classifier performance by network and pathway class;

[0055] FIG. 27 is a graph illustrating actual and predicted probabilities of candidates including change points;

[0056] FIGs. 28A-28D illustrate differential pathway topologies across functional genomics networks;

[0057] FIGs. 29A and 29B illustrate using network-specific topological rules to recapitulate pathway relationships;

[0058] FIG. 30 illustrates ranking the importance of pathway topological metrics across networks;

[0059] FIG. 31 shows plots of sensitivity versus specificity for classifiers trained on regulatory, signaling and metabolic pathways across networks.

DETAILED DESCRIPTION OF THE INVENTION

[0060] A description of example embodiments of the invention follows.

[0061] Network-based analyses are at the forefront of computational paradigms because of a need for a mathematical apparatus that can help map genotypes to phenotypes. Biology, including biological networks, can be complex and it has proven much more difficult to derive universal rules as compared to physics or mathematics. For example, a quantitative proteomic experiment using mass spectroscopy can yield data on 2,531 proteins and 14,657 potential interactions among the proteins. Currently, the number of nodes, e.g., genes, in a biological network can be in the range of about 12,000 to about 20,000 nodes, and the

number of connections, e.g., interactions, in the range of about 400,000 to about 2.5 million connections. In addition, the number of experimental and analytical options is increasing.

[0062] Different biological networks give different views of underlying biological processes. Biological networks can relate to protein-protein interactions, correlations in expression profiles, shared evolutionary history, synthetic lethality relationships in cancers, and cell perturbation profiles, among others. For example, InWeb+ relates to physical interaction among genes (based on a curated and quantitative process for assigning credibility to protein-interactions; this interaction network is supplemented with other biological interactions such as kinase-substrate interactions). The terms" InWeb+" and "InWeb" are used interchangeably herein. CMap/LINCS relates to cell perturabation profiles (i.e., for two genes A and B, over-expression of these induces a simlar set of genes to be expressed and the similar set of genes' expression to decrease). Achilles relates to co-dependency in cancer cells (i.e., similar synthetic lethality or "killing" of cancer cells across cancer cell lines). GEO relates to a standard repository for microarray experiments and relates to correlation in gene expression (i.e., the gene expression of gene A and B behave similarly across multiple samples). CLIME relates to similar evolutionary con/loss profiles (i.e., using phylogenic trees these genes conservation and loss profiles are similar). The relationships among a given set of genes may not look the same in all experiments and associated biological networks. Embodiments of the present approach allow data from different networks to be synthesized to give insights that were unobtainable from a single source.

[0063] Embodiments of method and systems described herein can address challenges of analyzing biological networks, such as inherent biological complexity, experimental noise and increase diversity and size of biological data.

[0064] Embodiments of the invention provide an analytic approach to compute on the above-described large networks to help build models of the pathway topology and filter experimental noise. Visualizations may be employed as cognitive aids to interpret and make sense of the remaining complexity and further elucidate the underlying biology. Embodiments of the present approach bring together the data, analytics, and visualization in an engaging and collaborative user experience where users can share data, insights, and analyses.

[0065] A biological pathway, as used herein, refers to set of molecules that work in a collaborative way to produce an outcome. Most genes, for example, exert their function in

concert with other genes in the same sub-networks that, from a cell biology perspective, often represent molecular machines, signaling circuits, enzymatic cascades, or rigid topological structures.

[0066] FIG.1 is a schematic diagram illustrating a method and system for analyzing biological networks in accordance with an example embodiment of the invention. At 110, a biological network 100, e.g., a genome scale network, is selected. For a given network, a set of pathways 105-1 to 105-p is selected, as illustrated at 120. For each pathway, one or more metrics 115-1 to 115-q are computed. Using the computed metrics, a classifier 142 is trained, as illustrated at 140. Using the classifier 142, the method and system can perform analyses on networks, pathways, and gene sets, as illustrated at 144. As illustrated and further described below, the method and system can be used to quantify signal in networks (150), score new gene sets (152), and expand new gene sets (154). Embodiments can rank the predictive power of the metrics across the several networks. Advantageously, embodiments can perform the analysis orders of magnitude faster than previous approaches.

[0067] FIGs. 2A and 2B schematically illustrate building a general classifier to predict pathway membership from networks according to an example embodiment of the invention.

[0068] As illustrated in FIG. 2A, for a given pathway, embodiments measure its topological properties (e.g., architectural properties) exemplified here with the 21 genes (210) of the AKT pathway 205 in the InWeb protein-protein interaction network. In this example, 18 topological properties are measured. In the resulting matrix 212, the 18 topological properties are shown as columns 214 and the corresponding values 216 for each of the 21 genes (210) in the AKT pathway (circles) as rows (metric values correspond to colors as indicated in the figure legend). One row in the matrix 212 corresponds to one row in the final modeling dataset. The same measurements are made for genes (220) in the context of the AKT pathway (squares). Only 2 of 2,449 context genes are shown in the illustration. This procedure is repeated for 853 pathways (205-1 to 205-k, where k=853) from which the modeling dataset used to train the classifier is derived, as illustrated at 230 in FIG. 2B. Resulting matrices 212-1 to 212-k are stacked to build the modeling dataset. At 240, a random forest classifier is trained for gene metrics 242 using a set of trees 244 (e.g., tree 1 to tree N). For any candidate gene in a network (e.g., a gene of a new gene set 205-x), the classifier can assign a probability 252 that the gene belongs to a pathway (e.g., the AKT

pathway) as defined by the candidates topological properties in the overall network (e.g., a row of matrix 212-x) and in relation to a specific set of genes (e.g., the 21 AKT genes).

[0069] FIG. 3 illustrates a known pathway 305, which is a network 300 imposed on a set of genes. In network 300, each node represents a gene and each edge represents a connection between two genes. Shown are pathway genes 310, context genes 320 and connections 315, 325 between or among pathway and context genes. Pathway 305, shown as an area encompassing genes in the network of genes 300, represents the current knowledge of implicated genes in the respective biological process or outcome of the process, e.g., phenotype, being investigated. Each pathway gene 310 is a known component of a molecular process. Each context gene 320 has a first order connection with a pathway gene, but is not, or not yet, implicated in the process or outcome being investigated. In general, a phenotype is an observable trait or collection of traits, like eye color, and is the result of gene expression and environment.

[0070] Information about a given pathway may be obtained from a database, such as the Molecular Signatures Database (MSigDB). One example of a pathway is any gene involved in WNT signaling. The present approach for analyzing biological networks can operate on data representing the pathway, including the pathway genes and context genes, to learn the topological profile(s) among and between the genes in the set, to allow the classification of pathway and context genes. Embodiments of the current approach may also identify new genes not previously implicated with the pathway.

[0071] FIG. 4 is a schematic flow diagram 400 illustrating a process and system of analyzing one or more biological networks in accordance with example embodiments of the invention. At block 410, data representing biological networks are obtained from one or more data stores 405. The biological networks are defined by respective nodes representing molecules and connections representing relationships between or among the molecules. As shown at 415, a user is allowed to select, and optionally to input, data representing the biological networks. For example, a user may specify a public database, such as MSigDB, as the source of the biological network data. Embodiments may also provide for the uploading of network data by users, the integration of networks (users can combine networks to make new bigger networks), and sharing of networks (shared networks which users can integrate into their own networks). At block 420, data representing biological pathways (e.g., pathways defined for the biological networks) are obtained. Optionally, pathway data are

obtained in response to, or as a result of, user input and/or user selection 425. For example, a user may upload the user's own pathway data or designate pathway data shared with another user. Each pathway represents any set of molecules that work in a collaborative way to produce an outcome. Next, at block 430, a computational model (e.g., a classifier) is generated based on the data representing the biological networks and the data representing the biological pathways. At 440, a set of molecules within or related to a given biological network are classified into pathway molecules and non-pathway molecules using the generated computational model. In an embodiment, the generated computational model can be stored to a model store 435. Further, the set of molecules can be classified using the generated computational model or a computational model retrieved from the store 435 in response to a user selection. In some embodiments, the molecules of the pathways are genes, the biological networks are genomic networks, and the set of molecules that is being classified is a gene set.

[0072] In block 430, generating the computational model can include learning respective connectivity profiles of the pathways resulting in learned models. In a particular example, each connectivity profile is a topological profile and the respective topological profiles are learned by evaluating connection characteristics of each pathway molecule (e.g., pathway gene) and each non-pathway molecule (e.g., context gene). This can include determining a node property of each pathway gene and each context gene (see also FIG. 5). Suitable node properties that may be employed in the learning of the topological profile can include a measure of the degree of connections (e.g., the number of first order connections, second order connections, etc.) of each node, a measure of the weighted degree of connections, an eigenvector centrality measure, a betweeness centrality measure, a closeness centrality measure, and local clustering coefficient. These node properties are further described below with reference to FIGs. 6-13 and in Examples 1 and 2. Other node properties may be determined in order to learn the connectivity profile. Each node property can be determined in a given pathway (see FIGs. 1-3 and Examples 1 and 2) and in a respective network (see FIG. 13 and Examples 1 and 2). Learning the topological profile can further include calculating, for each node property, a ratio of the node property as determined in the pathway to that determined in the network. Also, genes may be grouped based on their connectivity and the grouping(s) can be provided as an output to the user or used to learn the connectivity profile(s). Generating the computational model can include building a model data set by

stacking the connectivity profiles (see FIGs. 2B and 16) and, further, using machine learning techniques, for example, by performing a random forests analysis on the model data set (see FIGs. 2B and 17A).

[0073] Optionally, as illustrated at block 450 of FIG. 4, the classified genes may be assessed to determine at least one of significance of the gene set, structure of the gene set, components of the gene set, or relationship of the gene set to a known pathway. The classified genes can also be assessed to determine if the gene set is a variant of the known pathway or if the result of the classifying suggests a functional implication or association of the gene set to the known pathway. Optionally, at 460, the process may further include predicting candidate genes for inclusion in the gene set. As shown at 470, the process of analyzing biological networks can further include adjusting respective pathway(s) as a function of the classifying.

[0074] It should be readily appreciated by those of ordinary skill in the art that the aforementioned blocks are merely examples and that embodiments of the present invention are in no way limited to the number of blocks or the ordering of blocks described above. For example, some of the illustrated blocks may be performed in an order other than that which is described or include more or fewer blocks. Moreover, it should be understood that various modifications and changes may be made to one or more blocks without departing from the broader scope of embodiments of the present invention. It should also be appreciated that not all of the illustrated flow diagram is required to be performed, that additional flow diagram(s) may be added or substituted with other flow diagram(s).

[0075] FIG. 5 is a schematic diagram illustrating gene-by-gene learning of a topological profile. For each gene in the set of pathway genes 510 and the set of context genes 520, one or more node properties are determined. Shown here are the first order connections for each gene.

[0076] Embodiments learn the topological profiles of pathway and context genes by analyzing the characteristics of the pathway and context genes. One can use metrics from graph theory and social network analysis to establish the characteristics.

[0077] FIG. 6 illustrates determining a node property relating to a degree, e.g., the number of connections, of a node. As shown, node 610 has five first-order (direct) connections (611, 612, 613, 614, and 615) and node 620 has one first-order connection (621). Thus, in one example, the degree of node 610 is calculated to be 5 and the degree of node 620

is 1. Other measures of degree of connectivity can be used, including measures that take into account higher order connections, strength of respective connection, or combinations thereof. For example, a weighted degree measure is described with reference to FIG. 7. Additional information can be found in A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," PNAS 2004 101 (11) 3747-3752; published ahead of print March 8, 2004, doi:10.1073/pnas.0400087101. Connectivity measures as applied to molecules in biological networks are described in the article by Albert-László Barabási & Zoltán N. Oltvai, "Network biology: understanding the cell's functional organization," Nature Reviews Genetics 5, 101-113, February 2004, doi:10.1038/nrg1272.

[0078] FIG. 7 illustrates determining a node property relating to a weighted degree (weighted number of connections) of a node. This property can indicate how influential a molecule, e.g., a DNA molecule or a molecular unit, like a gene, is in a pathway. As illustrated, node 710 has many weak (low confidence) connections (711, 712, 713, 714 and 715) while node 720 has one strong (high confidence) connection (721). The strength of each connection, e.g., the level of confidence, can be associated with a numerical value. For connections 711 to 715 the respective weights are w1 to w5. For connection 721, the weight is w1'. For each node, respective weights can be summed to arrive at the weighted degree measure, i.e., *sum(w1, w2, w3, w4, w5)* and *sum(w1')*.

[0079] FIG. 8A illustrates determining a node property relating to an eigenvector centrality. This property can assess how influential a molecule is in a pathway, not just through direct interactions, but through second order interactions. A description of eigenvector centrality can be found in the article by Leo Spizzirri, "Justification and Application of Eigenvector Centrality," 2011, available online at https://www.math.washington.edu/~morrow/336_11/papers/leo.pdf; and the article by P.R. Gould, "On the Geographical Interpretation of Eigenvalues," Transactions of the Institute of British Geographers, No. 42 (Dec.) 1967, pp. 53-86. A mathematical expression from which a measure of eigenvector centrality can be derived is provided below.

[0080] $Ax = \lambda x$,

[0081] where A is the adjacency matrix and λ is the eigenvalue corresponding to the principal eigenvector.

[0082] FIG. 8B illustrates modeling of signaling networks. FIGs. 8A & 8B illustrate eigenvector centrality using both a network view (nodes and edges) 800 and a cartoon 830 (from: Susana R. Neves and Ravi Iyengar, "Modeling of signaling networks," Bioessays 2002). Node 810 is only connected to 815; however, 815 is an important biological component to propagate signals from two modules. Therefore, because 815 is important, this increases the importance of 810 even through it only has one connection to the network. Element 820 illustrates this signal propagation and relates it to the EGF induced MAP kinase activation in FIG. 8B.

[0083] FIG. 9 illustrates determining a node property relating to a betweeness centrality. This property can assess how signals in a network depend on a given node. Here, the shortest paths connecting respective nodes 915, 920, 925 and 930 to node 935 are all through node 910. Node 910 may be considered a "gate keeper" node. A mathematical expression for a measure of betweeness centrality, $C_B(i)$, is provided below.

[0084]
$$C_B(i) = \sum_{j<k} g_{jk}(i)/g_{jk} \quad ,$$

[0085] where g_{jk} is the number of geodesics connecting jk, and $g_{jk}(i)$ is the number of geodesics that actor i is on. Here, a graph geodesic is the shortest path between two nodes. A description of betweeness centrality and closeness centrality (see also FIG. 10 and below) can be found in Linton C. Freeman, "Centrality in Social Networks: Conceptual Clarification," Social Networks, 1 (1978/79) 215-239.

[0086] FIG. 10 illustrates determining a node property relating to a closeness centrality. This property can assess how influential a node is to communication amongst the rest of the pathway members. Here, node 1010 has only two first-order connection and several fourth- and one fifth-order connections among the shortest path connection to all the other nodes in the pathway. In contrast, node 1020 has five first-order connections and no fourth or fifth-order connections. The closeness centrally, which can be calculated as the inverse of minimum path length as shown in FIG. 10, is higher for node 1020 than for node 1010. A mathematical expression for a measure of closeness centrality, $C'_c(A)$, is provided below.

$$C_c'(A) = \left[\frac{\sum_{j=1}^{N} d(A,j)}{N-1} \right]^{-1},$$

[0087]

[0088] where $d(A,j)$ is the shortest path from node A to another node j in the pathway.

[Please confirm.]

[0089] FIG. 11 illustrates determining a node property relating to a local clustering coefficient, an indication of the 'embeddedness' of single nodes. This property can assess how connected the "neighbors" of a node are, suggesting a stronger functional relationship. In the example shown, node 1110 has a clustering coefficient (CC) of 1 while node 1120 has a clustering coefficient of 1/6. Both clustering coefficients are computed by taking the ratio of potential edges (connections) amongst the neighbors to the actual number of edges amongst the neighbors, of nodes 1110 and 1120 respectively. In inset 1130, the potential edges amongst the neighbors of node 1120 are shown as dashed lines and the actual edge as a solid line. Clustering coefficients have been described, for example, in an article by D. J. Watts and S.H. Strogatz, "Collective dynamics of 'small-world' networks," *393*(June) 1998, 440–442.

[0090] FIG. 12 illustrates a conventional method for grouping nodes representing genes based on connectivity. Shown are three communities of nodes 1210, 1220 and 1230. Each community is a group of nodes more connected to one another than they are to other groups of nodes. These nodes are labeled to be a part of the respective communities. This method is defined in Clauset et al. 2004 "Finding community structure in very large networks." A mathematical expression for a measure of community is provided below:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w).$$

[0091]

[0092] In the above, Q denotes the modularity, A_{vw} is an element of the adjacency matrix of the network, k_v and k_w are the respective degrees of the vertices (nodes) v and w, m is the number of edges in the graph, and the delta function $\delta(i, j)$ is 1 if $i=j$ and 0 otherwise.

[0093] As illustrated in FIG. 13, node properties may be determined in a given pathway (1310), to yield pathway properties, and may be determined in a respective network, e.g., to

yield overall properties (132). Furthermore, pathway and overall properties may be combined to create ratios of properties (1330).

[0094] FIG. 14 illustrates an example gene set including pathway genes 1410, context genes 1420 and connections (edges) 1415, 1425 between or among the genes. The example gene set is based on protein-protein interaction (PPI) data for the BIOCARTA RACCYCD pathway related to the cell cycle transition from G1 to S phase. The PPI data is obtained from MSigDB. As illustrated, there are 26 pathway genes, 2100 context genes, and 163,000 edges (connections). A diagram illustrating the cell cycle phases is provided at 1450, with the G1 to S transition shown at 1455. G1 to S(synthesis) transition is a "point of no return" for cell division whereby a cell goes from making mRNA and proteins to actually synthesizing new DNA for DNA replication. These genes in the gene set are all instrumental for the coordination of this process and the PPI shown are the interactions amongst these genes.

[0095] FIG. 15 shows a table of genes (nodes) 1535 and node properties (topological properties) 1545 for the gene set of FIG. 14. This table shows how node properties are structured to build the model. The order in which the properties are listed in the table (i.e., the order of columns) is not important. Genes 1535 are identified by their short name (e.g., "AKT1") and grouped by class 1530. Here, there are two classes: class 1510 ("Proteins Influential in G1 to S Transition") and class 1520 ("Context"). For genes of class 1510, which are considered pathway genes, the "In Pathway" column 1540 value is set to 1. For genes of class 1520, which are considered context genes, the value is set to 0. The values for each gene in the table of FIG. 15 may also be represented in matrix form, such as illustrated with matrix 212 in FIG. 2A.

[0096] FIG. 16 schematically illustrates stacking of results (connectivity profiles) obtained from many pathways to build a modeling dataset. The information contained across pathways can then be pooled to help build a general model across pathways. For simplicity, results from three pathways are illustrated having respective connectivity profiles for pathway genes 1610, 1611, 1612 and connectitiy profiles for context genes 1620, 1621, 1622. It will be understood that the number of pathways that can used in the analysis may be large and typically is only limited by the number of pathway data sets that are available.

[0097] FIG. 17A illustrates using machine learning techniques by performing a random forests analysis on the model data set to build the computational model. In one example, the

supervised machine learning ensemble method 1700 called random forests is used to build T (e.g., T=500) sub-optimal rule based models (e.g., trees) 1705-1 to 1705-T which all vote on the correct answer. A 70/30 split was used, i.e., the model was trained on 70% of the data and the remaining 30% were use to assess the classification power of the model. Using the generated computational model, a set of molecules (e.g., genes) within or related to a given biological network can be classified into pathway molecules (e.g., pathway genes) 1710 and non-pathway molecules (non-pathway genes, e.g., context genes) 1720. To this end, the computational model can output a probability P that a given molecule (e.g., gene) is a pathway molecule (e.g., pathway gene), as illustrated at 1760. The random forest analysis is advantageous because it is fast, shows enhanced predictive power over other conventional methods such as linear models, and is capable of identifying complex and non-linear relationships. Random forest analysis is described, for example, in the article by Leo Breiman, "Random Forests," Machine Learning 45 (1): 5–32, 2001, doi:10.1023/A:1010933404324.

[0098] FIG. 17B shows an example tree 1755 of a random forest analysis using structural metrics, such as weighted degree, degree in path, etc., as described above. The figure illustrates that trees can get complex and that these trees can represent copmlex and non-linear topological rules underlying pathway formation.

[0099] FIG. 18 shows a table of genes 1535 and node properties 1545 along with results 1850 of the classifying of the gene set of FIG. 14. The probabilities 1850 ("QuackP") can be used to segment potential pathway members in the original set, and predict new members. Here, the "ATK1" gene has a probability 0.96, as shown at 1855, which confirms the original classification of that gene as a pathway gene. In contrast, the "PAK1" gene, which is originally considered to be a pathway gene, only has a probability of 0.22, as shown at 1860. Conversely, the "CUL1" gene, originally considered a context gene, has a probability of 0.79, as shown at 1865.

[00100] Portions of the above-described embodiments of the present invention can be implemented using one or more computer systems, for example, to obtain data representing biological networks and data representing pathways, to generate a computational model, and to classify a set of molecules, e.g., genes, using the computational model. For example, the embodiments may be implemented using hardware, software or a combination thereof. When implemented in software, the software code can be stored on any form of non-transient

computer-readable medium and loaded and executed on any suitable processor or collection of processors, whether provided in a single computer or distributed among multiple computers.

[00101] Further, it should be appreciated that a computer may be embodied in any of a number of forms, such as a rack-mounted computer, desktop computer, laptop computer, or tablet computer. Additionally, a computer may be embedded in a device not generally regarded as a computer but with suitable processing capabilities, including a Personal Digital Assistant (PDA), a smart phone or any other suitable portable or fixed electronic device.

[00102] Also, a computer may have one or more input and output devices. These devices can be used, among other things, to present a user interface. Examples of output devices that can be used to provide a user interface include printers or display screens for visual presentation of output and speakers or other sound generating devices for audible presentation of output. Examples of input devices that can be used for a user interface include keyboards, and pointing devices, such as mice, touch pads, and digitizing tablets. As another example, a computer may receive input information through speech recognition or in other audible format.

[00103] Such computers may be interconnected by one or more networks in any suitable form, including as a local area network or a wide area network, such as an enterprise network or the Internet. Such networks may be based on any suitable technology and may operate according to any suitable protocol and may include wireless networks, wired networks or fiber optic networks.

[00104] Also, the various methods or processes outlined herein may be coded as software that is executable on one or more processors that employ any one of a variety of operating systems or platforms. Additionally, such software may be written using any of a number of suitable programming languages and/or programming or scripting tools, and also may be compiled as executable machine language code or intermediate code that is executed on a framework or virtual machine.

[00105] In this respect, at least a portion of the invention may be embodied as a computer readable medium (or multiple computer readable media) (e.g., a computer memory, one or more floppy discs, compact discs, optical discs, magnetic tapes, flash memories, circuit configurations in Field Programmable Gate Arrays or other semiconductor devices, or other tangible computer storage medium) encoded with one or more programs that, when executed

on one or more computers or other processors, perform methods that implement the various embodiments of the invention discussed above. The computer readable medium or media can be transportable, such that the program or programs stored thereon can be loaded onto one or more different computers or other processors to implement various aspects of the present invention as discussed above.

[00106] In this respect, it should be appreciated that one implementation of the above-described embodiments comprises at least one computer-readable medium encoded with a computer program (e.g., a plurality of instructions), which, when executed on a processor, performs some or all of the above-described functions of these embodiments. As used herein, the term "computer-readable medium" encompasses only a non-transient computer-readable medium that can be considered to be a machine or a manufacture (i.e., article of manufacture). A computer-readable medium may be, for example, a tangible medium on which computer-readable information may be encoded or stored, a storage medium on which computer-readable information may be encoded or stored, and/or a non-transitory medium on which computer-readable information may be encoded or stored. Other non-exhaustive examples of non-transitory computer-readable media include a computer memory (e.g., a ROM, RAM, flash memory, or other type of computer memory), magnetic disc or tape, optical disc, and/or other types of computer-readable media that can be considered to be a machine or a manufacture.

[00107] The terms "program" or "software" are used herein in a generic sense to refer to any type of computer code or set of computer-executable instructions that can be employed to program a computer or other processor to implement various aspects of the present invention as discussed above. Additionally, it should be appreciated that according to one aspect of this embodiment, one or more computer programs that when executed perform methods of the present invention need not reside on a single computer or processor, but may be distributed in a modular fashion amongst a number of different computers or processors to implement various aspects of the present invention.

[00108] Computer-executable instructions may be in many forms, such as program modules, executed by one or more computers or other devices. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. Typically, the functionality of the program modules may be combined or distributed as desired in various embodiments.

[00109] FIG. 19 illustrates a computer network or similar digital processing environment in which embodiments of the present invention may be implemented. Client computer(s)/devices 50 and server computer(s) 60 provide processing, storage, and input/output devices executing application programs and the like. Client computer(s)/devices 50 can also be linked through communications network 70 to other computing devices, including other client devices/processes 50 and server computer(s) 60. Communications network 70 can be part of a remote access network, a global network (e.g., the Internet), a worldwide collection of computers, Local area or Wide area networks, and gateways that currently use respective protocols (TCP/IP, Bluetooth, etc.) to communicate with one another. Other electronic device/computer network architectures are suitable.

[00110] FIG. 20 is a diagram of the internal structure of a computer (e.g., client processor/device 50 or server computers 60) in the computer network of FIG. 19. Each computer 50, 60 contains system bus 79, where a bus is a set of hardware lines used for data transfer among the components of a computer or processing system. Bus 79 is essentially a shared conduit that connects different elements of a computer system (e.g., processor, disk storage, memory, input/output ports, network ports, etc.) that enables the transfer of information between the elements. Attached to system bus 79 is I/O device interface 82 for connecting various input and output devices (e.g., keyboard, mouse, displays, printers, speakers, etc.) to the computer 50, 60. Network interface 86 allows the computer to connect to various other devices attached to a network (e.g., network 70 of FIG. 19). Memory 90 provides volatile storage for computer software instructions 92 and data 94 used to implement an embodiment of the present invention (e.g., generating a computational model based on data representing biological networks and data representing pathways, classifying a set of molecules within or related to a given biological network into pathway molecules and non-pathway molecules using the generated computational model, as detailed above and below, and further described in the Examples). Disk storage 95 provides nonvolatile storage for computer software instructions 92 and data 94 used to implement an embodiment of the present invention. Central processor unit 84 is also attached to system bus 79 and provides for the execution of computer instructions.

[00111] In one embodiment, the processor routines 92 and data 94 are a computer program product (generally referenced 92), including a non-transitory computer readable medium (e.g., a removable storage medium such as one or more DVD-ROM's, CD-ROM's, diskettes,

tapes, etc.) that provides at least a portion of the software instructions for the invention system. Computer program product 92 can be installed by any suitable software installation procedure, as is well known in the art. In another embodiment, at least a portion of the software instructions may also be downloaded over a cable, communication and/or wireless connection. In other embodiments, the invention programs are a computer program propagated signal product 107 embodied on a propagated signal on a propagation medium (e.g., a radio wave, an infrared wave, a laser wave, a sound wave, or an electrical wave propagated over a global network such as the Internet, or other network(s)). Such carrier medium or signals provide at least a portion of the software instructions for the present invention routines/program 92.

[00112] In alternative embodiments, the propagated signal is an analog carrier wave or digital signal carried on the propagated medium. For example, the propagated signal may be a digitized signal propagated over a global network (e.g., the Internet), a telecommunications network, or other network. In one embodiment, the propagated signal is a signal that is transmitted over the propagation medium over a period of time, such as the instructions for a software application sent in packets over a network over a period of milliseconds, seconds, minutes, or longer. In another embodiment, the computer readable medium of computer program product 92 is a propagation medium that the computer system 50 may receive and read, such as by receiving the propagation medium and identifying a propagated signal embodied in the propagation medium, as described above for computer program propagated signal product.

[00113] Generally speaking, the term "carrier medium" or transient carrier encompasses the foregoing transient signals, propagated signals, propagated medium, other mediums and the like.

[00114] Alternative embodiments can include or employ clusters of computers, parallel processors, or other forms of parallel processing, effectively leading to improved performance, for example, of generating a computational model.

EXAMPLE 1 – GENETS: A WEB PLATFORM FOR THE NETWORK-BASED ANALYSIS OF GENETIC VARIANTS

[00115] The interpretation of inherently complex and experimentally noisy cell circuitry data requires the convergence of big biological data, machine learning, visualization, and cutting-edge web technologies. This motivated the development of GENETS, the Broad

Institute Web Platform for the Network-Based Analysis of Genetic Variants, in accordance with an embodiment of the present invention. GENETS is positioned to become the future of web-based network analysis for the world-wide scientific community.

[00116] The field of machine learning has contributed to analyzing biology but has also been influenced by biological concepts (see Tarca AL et al. "Machine learning and its applications to biology," PLoS Comput Biol 3(6): e116, 2007. doi:10.1371/journal.pcbi.00301167). For example, one of the earliest machine learning fields, Artificial Neural Networks, has its roots in the modeling of neurons (see Warren McCulloch & Walter Pitts, "A Logical Calculus of Ideas Immanent in Nervous Activity," Bulletin of Mathematical Biophysics 5 (4): 1943, 115–133. doi:10.1007/BF02478259). GENETS harnesses machine learning in its analytic layer to learn systematic topological rules in pathways so that it can determine direct probability statements concerning pathway genes in a user-defined gene list.

Methods

1. Technology Stack

[00117] GENETS was developed using several modern database, back-end, and front-end technologies. FIG. 21 is a schematic diagram illustrating an example of interactions between components of a system 2100 for analyzing biological network data. A server 2110 can interact with a database 2115 and an interactive analysis and visualization module 2120. A user can interact with the system 2100 to have the system do at least two primary computations: (1) building one or more models (classifiers) based on networks and a set of pathways, and (2) running gene set analyses based on saved models. There are other features in the system and a user can, for example, just use the visualization features that can side-step computational wait time. One or more processors 2125 are configured to generate the model(s) and run the analysis and visualization procedures. Processor(s) 2125 may be distributed processors, and elements of system 2100 may communicate through one or more networks, including cloud-based data networks, as illustrated, for example, in FIG. 19. Models resulting from the analysis of the biological network data can be saved to data store 2130. Alternatively or in addition, models can be saved to database 2115. Model statistics, annotations or both can be saved to database 2115, as can gene set analysis results.

[00118] In an example embodiment, the server 2110 is JavaScript based Node.js and the database 2115 is NoSQL MongoDB. Node.js is a javascript based server and MongoDB is a

javascript based open-source document database that enables storage and retrieval of data that is not tabular as in relational databases. These technologies were chosen for the bottom two layers of the stack because a feature of GENETS is the visualization on the client-side (see FIGs. 22A-B), which, in one example, is completely developed with the JavaScript library d3.js. This consistency across all layers of the application facilitated rapid application development and coherency in the design. Furthermore, the utilization of the RMongoDB package in R (http://cran.r-project.org/web/packages/rmongodb/index.html) was a core functionality that enables direct communication between the back-end machine-learning algorithm(s) and the database that stores the results. The front-end development used Backbone.js (http://backbonejs.org/), jQuery (http://jquery.com/), Bootstrap.js (http://getbootstrap.com/), Underscore.js (http://underscorejs.org/), and d3.js (http://d3js.org/). In the particular embodiment illustrated, models are saved in ".RData" file format. The analysis and visualization module 2120 is GENEPATTERN, a platform for integrative genomics (http://www.genepattern.org/), described by Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP, "GenePattern 2.0," Nature Genetics 38 no. 5, 2006, pp. 500-501.

[00119] FIG. 22A is a screenshot 2200 of the GENETS pathway analysis platform, an embodiment of the present invention. A user can access the aforementioned networks, analyze (2205), visualize (2210) and maintain (2215) biological network data. As user can submit gene sets for pathway analysis. In addition, a user can upload and integrate the user's own network data and pathway list(s).

[00120] FIG. 22B illustrates several analysis and visualization tools available to a user.

[00121] For example, embodiments allow a user to interact with and explore gene sets, as illustrated at 2220. As shown, gene 2222 ("NUMA1") and gene 2224 ("SMC3") are connected to each other and are highlighted for analysis. A user can move the network, zoom in an out, and move nodes.

[00122] As illustrated at 2230 of FIG. 22B, embodiments allow for visual gene set enrichment. Shown is a set of genes 2232 that are cohesin complex related. The genes shown are compared with other functionally related gene sets and tested for significant overlap with these gene sets using hypergeometric testing. As illustrated, the genes are then encoded with 'bubbles' that visually indicate the shared membership with significantly overlapping functional gene sets.

[00123] As illustrated at 2240, embodiments provide for drag and drop annotations of genes. Here, genes 2242 ("SMC1A"), 2244 ("SMC3"), 2246 ("STAG2") and 2248 ("RAD21") are genes found to be significantly mutated across 21 tumor types (Lawrence, M. S. et al. "Discovery and saturation analysis of cancer genes across 21 tumour types," Nature 505, 495–501, 2014).

[00124] As illustrated at 2250 of FIG. 22B, embodiments can provide the user with faster PubMed lookup than conventional analysis platforms. For example, when a user clicks on (2252), or otherwise selects, the connection between genes 2222 and 2224, a window 2254 opens to display information regarding the genes, such as evidence for the genes, e.g., evidence of their identification and connection. The information can include publicly available information and methodological data.

[00125] As illustrated at 2260, another analysis tool is related to identifying drug-target interactions. Two example gene networks (2265-a, 2265-b) are shown, both including genes from a GWAS gene set and candidate genes. In network 2265-a, an established drug 2262 ("mebendazole") and its connection to a candidate gene ("TUBB2C") are shown. In network 2265-b, two novel drugs 2264 ("vorapraxar") and 2266 ("rusalatide") and their connection to a candidate gene ("E2RL3") are shown.

2. Process

[00126] Background

[00127] The last decade has exhibited a distinct trajectory from the "single gene hunt" strategy to a systems biology approach founded upon the recognition that macromolecules work in concert to produce phenotypes. The evolution of pathway analysis has transitioned through Over-Representation Analysis, Functional Class Scoring, to Pathway Topology Based approaches (see Khatri P, et al., "Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges," PLoS Comput Biol 8(2): e1002375, 2012, doi:10.1371/journal.pcbi.1002375). The analytics underlying GENETS can be classified in this 3rd generation of topology-based techniques. In the spirit of the "walks like a duck and talks like a duck" logic, the algorithm is named "Quack" because it is designed to learn universal topological rules, or profiles, of pathway genes to classify pathway from non-pathway genes. Similar applications of using known pathway as a reference set and also training a classifier on such a set have been predictive and highly successful in the gene expression space (see Subramanian A, et al., "Gene set enrichment analysis: a knowledge-

based approach for interpreting genome-wide expression profiles," Proc Natl Acad Sci USA. 2005 Oct 25;102(43):15545-50. Epub 2005 Sep 30) and clinical oncology (see Tamayo P, et al., "Predicting Relapse in Patients With Medulloblastoma by Integrating Evidence From Clinical and Genomic Features," J Clin Oncol. Apr 10, 2011; 29(11): 1415–1423. Published online Feb 28, 2011. doi: 10.1200/JCO.2010.28.1675), respectively.

[00128] The typical use-case in network based analysis begins with a gene set. This gene set is a function of prior beliefs and experimental data, be it, for example, a list of differentially expressed genes or a list of proteins that have been pulled down with a set of baits in a proteomics experiment. The user may want to determine any or all of the following: 1. Is this gene set 'significant' based on some definition of significance; 2. What is the structure and important components of the gene set; 3. If the gene set is a variant of a known pathway or the results suggest a functional implication or association, then also predict new genes that were not in the original list (i.e. candidates). Ultimately, the user may want to understand how the gene set fits into the cellular landscape and whether or not the resulting structure is suggestive of unknown biology that can be experimentally tested.

[00129] This use-case motivated the design of Quack to be predicated upon the learning of known pathway topologies so that one can provide a direct probability statement regarding the pathway profile of each gene in a gene set and use the same mechanism to predict new genes for follow-up study.

[00130] Learning Phase

[00131] The logic of Quack begins with a reference set of pathways and a reference network. The reference set of pathways is sampled from a data store, e.g., MSigDB curated pathways, based on the connectivity of the gene sets to ensure sufficient data for modeling. The user can also upload their own reference set or select a reference set based for a specific class of pathways, e.g., regulatory or signaling pathways. For each pathway, the topological profiles of pathway genes and their 1st order context (i.e. any adjacent node to any node in the pathway) are measured. Topological properties, also referred to as node properties, are described above in reference to FIGS. 6-13. Further illustrations of node properties can be found in Examples 1 and 2.

[00132] Next, Quack uses these profiles to build a computational model to differentiate between the pathway genes and their first order context. This computational model is built using the supervised machine learning technique of Random Forests. This methodology is

ubiquitous in machine learning applications and has been used successfully in biological network modeling such as Network Guided Forests (see Dutkowski J, Ideker T "Protein Networks as Logic Functions in Development and Cancer," PLoS Comput Biol 7(9): e1002180, 2011, doi:10.1371/journal.pcbi.1002180).

[00133] The binary response is defined as Y={1: molecule is a member of the pathway; 0: molecule is not a member of the pathway, but is directly connected to one or more of the members of the same pathway}. First, each of the topological metrics is computed using the connections amongst the pathway members. In this particular example, there are 18 topological metrics. Next, each non-member is added individually to the set to determine its characteristics assuming it is a member of the pathway. This is called the set of non-members, with direct connections to a pathway, the *context*. To ensure sufficient sample of both pathway members and context, one can down-sample to a maximum of 2,000 non-members for each pathway. The process can be continued across hundreds of pathways until a modeling data set is created. In this particular example, the random forest model uses 500 trees where each tree is built upon a subset of the data and a subset of the topological properties. These properties are used to define branches of the tree (an example path down the branches of this tree could be: degree is > 2 and the eigenvector centrality is <0.5 and the closeness centrality is > 0.7, this sub-population of pathway and context genes resulted in 30 pathway genes and 70 context genes, therefore the probability of being a pathway gene is 30% and if another gene meets these criteria it will be assigned a probability of 30% by this sub-tree, to be used in the "voting" with all the other sub-trees). Each of the sub-trees in the forest assigns a probability of pathway membership based on this sub-sample of data and predictors and this ensemble is used to predict an overall score which is the average of the probabilities across trees. Binary indicators can also be used for each tree such that if the probability > 0.5, the vote is 'yes' and votes are collected across the forest to compute the probability (i.e., % of trees that determined the gene is a pathway member). This forest is saved in the data-store and a new gene set is scored by computing the 18 topological properties for the members of the gene set and its context in the same manner as the training of the model (see, e.g., FIG. 15 and FIG. 18 and associated description). In this particular example, each member is assessed using each of the 500 sub-trees and each tree assigns a probability, then the probability is averaged across all trees in the forest for a final

probability. This process is done for each gene set member and each gene in the context of the new gene set being scored.

[00134] FIG. 23 illustrates Quack probability distributions of pathway genes (left panel, number of pathway genes = 25,129) and context genes (right panel, number of context genes = 600,898) for the analysis of data from 647 pathways. The figure illustrates that probability assignments are fundamentally different between pathway genes and context genes. Note that the Y-axes in the left and right panels are not the same.

[00135] Looking at the probability distributions of pathway genes, such as those illustrated in FIG. 23, one can formulate at least two hypotheses about learning pathway connectivity profiles. According to a first hypothesis, rules may be unique for each background network. According to a second hypothesis, rules may be unique within different classes of pathways.

[00136] FIG. 24 is a diagram illustrating pathway classes, including metabolic, signaling, regulatory, immune, and optimal selection pathways. For first three classes in the class list 2400, naïve text mining of gene set names in the MSigDB was conducted to determine class membership of regulatory, signaling and metabolic gene sets. A Venn diagram 2410 illustrates the class memberships for regulatory 2406, signaling 2404 and metabolic 2402 genes and their respective class overlap. For the fourth class, denoted immune pathways, immune data was derived from data of the Human Immunology Project Consortium. Immunological signatures were defined directly from microarray gene expression data from immunologic studies. The fifth class was optimal selection of pathways based on sufficient connectivity.

[00137] FIG. 25A illustrates results of the analysis of average path length by network and pathway class. Pathway classes are shown on the x-axis and are as defined above with reference to FIG. 24, i.e., immune (I), metabolic (M), optimal (O), regulatory (R), and signaling (S). Average path length ("aplCon") is shown on the y-axis. Data are grouped by networks, which are the networks described elsewhere herein, namely Achilles, CLIME, GEO, InWeb+ and L1000. For each network, data are grouped by pathways. Each data point represents the average path length between pathway genes in a specific pathway. Also shown are a reference line and a box plot of the distribution by network illustrating that the average path length varies by network and pathway class.

[00138] FIG. 25B illustrates results of the analysis of clustering coefficient by network and pathway class. The figure is similar to figure FIG. 25A, except that the value shown on the

y-axis is clustering coefficient ("gcc1"). The horizontal line is a reference line (clustering coefficient of 0.5). The figure illustrates that clustering coefficient varies by network and pathway class.

[00139] FIG. 25C illustrates results of the analysis of modularity by network and pathway class. The figure is similar to figure FIG. 25A, except that the value shown on the y-axis is a measure of modularity ("commNew"). Modularity as used herein is defined as the number of communities (i.e., significantly connected sub-network). The horizontal line is a reference line (modularity of 10). The figure illustrates that modularity varies by network and pathway class.

[00140] FIG. 25D illustrates results of the analysis of density by network and pathway class. The figure is similar to figure FIG. 25A, except that the value shown on the y-axis is density ("density"). Density as used herein is defined as the number of connections divided by the possible number of connections in the pathway. The horizontal line indicates reference density of 0.1 (10%.) The figure illustrates that density varies by network and pathway class.

[00141] FIG. 25E illustrates results of the analysis of context size by network and pathway class. The figure is similar to figure FIG. 25A, except that the value shown on the y-axis is context size ("context"). Context size as used herein is defined as the number of genes that are directly connected to the respective pathways. The horizontal line indicates reference context size of 3k, i.e., 3,000. The figure illustrates that density varies by network and pathway class.

[00142] FIG. 26 illustrates classifier performance by network and pathway class. Performance is measured as area under the receive operating characteristic curve (AUC), as described elsewhere herein. The AUC data are based on the 30% of the data not used for generating the model data set (30% holdout). The horizontal line is a reference line (AUC of 0.8). The figure illustrates that performance varies by network and pathway class.

[00143] FIG. 27 is a graph illustrating actual and predicted probabilities including changepoints. Actual (solid line) represents the actual true positive rate in that group, and predicted (dotted line) represents the average probability assigned for genes in that group. Changepoints (breakpoints) in the predicted curve are indicated with circles. The graph illustrates filtering to a "top tier" or highest likelihood set of candidates. Embodiments can include grouping genes based on predicted probabilities obtained using the computational

model. For a detailed description of changepoints, see Killick et al., "Optimal detection of changepoints with a linear computational cost," October 10, 2012.

EXAMPLE 2 – A SYSTEMATIC ANALYSIS OF DIFFERENTIAL PATHWAY TOPOLOGIES IN DIVERSE FUNCTIONAL GENOMICS NETWORKS

[00144] High-throughput technologies in genomics, genetics, epigenetics, transcriptomics, and proteomics have led to the generation of heterogeneous biological networks that connect genes if they are functionally correlated in any of the aforementioned data types. These networks share global design features by being scale-free, small world, and modular and have the potential to catalyze genomic interpretation, systems biology, and therapeutic intervention. However, it remains challenging to systematically map and understand how biological signal is organized within and between networks. Described is a general statistical method to quantify pathway relationships using 18 topological properties across 853 pathways in gene networks of correlated mRNA expression, phylogenetic patterns, cancer codependency relationships, cell perturbation profiles, and protein-protein interactions. Using this method, one can show that pathway topologies (e.g., architectures) diverge significantly between networks and between classes of pathways within each network (e.g., cell signaling, metabolism, and cell regulation), illustrating that despite similar global designs, pathway relationships are differentially organized in heterogeneous networks. A web platform (GeNets) is provided for harnessing network-specific pathway topologies to optimize biological discovery and for the scientific community to compare, visualize, and share genome-scale networks through a standardized framework.

[00145] **Introduction**

[00146] Following the technological breakthroughs of the human genome project (1), large-scale methods to map genomic, transcriptomic, and proteomic data have become ubiquitous in natural and biomedical sciences. Integrating the resulting large data sets into functional genomics networks (where nodes represent genes and edges represent some functional association between gene pairs) is a powerful and convenient way of representing the data to decipher complex biological relationships that would not emerge without a holistic view of the functional associations at genome scale (reviewed in 2,3). This is in part because these networks describe biologically relevant gene-gene relationships that have not been elucidated by smaller scale experiments.

[00147] Remarkably, diverse functional genomics networks share similar global design by being scale-free (meaning that most genes have a small number of connections and a few genes have a large number of connections (3, 4, 5)), small world (meaning that shortest path length between any gene pair in the networks is relatively small (6)), and modular (meaning that there are sub-networks of gene sets that connect more significantly to each other than to the rest of the network (2,7,8)). Analyses have further shown that these global rules reflect fundamental principles of molecular biology as most genes in the same subnetwork work in concert to execute specific cellular functions - i.e., as molecular machines, signaling circuits, enzymatic cascades or rigid topological structures.

[00148] Understanding the global network-based location and properties of genes have important implications for biology and phenotype-to-genotype relationships. For example, genome-wide association studies or exome-sequencing studies both in somatic cancers, Mendelian diseases, and common complex disorders have revealed that genes in close proximity of each other (and often in the same sub-network or module) in these networks often are affected by genetic variation linked to the same disease. These results illustrate that functional genomics networks can serve as a model to interpret large genomic data sets and suggest targeted and cost-efficient follow up experiments (9, 10, 11, 12, 13, 14 and reviewed in 15). Other analyses of global topologies have indicated that mutated genes central in the networks are more often embryonic lethal and have deleterious effects on gene function, and that genes involved in diseases are often located more in the periphery of functional genomics networks because mutations in these genes are tolerated from an embryonic developmental perspective, but can affect human health later in life (16).

[00149] However, as many complementary functional genomics network become available, and as they grow in size, their inherent complexity (both in terms of nodes and edges) makes it challenging to understand and quantitatively map how biological signal is organized within and between different types of networks at a higher resolution than allowed for by conventional stand-alone approaches involving centrality, proximity, or clustering. It is also a significant technological challenge to provide standardized and computationally efficient statistical methods with the needed flexibility to extract the network-specific signals that will optimize biological discovery from any genome-scale network.

[00150] Here, a Random Forest (RF) model is presented to quantify pathway relationships using 18 topological properties of pathways from the Molecular Signatures Database

(MSigDB) across heterogeneous networks of gene-gene correlations based on i) mRNA expression patterns in tissue samples from the Gene Expression Omnibus (22), ii) cancer codependency relationships from project Achilles (23), iii) phylogenetic patterns from inferred models of evolution (24), iv) cell perturbation profiles of eight cell lines from the LINCS project (25,26,27), and v) protein-protein interactions between 12,509 human proteins from InWeb (9). Specifically, the present approach sought to: i) quantify and compare how biological pathways are organized across heterogeneous networks frequently used in medical sciences, ii) quantify and compare the differential architecture of classes of pathways (e.g., cell signaling, metabolism and regulation) within each of the five network types, and iii) formulate a Random Forest (28) classifier that is capable of harnessing network- and pathway-specific topological rules to optimize biological discovery from any functional genomics network.

[00151] Results described here show that despite many biological networks having similar global design principles, biological signal at the pathway level is significantly differentially organized between heterogeneous networks and that this differential organization can be systematically deciphered and harnessed to optimize biological discovery from any type of network. To make these insights actionable and widely accessible to the biomedical community, a unified web platform named GeNets was designed which enables the analysis of complicated pathway relationships based on any functional genomics network. GeNets can be used to quantify and visualize pathway relationships from any network to interpret the functional significance between genes emerging from the newest genomic platforms.

[00152] **Results**

[00153] **Analyzing eighteen topological pathway metrics across networks**

[00154] For a given network, it was hypothesized that genes in a common pathway would share pathway-specific topological properties that can be systematically deciphered and harnessed to distinguish them from genes that are not part of the pathway in question.

[00155] Embodiments of the present approach created and used pre-existing networks from the following sources (Methods): 1. gene-gene correlations based on mRNA expression patterns in 19,019 tissue samples from the Gene Expression Omnibus (GEONet, hereafter); 2. Cancer codependency relationships across 216 cancer cell lines from project Achilles (AchillesNet, hereafter); 3. Phylogenetic relationships from 'clustering of inferred models of evolution' between genes in 502 species (CLIMENet, hereafter); 4. Cell perturbation profiles

from eight cell lines from the LINCS project (LINCSNet, hereafter); and 5. 428,429 protein-protein interactions between 12,509 human proteins (from the InWeb database). Table 1 provides a summary of final network sizes after pre-processing. For details of the pre-processing, including the removal of indirect edges from matrix data (29, 30), thresholding edge scores and optimizing sparse network sizes, see Methods and Supplementary Notes, 1-4.

Table 1: Final Network Sizes

Network	# of nodes	# of edges
InWeb	12,357	394,069
AchillesNet	9,219	500,000
LINCSNet	6,721	500,000
GEONet	12,390	500,000
CLIMENet	8,279	500,000

[00156] Using the InWeb network and the Biocarta AKT pathway, six topological metrics (e.g., topological properties) were defined that describe the relationships of a gene (e.g., AKT1) to other genes in the same pathway (i.e., betweenness centrality in pathway, weighted degree in pathway, clustering coefficient in pathway, closeness centrality in pathway, eigenvector centrality in pathway, and degree in pathway; see Methods for a detailed description of these metrics). The analogous six metrics for AKT1 in the overall InWeb network (e.g., the betweenness centrality in the overall network) were also computed and a ratio between the pathway-specific metric and the overall network metric was derived (e.g., betweenness centrality in pathway / betweenness centrality in overall network). Expanding this calculation to all genes in the AKT pathway resulted in a total of 18 metrics being calculated for each of the 21 AKT pathway genes. To look for topological properties that systematically distinguished AKT pathway genes from other genes in InWeb, we also computed these metrics for 2,449 genes that are in the context of the AKT pathway. Hereafter, we define the context of a specific pathway (e.g., the AKT pathway) in a specific network (e.g., InWeb) as all genes that are not part of that pathway set, but have least one connection to a gene in the pathway under investigation. This resulted in a set of 21 data points for each topological metric for the AKT pathway genes and 2,449 data points for each

topological metric for the AKT context genes. This data was then used to show the topological differences between the AKT pathway members and context genes (see FIG. 28A).

[00157] To systematically map the topologies of many pathways in InWeb, the analysis above was repeated for 853 pathways from the MSigDB database (after ensuring that this set was non-redundant). A univariate analysis of the distributions of scores for pathway genes versus context genes for each of the 18 metrics (FIG. 28B) confirms the hypothesis that there are topological signatures that clearly distinguish genes that together form a pathway in InWeb, from genes that are not part of the pathway in question.

[00158] Expanding this analysis to all five networks revealed two pathway topological principles: First, in all networks, the distributions of these metrics are generally different between pathway genes and context genes (see, e.g., 6 of 18 metrics illustrated in FIGs. 28B-28C). This means that when considered on the background of a complex set of network properties, genes in a common pathway have an topological signature that distinguishes them from other genes in the network. Second, one can observe differential pathway topologies in the five networks, meaning that for each network, the distributions of topological metrics for pathway members form a network-specific signature (see, e.g., partial signature with 6 of 18 metrics illustrated in FIG. 28D).

[00159] FIGs. 28A-28D illustrate differential pathway topologies across functional genomics networks. As shown in FIG. 28A, for a given pathway, its topological properties are measured, exemplified here with the 22 genes of the AKT pathway in the InWeb protein-protein interaction network. The same measurements can were made for all genes in the AKT pathway context set (squares), in this case 2,449 genes (only 2 of which are shown for illustration) that have at least one connection to an AKT gene in InWeb. The distributions for 4 of 18 topological properties [Weighted Degree Ratio, Closeness Ratio, Eigenvector (P) and Closeness (P)] are shown and illustrate the differences between pathway (dark) and context (light) distributions. This procedure is repeated for 583 non-redundant pathways in the InWeb network. As illustrated in FIG. 28B, the distributions of the broader population show that genes in a common pathway have an topological signature that distinguishes them from context genes. As illustrated in FIG. 28C, repeating the procedure (detailed with respect to FIG. 28B) for the other four networks shows this is a general principle. When quantified and compared, as illustrated in FIG. 28D, it is clear that each network has a unique distribution of

topological metrics (shading in FIG. 28D is as indicated in FIGs. 28B and 28C). In all panels of FIGs. 28A-28D, the x-axis denotes the respective metrics and the y-axis is the relative frequency (density) of observations. The following abbreviations are used in the figures: interaction (int.), member (Mbr.), distribution (dist.), weighted (Wt.), pathway (P), overall network (N); e.g. Eigenvector (P) denotes the Eigenvector centrality in the pathway.

[00160] Learning biological signal across heterogeneous networks

[00161] A Random Forest classifier was employed to assess whether the complex topological rules observed in FIGs. 28A-28D can be harnessed to learn biological signal across any type of network (see Methods).

[00162] FIGs. 2A-2B, described above, illustrate building a general classifier to predict pathway membership from networks. Using the InWeb network and 853 pathways, the following calculations are repeated: First, compute all 18 topological properties for each gene in the pathway in question (FIG. 2A) and add these observations to the modeling dataset, where the gene is indexed as a row and the eighteen metrics as columns in that row. Second, determine the context gene set for the pathway in question. For each gene in the context set, compute the 18 topological properties under the assumption that the gene is a member of the pathway (FIG. 2A) and add a subset of these observations as one row per gene in the modeling dataset to ensure a more balanced representation (31) of pathway members to context genes (see Methods). The pathway member and context gene observations from all 853 pathways are then combined to obtain a modeling dataset of 752,172 rows with each eighteen columns (FIG. 2B). Randomly sample 70% of the pathways as the training set and use the remaining 30% of the pathways for validation.

[00163] A Random Forest classifier was constructed using the training data to build an ensemble of 500 smaller classifiers (i.e., the forest). In each of these, a subset of the topological metrics is used to construct a tree that maximizes the segmentation of pathway members from context genes (FIG. 2B). To assign a probability that a gene in the validation dataset belongs to a specific pathway, the gene's eighteen topological metrics (in relation to that specific pathway) are used as input to the forest, whereby each of the 500 trees each cast a vote. The probability that a gene belongs to the pathway in question is the proportion of trees that votes the gene as a pathway member (FIG. 2B). Because the method predicts whether a gene is part of a pathway based on the way it 'walks and talks,' the classifier is named "Quack" and will be referred to as such hereafter.

[00164] FIGs. 29A and 29B illustrate using network-specific topological rules to recapitulate pathway relationships. FIG. 29A illustrates that the Quack classifier can efficiently learn the network-specific topological organization of pathways and consistently classifies pathway genes across networks with the following areas under the receiver operating characteristics curve (AUC) in a 30% holdout analysis (InWeb, 0.93 C.I. [0.93 - 0.94]; CLIMENet, 0.86 C.I. [0.85 - 0.87]; AchillesNet, 0.85 C.I. [0.84 - 0.86]; GEONet, 0.84 C.I. [0.83 - 0.85]; LINCSNet, 0.84 C.I. [0.84 - 0.85]). Referring to FIG. 29B, a leave-one-out cross-validation approach similarly illustrates the predictive power of Quack as the held-out genes are generally ranked >95th percentile regardless of the type of network it is trained on.

[00165] When using this approach to differentiate between pathway and context genes in the validation data, the area under the receiver operating characteristics curves (AUCs) range from 0.84 to 0.93 for the five networks (FIG. 29A). To further assess the predictive power, one can also remove one of the pathway members from the pathway set to assess whether the trained classifier can recover this gene (i.e., assign a high probability that the gene is, in fact, part of the pathway from which it was held-out). Proceeding in this manner, one can observe that the held-out genes are consistently ranked above the 90th percentile as compared to context genes (FIG. 29B). To further assess if the forest size impacted the results, a five-fold cross-validation was conducted using forest sizes of 100 to 500 trees. It was found that larger forest sizes had little impact on learning network signal. It was also found that edge weights (i.e., the relative strength of connection between two genes in the network being tested) provided a marginal increase in AUC percentage points of 2-3% when compared to the same networks with only binary (i.e., 0 of 1) edges.

[00166] These results illustrate that, despite fundamentally different topological signatures across networks, the approach described herein effectively learns how to harness network-specific biological signals to identify pathway relationships.

[00167] Differential pathway topologies in heterogeneous network data

[00168] FIG. 30 illustrates ranking the importance of pathway topological metrics across networks. By permuting the values of each topological metric being evaluated it is possible to estimate the overall importance of each topological metric across networks. Here, the topological properties are in descending order by their average rank across networks. The y-axis is the rank (1-18), where 18 is most important metric for distinguishing pathway

members and 1 is least important. The following abbreviations are used: weighted (Wt.), pathway (P), overall network (N) and local clustering coefficient (LCC), so that LCC (P) and LLC (N) means local clustering coefficient in the pathway and network, respectively. Closeness and Eigenvector centrality are consistently important across networks (column 18 and 17, respectively), while there is significant variation in the predictive power of, e.g., the local clustering coefficient in the network [LCC (N), column 8]. One can also observe that some metrics such as the degree in the pathway (column 1), are less important in all networks when controlling for others topological metrics.

[00169] As noted above, to quantify the relative importance of the various topological metrics in defining pathway relationships, one can permute the values of each topological metric being evaluated within each tree and compare the average classification error across trees before and after permutation. This provides an estimate of the overall importance for each topological metric when determining pathway relationships in the network under investigation.

[00170] This analysis was applied to quantify the extent to which the diverging signatures (illustrated in FIGs. 28A-28D) are mirrored in varying importance of topological metrics when Quack is applied to different networks and found that although there are metrics for which the predictive power is generally high or low, there is considerable variation across networks (FIG. 30). For example, the eigenvector centrality within a pathway is a consistently top-ranked metric across networks. Interestingly, most of the ratios derived from the respective pathway-to-network metrics were also ranked highly (four of six metrics based on ratios are ranked in the top 10). For example, the eighth most important metric (based on the average rank) is the ratio of the degree in the pathway of a gene to that of its degree in the overall network. In contrast, the degree in the pathway evaluated alone is unimportant across all networks (ranked 18th, i.e., the least important metric overall). This proves the intuition that - in any network - the number of connections a gene has within a given gene set (which is sometimes taken as naive support for a gene's role in a biological process of interest) can be highly misleading if not evaluated on the background of its total number of connections in the network in question.

[00171] Conversely, there is significant variation in the predictive power of the local clustering coefficient in the network, degree in the network, and the betweenness and eigenvector centrality in the network. Where betweenness centrality in network was the most

important metric for gene networks based on coexpression across gene expression omnibus (GEONet) and protein-protein interactions (InWeb), it is one of the least important for those based on cancer synthetic lethalities (AchillesNet), phylogenetic relationships (CLIMENet), and cell perturbation profiles (LINCSNet). Similarly, the local clustering coefficient in network is the most important metric for CLIMENet, and the fourth most important for InWeb, while being relatively unimportant in AchillesNet, LINCSNet, and GEONet.

[00172] To illustrate differences in the organization of pathways across networks in the context of Notch signaling, PI3K signaling, glucolysis and gluconeogenesis and oxidative phosphorylation, the eigenvector centralities were mapped to genes in these pathways across heterogeneous networks. Despite the consistently high rank of this metric across networks, there is considerable divergence in the patterns and strengths of the values across networks, which illustrates how pathway relationships manifest differently across networks.

[00173] **Pathway stratification into subclasses improves biological discovery**

[00174] Three major classes of cellular pathways in MSigDB are those involved in metabolism, the regulation of genes and cellular processes, and the transmission of signals inside and between cells. One can hypothesize that, within a network, the organization of different classes of pathways - despite the broadness of these classes - would also exhibit differential topologies. If so, there would be an opportunity to improve biological discovery through pathway stratification.

[00175] In an embodiment, pathways were classified based on text-mining which was validated with manual curation (Methods) to ensure robust class assignment which resulted in 323 signaling pathways (e.g., PID Wnt Signaling Pathway), 230 metabolic pathways (e.g., Reactome Purine Catabolism), and 397 regulatory pathways (e.g., KEGG Regulation of Autophagy). For details, see Methods and FIG. 24. First, the overall importance of pathway class was tested by considering it as another feature in the classifier. A high rank of this feature indicates that there are differences in the topologies of classes of pathways. Conversely, if the feature is not ranked highly, that can be an indication that topological metrics independently of the pathway class provide sufficient information. Here, it is found that pathway class is ranked 6th amongst other topological metrics when modeling pathways in InWeb and CLIMENet, whereas in GEONet, AchillesNet, and LINCSNet the pathway class does not make it in the top 10 features of the model.

[00176] To investigate and quantify differential topologies between classes of pathways and their implications when predicting biological signal across the five different networks, the above-described modeling process (FIGs. 2A-2B) was repeated for each pathway class within a network. One can observe a notable tendency for the performance to be suboptimal when attempting to predict pathway membership in a given class such as regulatory pathways, when the classifier was trained on the data of another class such as metabolic pathways (see FIG. 31).

[00177] FIG. 31 illustrates that regulatory models show improved biological discovery of regulatory pathway members. Classifiers trained on regulatory pathways across networks improve discovery of regulatory pathway members when compared to classifiers trained on other classes of pathways.

[00178] Across all networks (e.g., InWeb, CLIMENet, GEONet, AchillesNet, and LINCSNet), it is a general observation that predicting biological signal can be optimized by applying specific topological features inherent to the type of the pathway under investigation.

[00179] Furthermore, while there is evidence that metabolism, regulation and signaling exhibit differential topologies in InWeb and CLIMENet, there is less distinction in architecture between different pathway classes in GEONet, AchillesNet and LINCSNet.

[00180] **Discussion**

[00181] Through the systematic investigation of 18 topological properties across 853 pathways in gene networks of correlated mRNA expression, phylogenetic patterns, cancer codependency relationships, cell perturbation profiles, and protein-protein interactions, embodiments of the present approach have quantified and compared the network-specific organization of pathway information across heterogeneous networks that are widely used in the biomedical community.

[00182] The above analysis and results have shown that the organization of pathway signal is network-specific and harness this fundamental principle to formulate a machine learning approach, Quack, that can accurately assign probabilities that genes, in any gene set, form a pathway in any network. Given a set of genes, the Quack classifier can also predict likely new pathway genes amongst context genes. The complexity of biological networks and the pathway relationships they encode is illustrated by the fact that the median size of pathways is 48 genes, while the contexts of most pathways are several orders of magnitude higher. This inherent complexity of biology underscores the importance of developing scalable and

efficient machine learning strategies for elucidating functional insights from genome-scale networks.

[00183] It could be asserted that Quack is biased by knowledge contamination (i.e., the idea that genes that are part of the same pathway are more studied in relation to each other and therefore have more connections in a network) when applied to the InWeb protein-protein interaction network. However, the four other networks used in this analysis (CLIMENet, AchillesNet, GEONet and LINCSNet) are built from systematically exploring gene expression datasets or the the alignment of genomes and are therefore completely independent of this form of knowledge contamination. Therefore, it is all the more noteworthy that the biological principles observed for InWeb are mirrored in analogous observations in CLIMENet, AchillesNet, GEONet and LINCSNet. Although the performance of Quack on the InWeb protein-protein interaction network is higher than most other networks, the approach is able to accurately predict pathway relationships in all five networks making it unlikely that potential study bias of proteins in InWeb has any major impact on our results or observations.

[00184] Correlation-based networks such as CLIMENet, AchillesNet, GEONet, and LINCSNet can be influenced by indirect relationships (i.e., the property that if nodes pairs A-B and B-C are strongly correlated in any of these networks independently of each other, this may create a link between A-C even if there is no direct biological information flow between A and C). Two processes were applied that eliminate indirect effects to all of the networks and the analyses were repeated (Supplementary Notes). This analysis shows that indirect effects do not significantly influence results and that Quack performs similarly with or without these preprocessing steps. Given that it can be computationally demanding (for example, it may take several days to eliminate indirect effects from large matrices and select optimal networks based on various potential thresholds), it is desirable from a technology standpoint that this pre-processing step is not necessary for the analytical framework described here.

[00185] In embodiments, analysis of pathway topologies is based on a comprehensive set of metrics encapsulating a gene's topological relationships in the pathway and in the entire network. This approach can also be extended to include information about the other types of information such as whether the pathway metabolic, regulatory, or involved in signaling. As shown, such a pathway stratification step can improve biological discovery and that the

distinction between pathway class topologies varies across networks with the greatest differences observed in physical interactions (InWeb) and phylogenetic patterns (CLIMENet).

[00186] To make results actionable for the biomedical community, a free web platform (GeNets) is provided for quantifying and expanding the pathway relationships within user-defined gene sets. Since the pathway models are trained prior to the submission of user-defined gene sets, the time required for scoring and expanding these gene sets are on the order of minutes, which facilitates the fast and iterative workflow needed to efficiently glean biological insights from large functional genomics datasets. Furthermore, GeNets enables advanced users to train their own pathway models by providing a network and pathway set. With GeNets, users across the biomedical spectrum can: (i) easily upload and analyze any network or select from the set of pre-packaged publicly available networks described in this publication, (ii) build pathway models based on our machine learning approach for any network and any pathway set, (iii) calculate the probability that genes in any user-defined set participate in the same pathway (based on any network and pathway model), (iv) interactively visualize and investigate the pathway relationships quantified by the models, and (v) share interactive networks and results with collaborators. Advantageously, the web platform scales well with the size of the networks it analyzes. Furthermore, built-in visualization tools enable the interpretation of complicated networks in a user friendly and intuitive manner, particularly useful for scientists that are not experts in network biology.

[00187] **Methods**

[00188] **Functional Genomics Data**

[00189] The following functional genomics data were employed by example embodiments, it being understood that other suitable data may be used.

[00190] Protein-protein interaction data: Reported interaction from the literature with associated credibility scores were used from the InWeb database (9).

[00191] Gene expression correlation data: Co-expression correlations were derived using several metrics reported from AFFYMETRIX arrays obtained from the Gene Expression Omnibus (22) (Supplementary Note 2).

[00192] Cell perturbation data: Similar global transcriptional responses (due to perturbations) were downloaded from the LINCS Connectivity Map (25, 26, 27)(Supplementary Note 2).

[00193] Cancer codependency data: Cancer codependency correlations from 216 cancer lines (23) were downloaded (Supplementary Note 3).

[00194] Phylogenetic pattern data: Similarity in phylogenetic patterns based on connections defined by the method of clustering by inferred models of evolution (24) were downloaded (Supplementary Note 4).

[00195] **Topological Metrics**

[00196] A set of topological metrics are described that can be used with embodiments of the present approach, it being understood that other metrics may be employed.

[00197] Let G=(V,E) be a graph with vertex (e.g., node) set V and edge (e.g., connection) set E. |V| = N is the number of vertices in the graph and |E| = M is the number of edges. Let A be defined as the adjacency matrix of G, i.e., the N × N matrix such that non-diagonal entries a_vw are positive real numbers (which depends on the network, see Methods - Functional Genomics Data for interpreting edge weights), and the diagonal elements are all zero (in all networks edges between the same gene [self interactions or self loops] are disregarded).

1) *Degree*: the degree of a vertex v is defined as the number of vertices directly connected to v (i.e., direct neighbors or just "neighbors").
2) *Weighted degree*: the weighted degree, also called the "strength", is defined as the sum of the weights of the edges which connect the neighbors to v..
3) *Clustering coefficient:* the clustering coefficient of a vertex v relates to the tendency of its first order interactors to also interact with each other. Technically it is defined as C_v = 1/(s_v*(k_v-1))*sum((wgt_vw+wgt_vu)/2 *a_vw * a_vu * a_wu) across w, u. Here, s_v is the strength of vertex v, 1/(s_v*(k_v-1)) is the normalization factor, a_vw is an adjacency indicator a_vw={0: no edge; 1: edge exists, k_v is the vertex degree, wgt_vw are the weights. C_v is continuous on [0,1]. As C_v approaches 1, the neighbors of v are becoming fully connected to one another. As C_v approaches 0, the neighbors of v are not well connected (i.e., a star with v in the middle has C=0).
4) *Closeness centrality*: the closeness centrality of vertex v is a measure of how close it is to all other vertices in the network. It is defined as (N-1)/sum(shortest_path(v,w), v != w), the inverse of the average shortest path length to all the other vertices w in the graph.
5) *Betweenness centrality*: the betweenness of vertex v is a measure of how many shortest paths between the graphs vertices go through v. It is defined as sum(

spath_uvw / spath_uw, u!=w,u!=v,w!=v), where spath_uw is total number of shortest paths from node u to node w and spath_uvw is the number of those paths that pass through v.

6) *Eigenvector centrality*: the eigenvector centrality of the vertex v is defined as x_v = 1/lambda * sum(a_vw*x_w) where lambda is the eigenvalue corresponding to the principal eigenvector (the eigenvector for which all entries are positive), a_vw is the value of the adjacency matrix corresponding to vertices v and w, and x_w is the component of the principal eigenvector corresponding to vertex w.

[00198] These six topological metrics are computed for genes (i.e., vertices) both (1) within a pathway using only the sub-network formed by the pathway genes, and (2) for the genes using the entire functional network. Further (3), the ratios of (within pathway / entire network) are computed for these metrics as well. When the denominator is zero the ratio is set to zero, otherwise, the natural logarithm ln(ratio) is computed. Therefore, in total 6 x 3 = 18 metrics are calculated for each gene.

[00199] **Training and Benchmarking Quack**

[00200] As described above, a pathway gene classifier was developed that uses the Random Forest methodology (28) to identify topological properties of pathway genes that can be used to differentiate between pathway and non-pathway members. The pathway membership is determined by gene sets in MSigDB curated pathways. The topological metrics listed above, which are well established in graph theory and social network analysis, were used to measure the characteristics of genes both in the context of pathways and their connectivity in the entire network under consideration. A set of new metrics was included by taking ratios of local and global topological properties (e.g., the number of interactions a protein has in a pathway versus how many interactions it has in the entire network) with the hypothesis that these relativities would improve the classification power. An expanded description of the metrics can be found in Supplementary Note 5. The binary response is defined as Y={1: is a member of the MSigDB gene set; 0: is not a member of the gene set, but is directly connected to one or more of the members}. First, each of the 18 topological properties was computed using the connections amongst the pathway members. Next, each non-member is added individually to the set to determine the non-member's characteristics assuming it is a member of the pathway. This set of non-members, with direct connections to a pathway, is called the context. These contexts vary in size with a median of 3,400. To

ensure sufficient sample of both pathway members and context, one can downsample to a maximum of 1,500 non-members for each pathway. The process is continued across many, e.g., hundreds of, pathways until a modeling data set of size 600,000 observations is created. The random forest model is trained on 70% of this data and the Area Under the ROC Curve (AUC) is used to assess the classification power using the 30% holdout. The number of trees is a parameter of the method and here we used 500. To assess if forest size impacted the results, a five-folded cross-validation was conducted using forest sizes of 100 to 500 trees. It was found that larger forest sizes had little impact on learning network signal.

[00201] Supplementary Note 1 - Details on Creating GEONet Gene Expression Network.

[00202] A co-expression network was derived from the Gene Expression Omnibus (22). This expression matrix was derived using several metrics reported from AFFYMETRIX arrays. Thresholding based on the distributions of the metrics was conducted such that extreme outliers would be excluded: First, the ratio of signal coming from the 3' versus the 5' end of the transcript of the gene beta-actin, a standard control gene. The reason for measuring this ratio is that transcription runs from the 3' to the 5' end of a gene. If mRNA samples are of good quality, one should observe equal amounts of signal from the 3' and 5' ends. However, if the 5' end is underrepresented, this suggests that the mRNA was degraded or wasn't labeled completely. A filter of ratio <= 2 was used based on the empirical distribution. Second, the same filter was applied to the gene GAPDH with a rule of ratio <= 2 based on the empirical distribution. Third, the average signal intensity across all genes from a given sample with a filter for samples with value >= 150 based on the empirical distribution. Fourth, the fraction of genes that the AFFYMATRIX platform has called as being reliably detected, or "present", was used to filter for samples with value >= 30. Finally, a power law regression is fit to 80 "invariant" genes in each sample. A goodness of fit for this regression was estimated and filtering was conducted for samples whose goodness of fit was at least 4 based on the empirical distribution. This process concluded in a matrix of 22,268 probes by 19,019 samples. Probes were collapsed to HUGO gene symbols by averaging and the resulting gene-by-gene correlation matrix was of size 12,716 by 12,716. Finally, we applied both global silencing (29) and network deconvolution (30) to this matrix and compared performance of the original matrix and silenced solutions at various network sizes (i.e., edge thresholds) as

outlined in Methods which resulted in the selection of the deconvoluted matrix thresholded to the strongest positive 1M gene expression correlations.

[00203] Supplementary Note 2 - Details on Creating the LINCSNet Cell perturbation Profile Network.

[00204] Here, L1000 data was utilized. L1000 is a high-throughput, bead-based gene expression assay in which mRNA is extracted from cultured human cells treated with various chemical or genomic perturbagens (small molecules, gene knockdowns, or gene over-expression constructs) as previously described (26). This mRNA is reverse transcribed into first-strand cDNA. Gene specific probes containing barcodes and universal primer sites are annealed to the first strand cDNA. The probes are ligated to form a template for PCR. The template is PCR amplified with biotinylated-universal primers. The end products are biotinylated, fixed length, barcoded amplicons. The amplicons can then be mixed with Luminex beads that contain complementary barcodes to those encoded in each of the 1000 amplified landmark genes. These 1000 landmark genes were chosen as a reduced representation of the transcriptome and account for the majority of expression variation across many cellular contexts (Subramanian, et al., manuscript in preparation). These beads are then stained with fluorescent streptavidin-phycoerythrin (SAPE) and detected in 384-well plate format on a Luminex FlexMap flow cytometry-based scanner. The resulting readout is a measure of mean fluorescent intensity (MFI) for each landmark gene. The raw expression data are log2-scaled, quantile normalized, and z-scored, such that a differential expression value is achieved for each gene in each well. In the standard L1000 protocol, each well corresponds to a different perturbagen and these differential expression values are collapsed across replicate wells to yield a differential expression signature for each perturbagen. The signatures of different perturbagens can then be compared to identify those that result in similar or dissimilar transcriptional responses as previously described (17, 25, 27). The similarity between all pairwise combinations of the roughly 460,000 signatures in the CMap database was computed. Then, a summary of these query results was performed to arrive at a more perturbagen-centric view of connectivity. To summarize, the query result is first grouped by cell line and perturbagen type (small molecule, gene knockdown, or overexpression). The connectivity scores are then normalized by dividing by the signed mean score of each group. The scores are converted to percentile ranks within each group. The perturbagens are then ranked according to the direction of connectivity. Positive connections

are ranked highest and negative connections ranked lowest. For each unique perturbagen, the average percentile rank was considered in the four cell lines for which the connection to the query was strongest. The matrix was then transformed into a symmetric matrix by averaging the (i,j) and (j,i) values. Next, both global silencing (29) and network deconvolution (30) was applied to this matrix and performance of the original matrix and silenced solutions compared at various network sizes (i.e., edge thresholds) as outlined in Methods, which resulted in the selection of the original matrix (i.e., non-silenced) thresholded at the strongest 1M edges.

[00205] Supplementary Note 3 - Details on Creating AchillesNet Cancer Codependency Network.

[00206] A codependency network was derived from the Project Achilles dataset v2.4.3 (23). This dataset is from RNAi screens of 216 cancer cell lines, each one infected with a pool of over 54,000 shRNAs each designed to knockdown one gene, for a total of more than 17,000 genes. These data were processed by the ATARiS method to yield dependency scores representing the degree of dependence of each cell line on each gene. Pearson correlation coefficients were computed from the dependency profiles of all pairs of targeted genes, resulting in a 17Kx17K correlation matrix. Finally, both global silencing (29) and network deconvolution (30) was applied to this matrix and performance of the original matrix and silenced solutions compared at various network sizes (i.e., edge thresholds) as outlined in Methods, which resulted in the selection of the deconvoluted matrix thresholded to the strongest 1M co-dependencies.

[00207] Supplementary Note 4 - Details on Creating CLIMENet Gene Coevolution Network.

[00208] CLIMENet was created using the algorithm CLIME (clustering by inferred models of evolution)(24) applied to 1,025 curated human gene sets from GO and KEGG (downloaded from http://www.gene-clime.org). CLIME predicts functionally related genes based on three inputs: a functionally-related gene set G, a species tree S, and a phylogenetic matrix containing presence/absence vectors of all genes in a reference genome across all species in S. Briefly, CLIME first partitions G into disjoint evolutionarily conserved modules (ECMs), and then scores all other genes in the genome a log-likelihood ratio (LLR) score to quantify the possibility that each gene has arisen under the ECM's inferred model of evolution compared to a background null model. The "expanded ECM" or ECM+ is the set of all other genes (not in G) with LLR>0. ECM+ genes share a similar evolutionary history

with a subset of G and thus may be functionally related. To create CLIMENet, CLIME was applied to all 909 GO cellular components and 116 KEGG pathways (using a human centric phylogenetic matrix and a 138-species eukaryotic tree, see www.gene-clime.org) – resulting in a total of 13,307 ECMs. Of these 13,307 ECMs, 10,606 are singletons (containing only one gene within G, indicating no shared evolutionary history with other pathway members). The resulting 13,307 ECM+ expansion sets contain a total of 667,592 genes, many redundant. In CLIMENet, genes A and B are connected by an edge if A and B occur together in at least one ECM+, and the edge weight is assigned the mean LLR scores for A and B across all ECM+s that contain both genes (LLR>0).

[00209] For example, the GO "voltage-gated potassium channel complex" contains 70 genes, which CLIME partitions into 28 ECMs (including 10 singletons). ECM1 contains 9 known potassium channel genes and the ECM1+ expansion contains 45 predictions with LLR>0 (most sharing the pfam domain "IRK" – which causes these genes to have a similar phylogenetic profile across 138 eukaryotic species). All gene pairs in ECM1+ (45*44=1,980 gene pairs) will be connected in CLIMENet due to the shared phylogenetic history (in this case due to the IRK protein domain). Note that the original 9 potassium channel genes in ECM1 are not used in the creation of CLIMENet – only the ECM+ genes. This avoids any circularity in assessing CLIMENet performance on MSigDB C2 (which has substantial overlap with the 1025 curated sets used to create CLIMENet).

[00210] Finally, the process applied both global silencing (29) and network deconvolution (30) to this matrix and compared performance of the original matrix and silenced solutions at various network sizes (i.e., edge thresholds) as outlined in Methods, which resulted in the selection of the original matrix (i.e., non-silenced) thresholded at the strongest 1M edges.

[00211] Supplementary Note 5 - A List of Network Metrics.

1. Eigenvector Centrality in the Pathway
2. Weighted Degree in the Network
3. Weighted Degree in the Pathway
4. Ratio of the Weighted Degree in the Pathway to that of the Network
5. Ratio of the Pathway Degree to that of the Network
6. Eigenvector Centrality in the Network
7. Closeness Centrality in the Pathway
8. Ratio of the Pathway Eigenvector Centrality to that of the Network

9. Ratio of the Pathway Closeness Centrality to that of the Network
10. Local Clustering Coefficient in the Network
11. Degree in the Network
12. Betweenness Centrality in the Network
13. Closeness Centrality in the Network
14. Betweenness Centrality in the Pathway
15. Degree in the Pathway
16. Local Clustering Coefficient in the Pathway
17. Ratio of the Pathway Local Clustering Coefficient to that of the Network
18. Ratio of the Pathway Betweenness Centrality to that of the Network

[00212] Supplementary Note 6 - Silencing indirect links (or deconvoluting) networks.

[00213] Matrices of functional genomics data contain both direct and indirect associations between gene pairs. To remove indirect effects and focus on direct functional relationships between pairs of genes, both global silencing (29) and network deconvolution (30) were applied. For each dense matrix (all functional genomics data previously listed with the exception of protein-protein interactions), the connections were ranked in descending order by the original connections weights, the deconvoluted weights, and the globally silenced weights. The process then filtered to the top 500K, 750K, 1M, 1.25M, and 1.5M connections. To assess what size networks to use in the next step of pathway modeling, both conventional permutation testing on 1,300 canonical gene sets from MSigDB was conducted and the classification power (AUC) of the classifier assessed for each of networks and sizes. For the permutation approach, for each gene set, the global clustering coefficient (with disconnected genes set to zero) was computed and then 500 gene sets with similar overall degree distributions were sampled from the respective network. The empirical p-value for each pathway can be determined by assessing the pathways clustering coefficient to that of null distribution, i.e., the % of randomly sampled gene sets with global clustering coefficient greater than or equal to the pathway under consideration. One can then plot the % of pathways among the 1,300 that are significantly connected (using a significance level of 0.10) and assess the behavior of this metric as the threshold for the top connections (i.e., increase the size of the network) is relaxed. Next, the process assessed the behavior of the AUC for the classifier as the network size for each of these networks was relaxed. It was

observed that the AUC's were consistently stable across networks at 1M edges. Finally, the learnings from both methods were incorporated and the top 1M edges selected from the deconvoluted cancer codependency network, original phylogenetic pattern network, the original cell perturbation network, and the deconvoluted gene expression network.

[00214] Supplementary Note 7 - Improvement of predictive power through pathway stratification.

[00215] To test whether pathway stratification both enabled the identification of specific topological rules and improved predictive power, pathways from MSigDB C2 (curated gene sets) and C5 (GO gene sets) were stratified by text mining the pathway names. The process classifies signaling pathways as those that contain the strings "signal" "ERK","MAPK"; metabolic pathways as those that contain "metabol", "catabol", "glyco","pentose"; and regulatory pathways as those that contain "regulat", "upreg", "activ", or "inhibi". These rules result in 323 signaling pathways (e.g., PID Wnt Signaling Pathway, KEGG Toll Like Receptor Signaling Pathway, and Reactome Signaling by Notch), 230 metabolic pathways (e.g., Reactome Purine Catabolism, and KEGG Starch and Sucrose Metabolism), and 397 regulatory pathways (e.g., Reactome Regulation of KIT Signaling, and KEGG Regulation of Autophagy). FIG. 24 shows a Venn Diagram illustrating that very few of these pathways are assigned to multiple classes. For each set of pathways and each of the aforementioned functional networks, a Quack classifier was trained and 30% of the data held out. The process then used each classifier (e.g., the signaling classifier) on the holdout data sets from the other two classes (e.g., the regulatory and metabolic pathway holdout datasets) to assess whether the predictive power of the classifier corresponding to the holdout data was significantly better than the classifiers trained on the others classes of pathways. The Area Under the Curve (AUC) for classifying pathway membership on the holdout data was shown to be significantly greater for the classifiers that correspond to the respective holdout.

[00216] Supplementary Note 8 - Expanding therapeutic hypotheses in 784 GWAS gene sets

[00217] Biological networks are the middle-layer between phenotype and genotype. Armed with the discovery power of Quack, one can set out to expand 784 genome-wide association study (GWAS) gene sets (e.g., genes implicated in Alzheimer's, IBD, and Prion diseases) to identify new pathways from which these phenotypes manifest and provide additional therapeutic hypotheses based on the genes discovered.

[00218] Protein-protein interactions are considered the most interpretable and actionable networks to test therapeutic hypotheses. Therefore, the InWeb network and corresponding trained classifier were used to score and expand the GWAS gene sets. In 98% of these gene sets, Quack was able to identify candidates that physically interact with these proteins to form networks that are derived from learned pathway topological rules. Next, using data on established drug targets, it was found that of the 784 gene sets, 490 (63%) of these included established targets. However, after expanding these gene sets, 578/784 (75%) contained drug targets. In total, the expanded gene sets contained 63% more targets found across the original GWAS gene sets, significantly increasing the potential for therapeutic intervention.

[00219] References

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(February), 860–921. http://doi.org/10.1038/35057062
2. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68 (2011).
3. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–13 (2004).
4. Hein, O., Schwind, M., & König, W. (2006). Scale-free networks. *Wirtschaftsinformatik*, *48*, 267–275. http://doi.org/10.1007/s11576-006-0058-2
5. Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. *Science (New York, N.Y.)*, *325*, 412–413. http://doi.org/10.1126/science.1173299
6. Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks, *393*(June), 440–442.
7. Mitra, K., Carvunis, A.-R., Ramesh, S. K., & Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews. Genetics*, *14*(10), 719–32. http://doi.org/10.1038/nrg3552
8. Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, a W. (1999). From molecular to modular cell biology. *Nature*, *402*(December), C47–C52. http://doi.org/10.1038/35011540
9. Lage,K., Karlberg,E.O., Storling,Z.M., Olason,P.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tumer,Z., Pociot,F., Tommerup,N. et al. (2007) A human phenome-

interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol., 25, 309–316.

10. Lage, K., Tue, N., Karlberg, E. O., Eklund, A. C., Roque, F. S., Donahoe, P. K., … Brunak, S. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes, 1–6.
11. Lundby, A., Rossin, E. J., Steffensen, A. B., Rav Acha, M., Newton-Cheh, C., Pfeufer, A., … Olsen, J. V. (2014). Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics.*Nature Methods*, *11*(8), 868–874. doi:10.1038/nmeth.2997
12. Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., … Daly, M. J. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, *485*(7397), 242–245. http://doi.org/10.1038/nature11011
13. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485: 246–250 doi:10.1038/nature1098922495309
14. Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., … Daly, M. J. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genetics*, *7*(1), e1001273. http://doi.org/10.1371/journal.pgen.1001273
15. Lage, K. (2014). Protein-protein interactions and genetic diseases: The interactome. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1842*(10), 1971–1980. http://doi.org/10.1016/j.bbadis.2014.05.028
16. Goh, K., Cusick, M. E., Valle, D., Childs, B., & Vidal, M. (2007). The human disease network. Proc. Natl. Acad. Sci. USA, 104, 8685–8690.
17. Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci USA 2005;102:15545-15550.
18. Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).
19. Darryl Nishimura. Biotech Software & Internet Report. June 2001, 2(3): 117-120. doi:10.1089/152791601750294344.

20. Carl F. Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay & Kenneth H. Buetow. PID: The Pathway Interaction Database. Nucleic Acids Res.37, D674-9 (2009)
21. Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., ... D'Eustachio, P. (2014). The Reactome pathway knowledgebase. Nucleic Acids Research, 42(Database issue), D472–7. doi:10.1093/nar/gkt1102
22. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10
23. Cowley, G. S. *et al.* Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* 1, 140035 (2014).
24. Li, Y., Calvo, S. E., Gutman, R., Liu, J. S. & Mootha, V. K. Expansion of biological pathways based on evolutionary inference. *Cell* 158, 213–25 (2014).
25. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–35 (2006).
26. Peck, D., Crawford, E. D., Ross, K. N., Stegmaier, K., Golub, T. R., Lamb, J. (2006) A method for high throughput gene expression signature analysis. Genome Biol. 7, R61
27. Lamb, J., (2007) The Connectivity Map: a new tool for biomedical research. Nat. Rev. Cancer 7, 54–60.
28. Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32. doi:10.1023/A:1010933404324.
29. Barzel, B. & Barabási, A.-L. Network link prediction by global silencing of indirect correlations. *Nat. Biotechnol.* 31, 720–5 (2013).
30. Feizi, S., Marbach, D., Médard, M. & Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* 31, 726–33 (2013).
31. Japkowicz, N. (2000). Learning from Imbalanced Data Sets. *Papers from AAAI Workshop*, *21*(9), 10–15. Retrieved from http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-003.pdf

[00220] The teachings of all patents, published applications and references cited herein are incorporated by reference in their entirety.

[00221] While this invention has been particularly shown and described with references to example embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the scope of the invention encompassed by the appended claims.

CLAIMS

What is claimed is:

1. A method for analyzing biological networks, the method comprising:

obtaining data representing biological networks from one or more data stores, the biological networks being defined by respective nodes representing molecules and connections representing relationships between or among the molecules;

obtaining data representing biological pathways, each pathway representing any set of molecules that work in a collaborative way to produce an outcome;

in one or more processors, generating a computational model based on the data representing the biological networks and the data representing the pathways; and

classifying a set of molecules within or related to a given biological network into pathway molecules and non-pathway molecules using the generated model.

2. The method of claim 1, wherein the molecules are genes, the set of molecules is a gene set, and the biological networks are genomic networks.

3. The method of claim 2, further comprising assessing the classified genes to determine at least one of: significance of the gene set, structure of the gene set or relationship of the gene set to a known pathway.

4. The method of claim 3, wherein the classified genes are assessed to determine if the gene set has topological characteristics of a pathway.

5. The method of claim 3, further comprising predicting candidate genes for inclusion in the gene set.

6. The method of claim 2, wherein each pathway includes pathway genes and context genes.

7. The method of claim 6, wherein each pathway gene is a known component of a molecular process.

8. The method of claim 6, wherein each context gene has a first order connection to a pathway gene.

9. The method of claim 6, wherein generating the computational model includes learning respective connectivity profiles across the pathways resulting in learned models.

10. The method of claim 9, wherein each connectivity profile is a topological profile.

11. The method of claim 10, wherein learning the topological profiles includes evaluating connection characteristics of each pathway gene and each context gene.

12. The method of claim 11, wherein evaluating the connection characteristics includes determining a node property of each pathway gene and each context gene.

13. The method of claim 12, wherein the node property includes at least one of a number of connections, a weighted degree, an eigenvector centrality, a betweeness centrality, a closeness centrality, and a local clustering coefficient.

14. The method of claim 13, further including grouping genes based on their connectivity.

15. The method of claim 14, wherein grouping genes includes grouping the genes based on predicted probabilities from the generated computational model.

16. The method of claim 12, wherein each node property is determined in a given pathway and in a respective network.

17. The method of claim 9, wherein generating the computational model further includes building a model data set by stacking the connectivity profiles.

18. The method of claim 17, wherein generating the computational model includes using machine learning techniques by performing a random forests analysis on the model data set.

19. The method of claim 2, wherein the relationships between or among the genes are based on at least one of: physical interaction, similar global transcriptional response, co-dependencies in cancer cell lines, and correlation in gene expression.

20. The method of any one of claims 1 to 19, further comprising adjusting a respective pathway as a function of the classifying.

21. The method of any one of claims 1 to 19, wherein obtaining the data representing biological networks includes allowing a user to select the data.

22. The method of any one of claims 1 to 19, wherein obtaining the data representing the biological pathways includes allowing a user to input the data.

23. The method of any one of claims 1 to 19, further comprising storing the generated computational model to a model store.

24. The method of claim 23, wherein the set of molecules is classified using the generated computational model or a computational model retrieved from the model store in response to a user selection.

25. A system for analyzing biological networks, the system including elements configured to perform the method of any one of the preceding claims.

26. A system for analyzing biological networks, the system comprising:

a network module configured to obtain data representing biological networks from one or more data stores, the biological networks being defined by respective nodes representing molecules and connections representing relationships between or among the molecules;

a pathway module configured to obtain data representing biological pathways, each pathway representing any set of molecules that work in a collaborative way to produce an outcome;

one or more processors configured to generate a computational model based on the data representing the biological networks and the data representing the pathways; and

a classifier module configured to classify a set of molecules within or related to a given biological network into pathway molecules and non-pathway molecules using the generated model.

27. The system of claim 26, wherein the molecules are genes, the set of molecules is a gene set, and the biological networks are genomic networks.

28. The system of claim 27, further comprising an assessment module configured to assess the classified genes to determine at least one of: significance of the gene set, structure of the gene set or relationship of the gene set to a known pathway.

29. The system of claim 28, wherein the classified genes are assessed to determine if the gene set has topological characteristics of a pathway or if the result of the classifying suggests a functional implication or association to a known pathway.

30. The system of claim 29, further comprising a predictor configured to predict candidate genes for inclusion in the gene set.

31. The system of claim 27, wherein each pathway includes pathway genes and context genes.

32. The system of claim 31, wherein each pathway gene is a known component of a molecular process.

33. The system of claim 31, wherein each context gene has a first order connection to a pathway gene.

34. The system of claim 31, wherein generating the computational model includes learning respective connectivity profiles across the pathways resulting in learned models.

35. The system of claim 34, wherein each connectivity profile is a topological profile.

36. The system of claim 35, wherein learning the topological profiles includes evaluating connection characteristics of each pathway gene and each context gene.

37. The system of claim 36, wherein evaluating the connection characteristics includes determining a node property of each pathway gene and each context gene.

38. The system of claim 37, wherein the node property includes at least one of a number of connections, a weighted degree, an eigenvector centrality, a betweeness centrality, a closeness centrality, and a local clustering coefficient.

39. The system of claim 38, wherein the one or more processors are further configured to group genes based on their connectivity.

40. The system of claim 39, wherein the genes are grouped based on predicted probabilities from the generated computational model.

41. The system of claim 37, wherein each node property is determined in a given pathway and in a respective network.

42. The system of claim 34, wherein generating the computational model further includes building a model data set by stacking the learned connectivity profiles.

43. The system of claim 42, wherein generating the computational model includes using machine learning techniques by performing a random forests analysis on the model data set.

44. The system of claim 27, wherein the relationships between or among the genes are based on at least one of: physical interaction, similar global transcriptional response, co-dependencies in cancer cell lines, and correlation in gene expression.

45. The system of any one of claims 26 to 44, further comprising an adjustment module to adjust a respective pathway as a function of the classifying.

46. The system of any one of claims 26 to 44, wherein the network module is configured to obtain the data representing biological networks by allowing a user to select the data.

47. The system of any one of claims 26 to 44, wherein the pathway module is configured to obtain the data representing the biological pathways by allowing a user to input the data.

48. The system of any one of claims 26 to 44, further comprising a model store, and wherein the one or more processors are configured to store the generated computational model to the model store.

49. The system of claim 48, wherein the set of molecules is classified using the generated computational model or a computational model retrieved from the model store in response to a user selection.

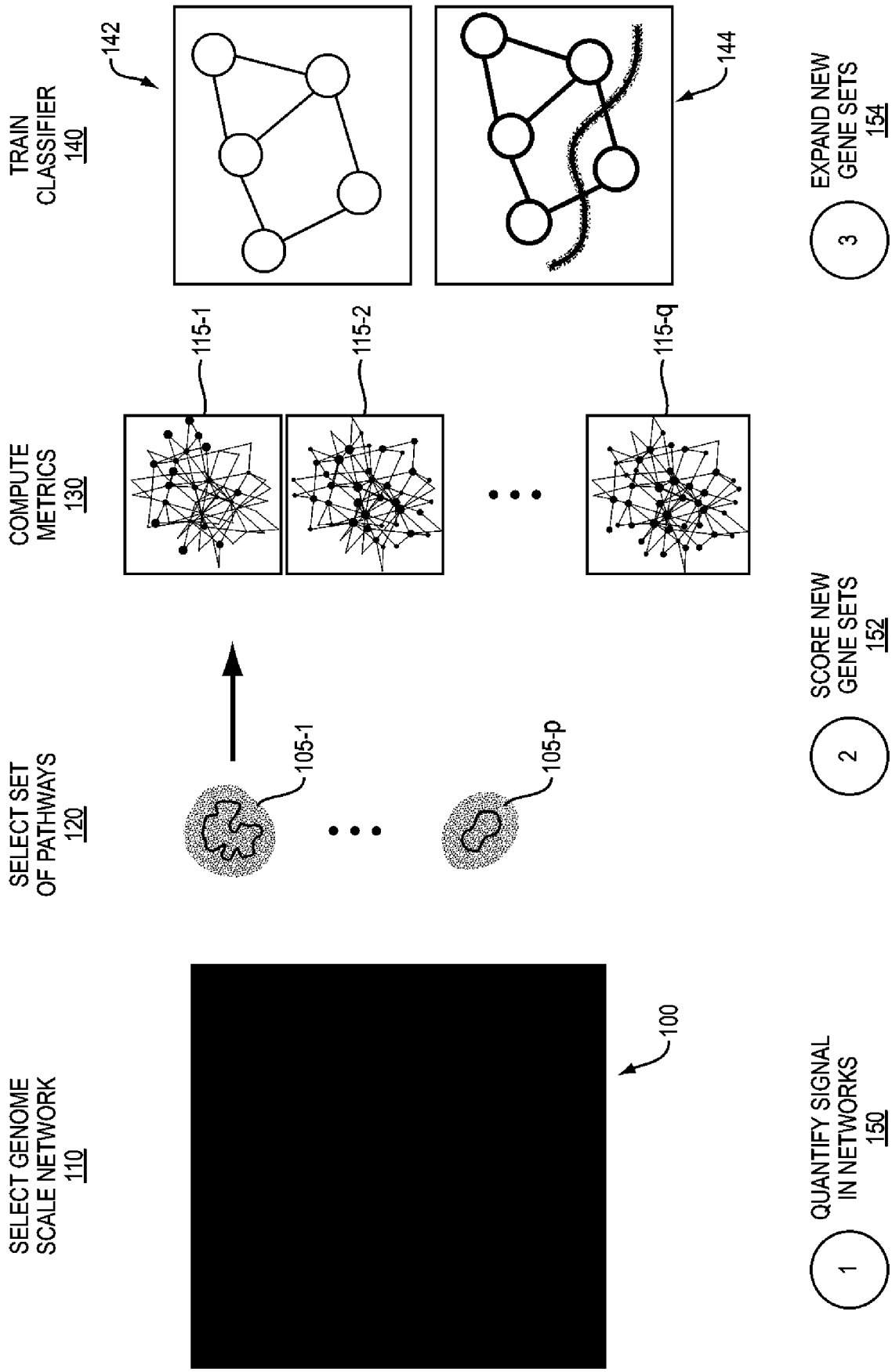

SELECT GENOME SCALE NETWORK 110
100
SELECT SET OF PATHWAYS 120
105-1
105-p
COMPUTE METRICS 130
115-1
115-2
115-q
TRAIN CLASSIFIER 140
142
144
1 QUANTIFY SIGNAL IN NETWORKS 150
2 SCORE NEW GENE SETS 152
3 EXPAND NEW GENE SETS 154

FIG. 1

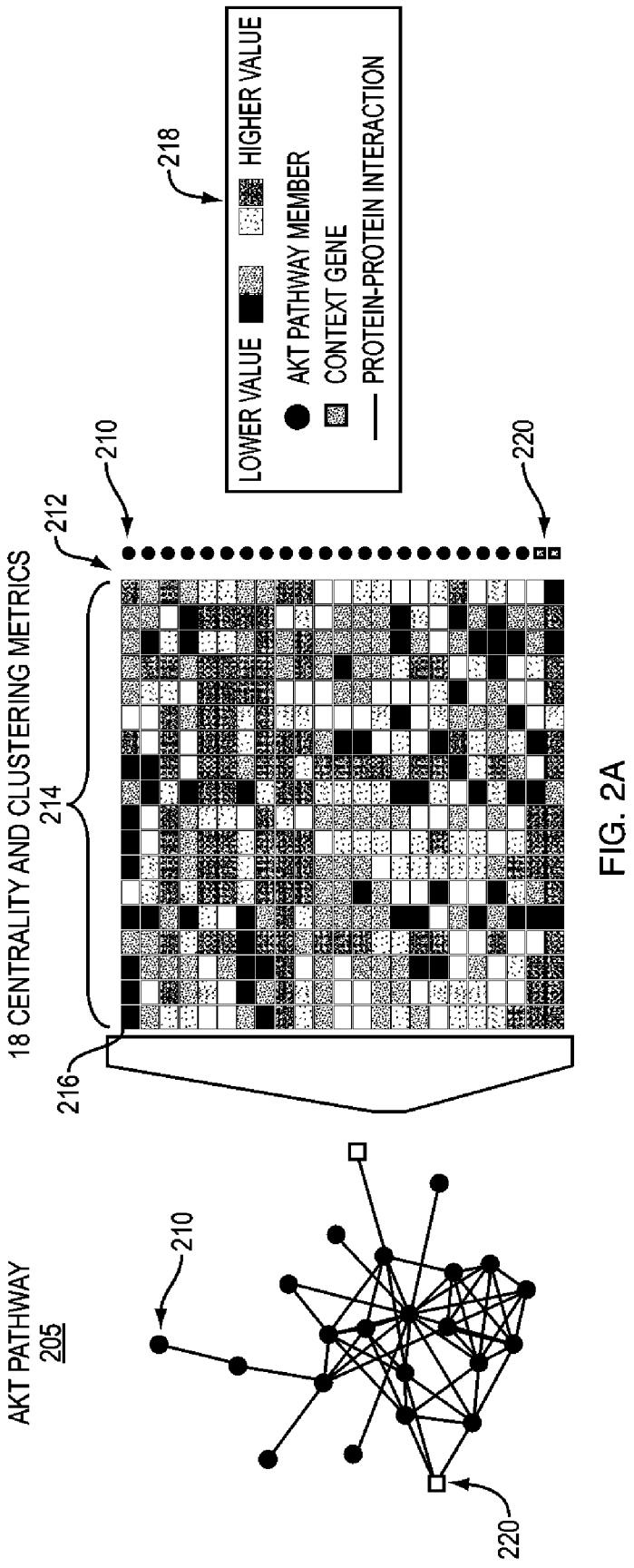

AKT PATHWAY
205
210
220
18 CENTRALITY AND CLUSTERING METRICS
214
212
210
216
218
LOWER VALUE
HIGHER VALUE
AKT PATHWAY MEMBER
CONTEXT GENE
PROTEIN-PROTEIN INTERACTION
220

FIG. 2A

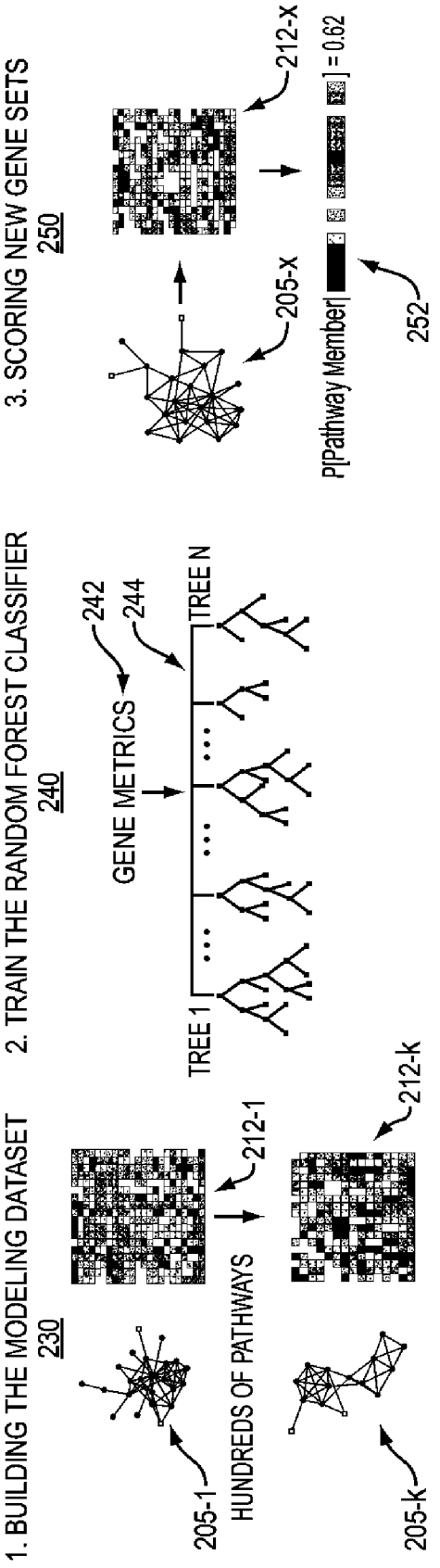

1. BUILDING THE MODELING DATASET
230
HUNDREDS OF PATHWAYS
205-1
212-1
205-k
212-k
2. TRAIN THE RANDOM FOREST CLASSIFIER
240
GENE METRICS
242
244
TREE 1
TREE N
3. SCORING NEW GENE SETS
250
205-x
212-x
P[Pathway Member] = 0.62
252

FIG. 2B

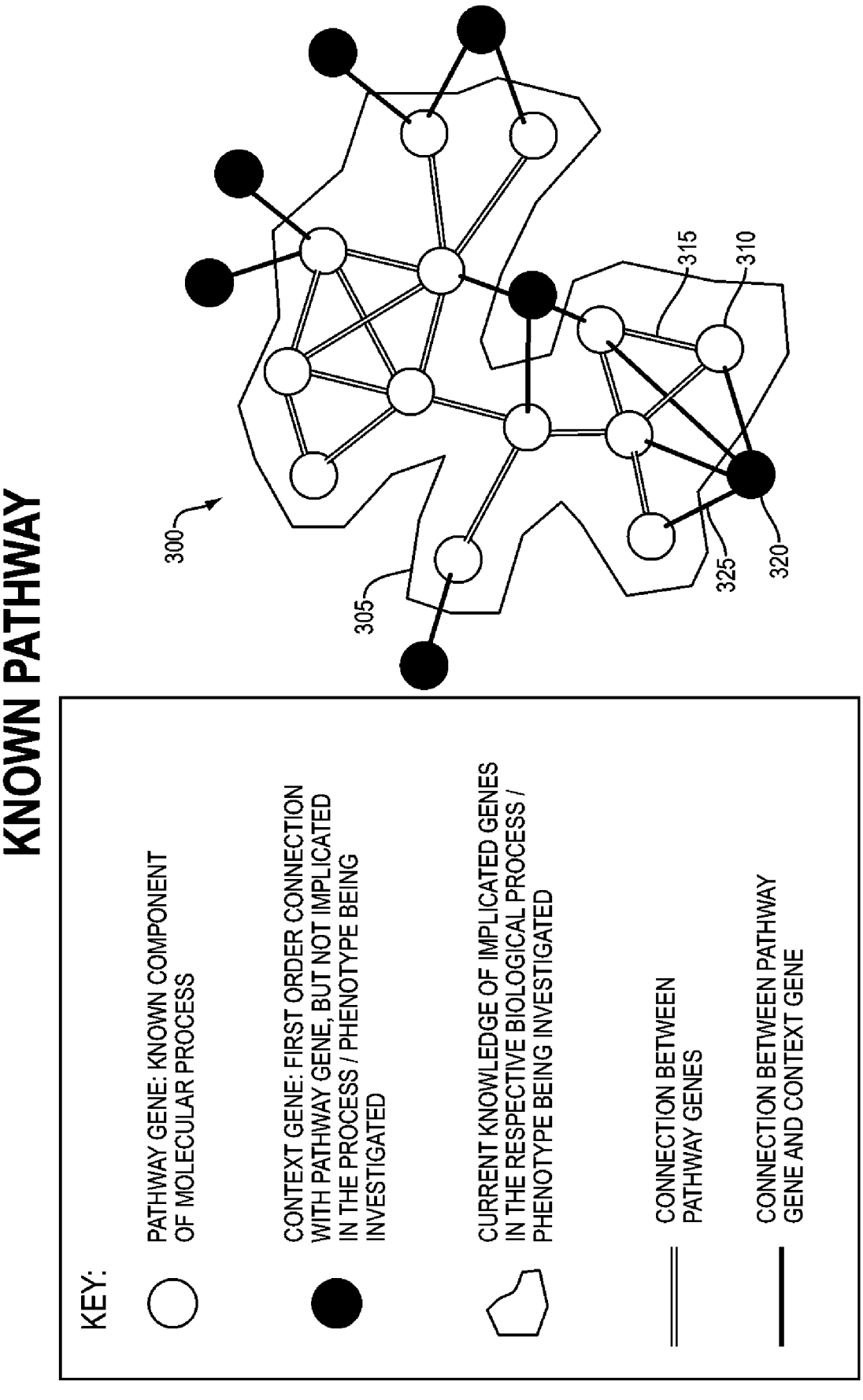

KNOWN PATHWAY
KEY:
PATHWAY GENE: KNOWN COMPONENT OF MOLECULAR PROCESS
CONTEXT GENE: FIRST ORDER CONNECTION WITH PATHWAY GENE, BUT NOT IMPLICATED IN THE PROCESS / PHENOTYPE BEING INVESTIGATED
CURRENT KNOWLEDGE OF IMPLICATED GENES IN THE RESPECTIVE BIOLOGICAL PROCESS / PHENOTYPE BEING INVESTIGATED
CONNECTION BETWEEN PATHWAY GENES
CONNECTION BETWEEN PATHWAY GENE AND CONTEXT GENE
300
305
315
310
325
320

FIG. 3

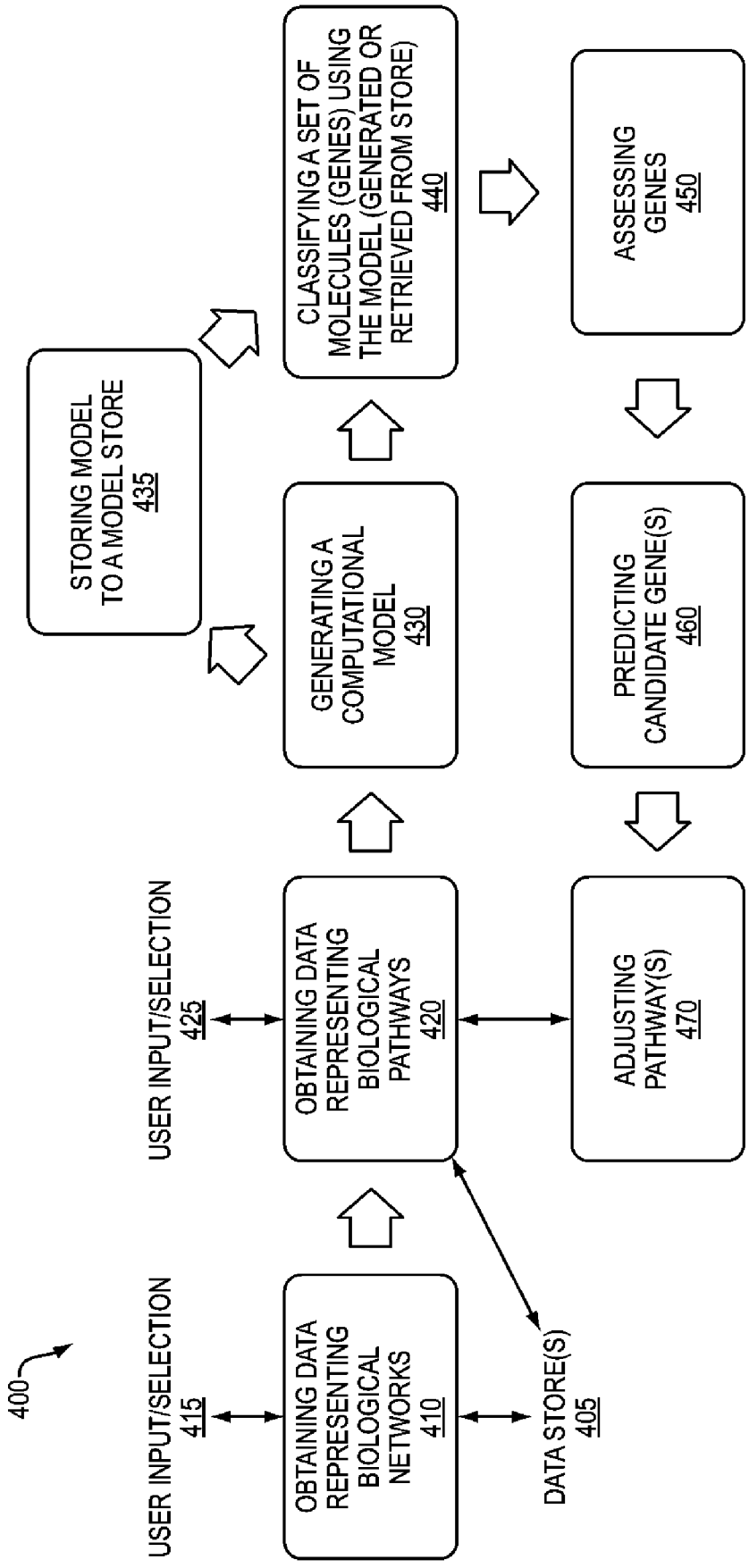

400
USER INPUT/SELECTION 415
OBTAINING DATA REPRESENTING BIOLOGICAL NETWORKS 410
DATA STORE(S) 405
USER INPUT/SELECTION 425
OBTAINING DATA REPRESENTING BIOLOGICAL PATHWAYS 420
ADJUSTING PATHWAY(S) 470
GENERATING A COMPUTATIONAL MODEL 430
PREDICTING CANDIDATE GENE(S) 460
STORING MODEL TO A MODEL STORE 435
CLASSIFYING A SET OF MOLECULES (GENES) USING THE MODEL (GENERATED OR RETRIEVED FROM STORE) 440
ASSESSING GENES 450

FIG. 4

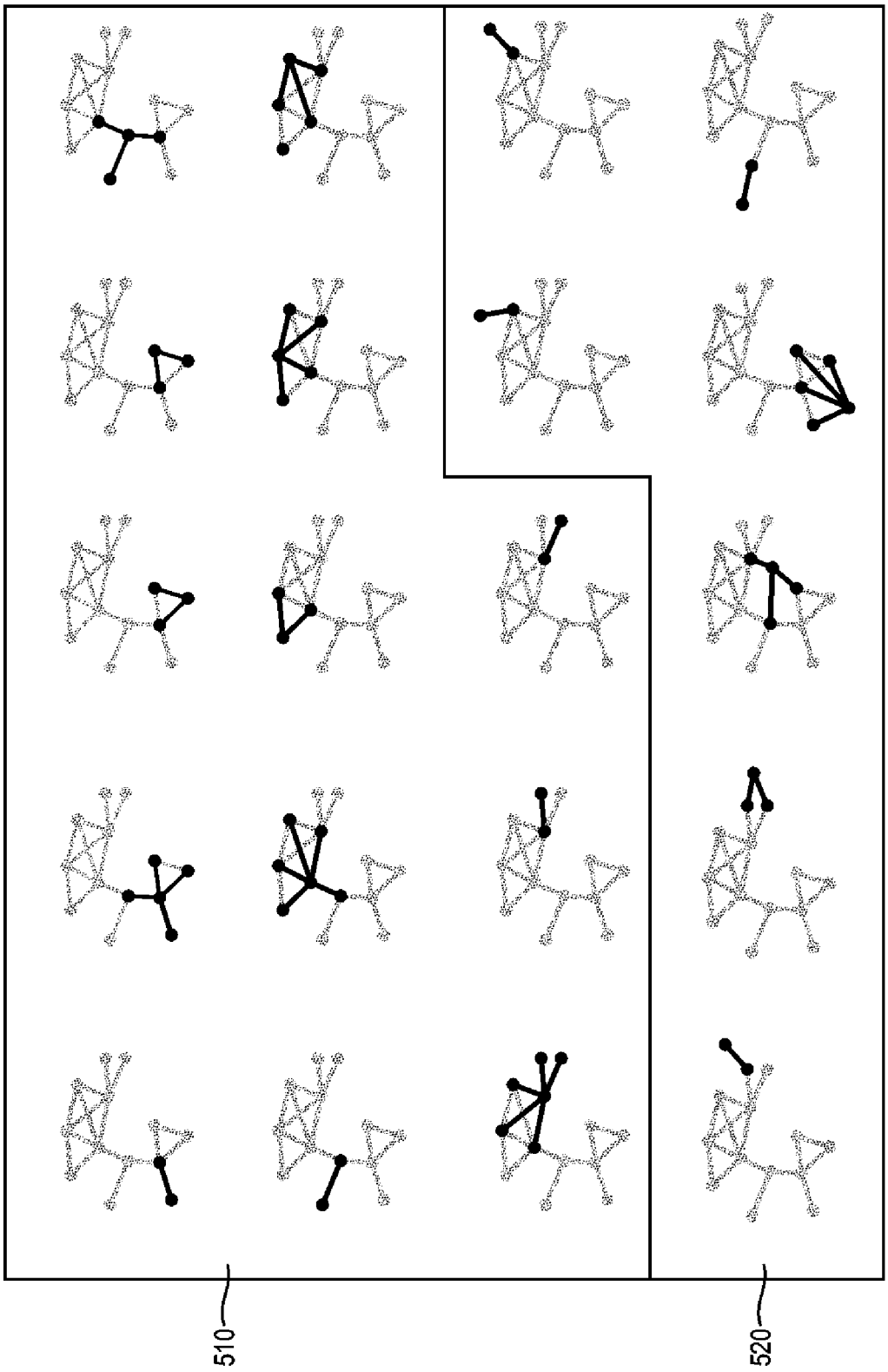


FIG. 5

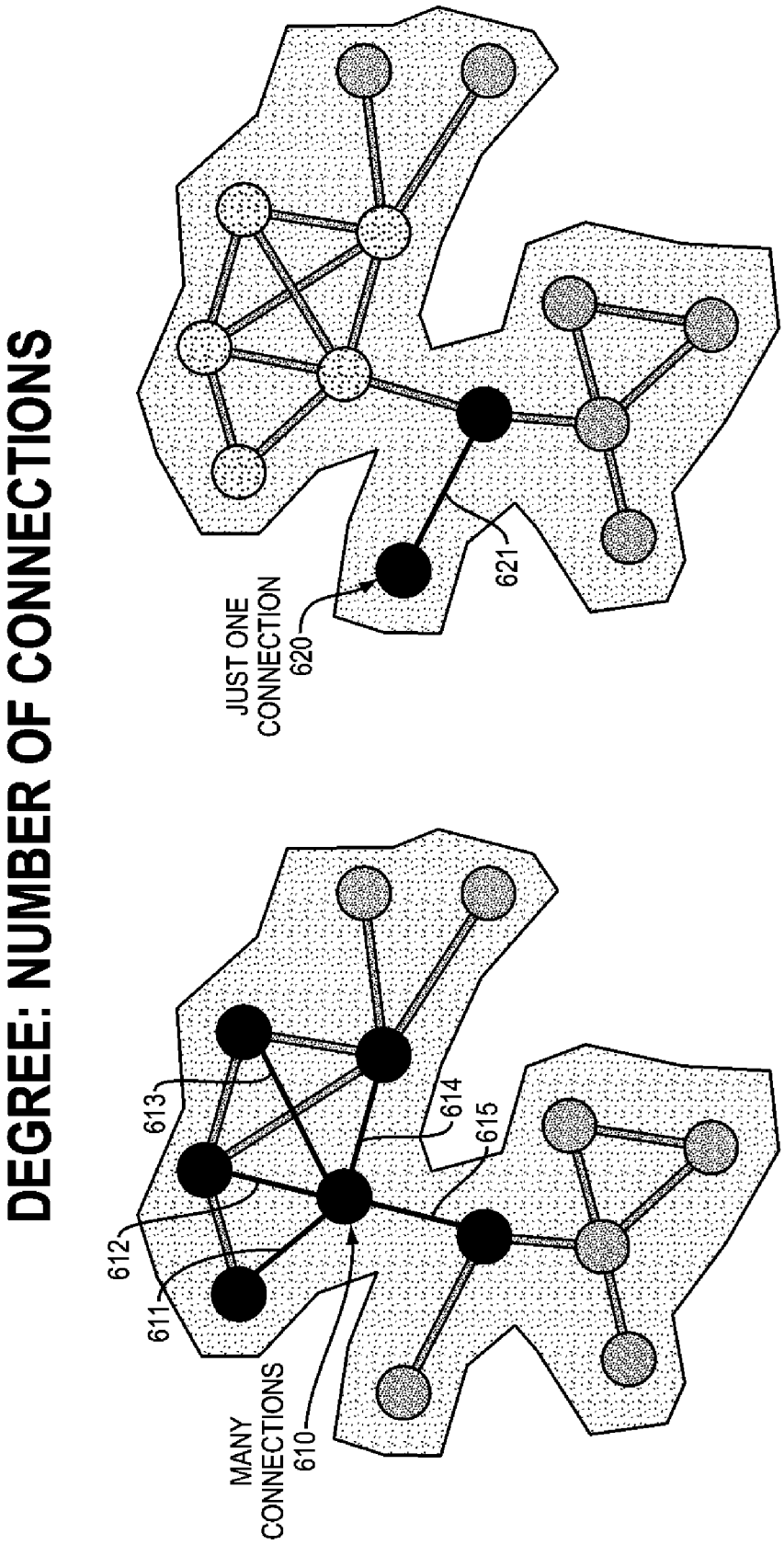

DEGREE: NUMBER OF CONNECTIONS
JUST ONE CONNECTION
620
621
MANY CONNECTIONS
610
611
612
613
614
615

FIG. 6

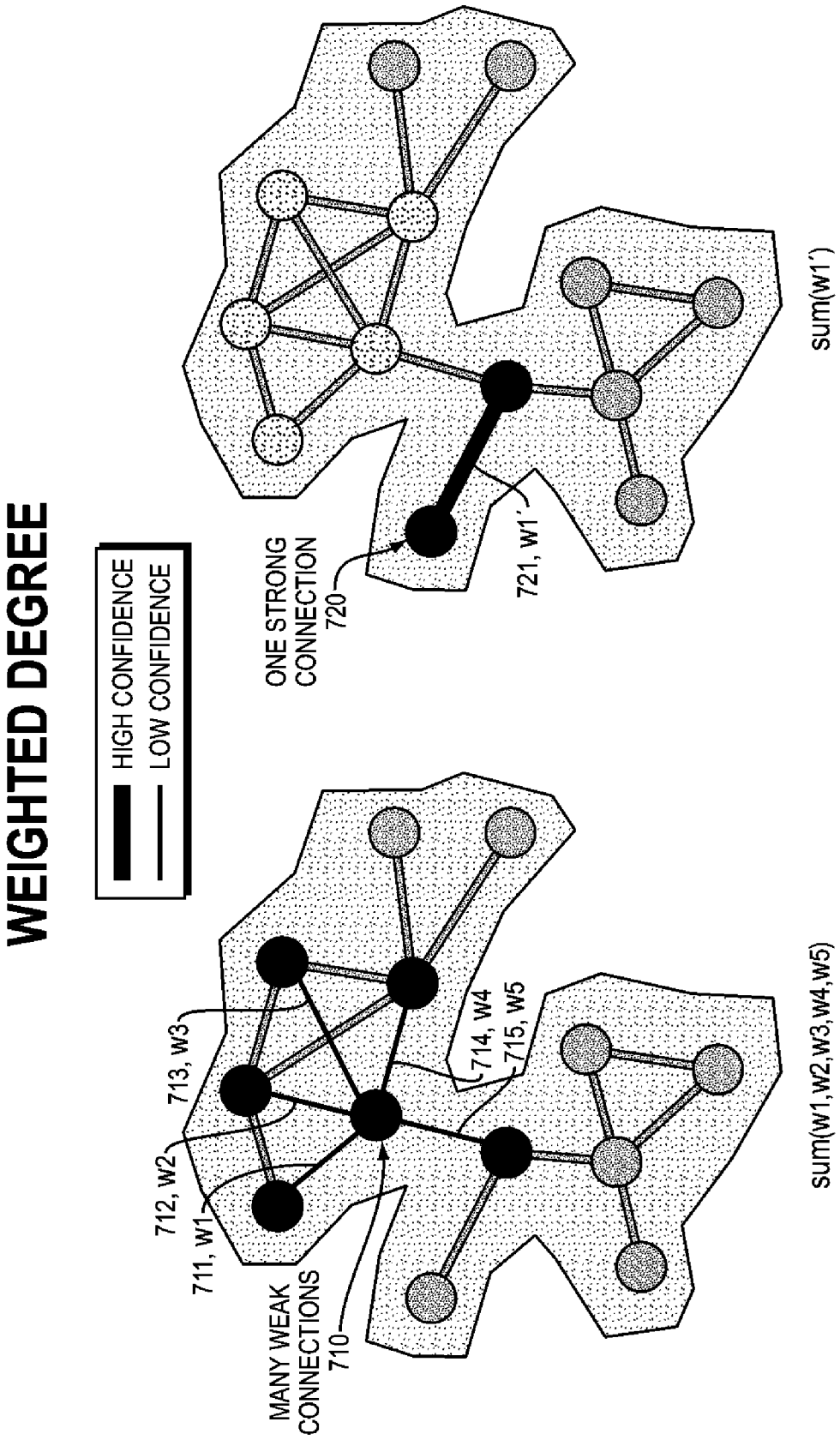

WEIGHTED DEGREE
HIGH CONFIDENCE
LOW CONFIDENCE
ONE STRONG CONNECTION
720
721, w1'
sum(w1')
MANY WEAK CONNECTIONS
710
711, w1
712, w2
713, w3
714, w4
715, w5
sum(w1,w2,w3,w4,w5)

FIG. 7

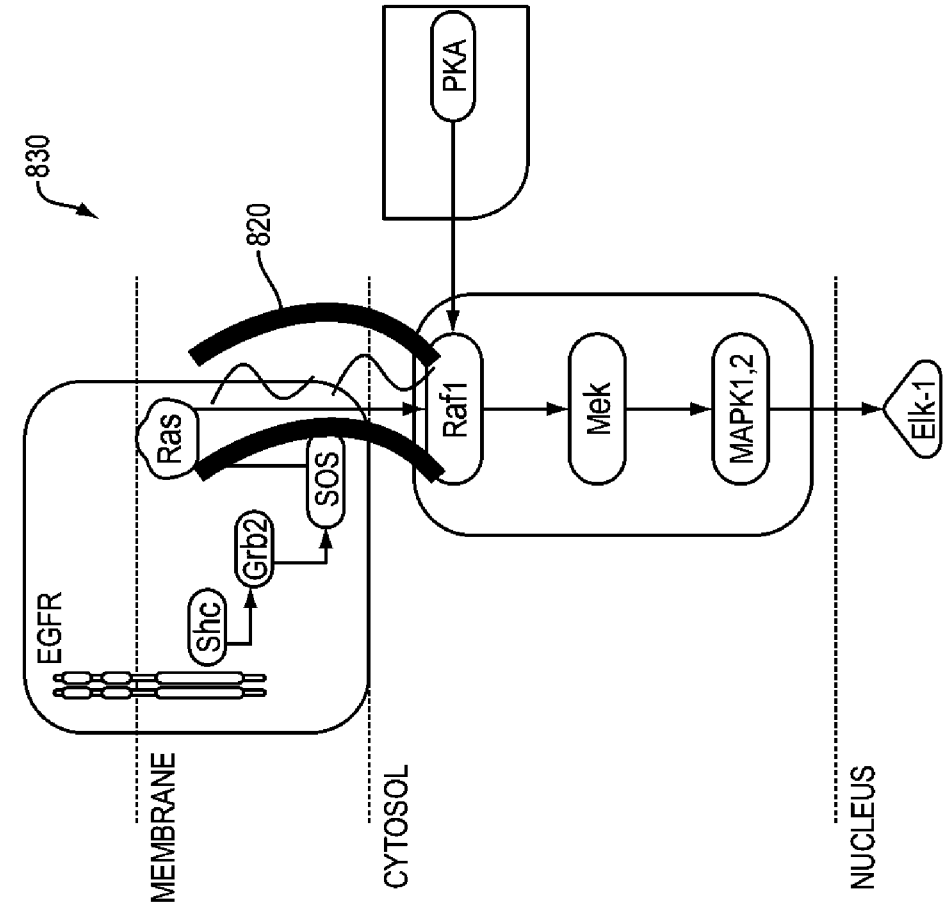

830
820
PKA
Ras
SOS
Grb2
Shc
EGFR
Raf1
Mek
MAPK1,2
Elk-1
MEMBRANE
CYTOSOL
NUCLEUS

FIG. 8B

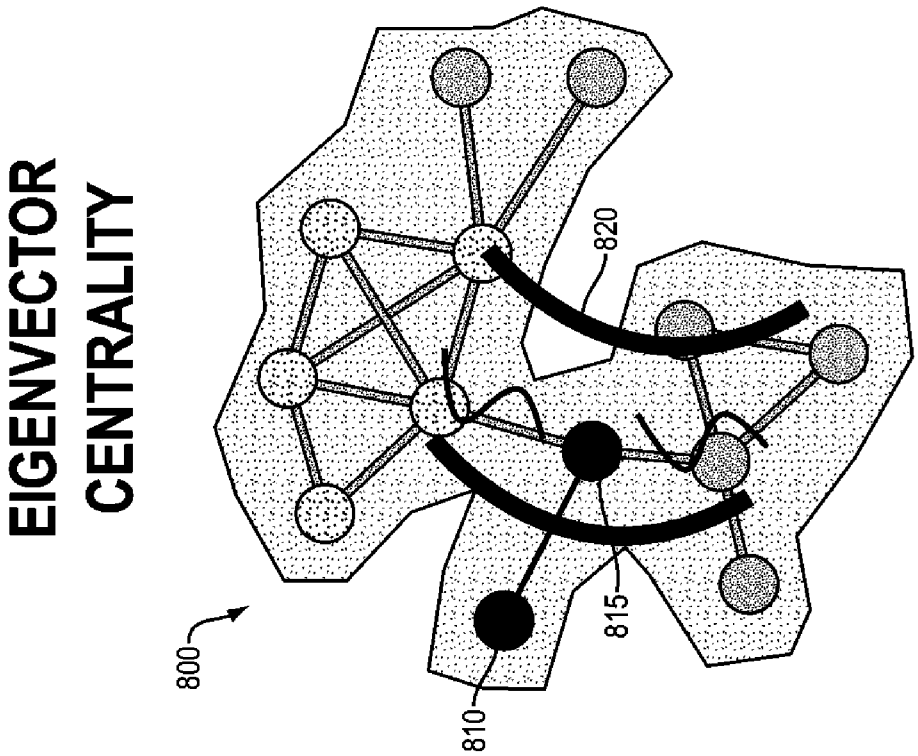

EIGENVECTOR CENTRALITY
800
810
815
820

FIG. 8A

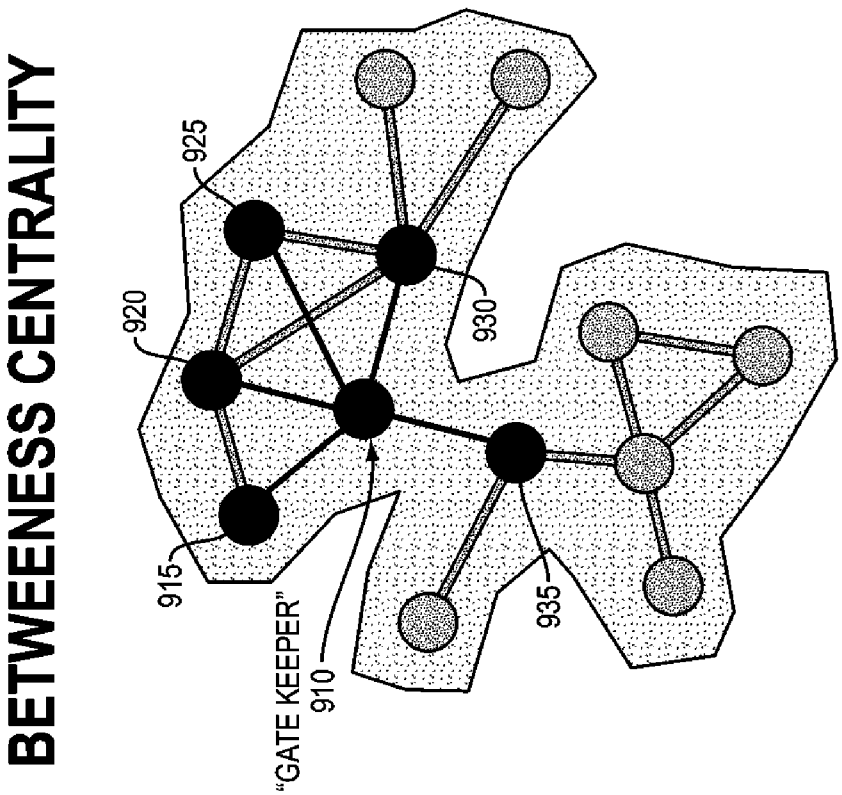


FIG. 9

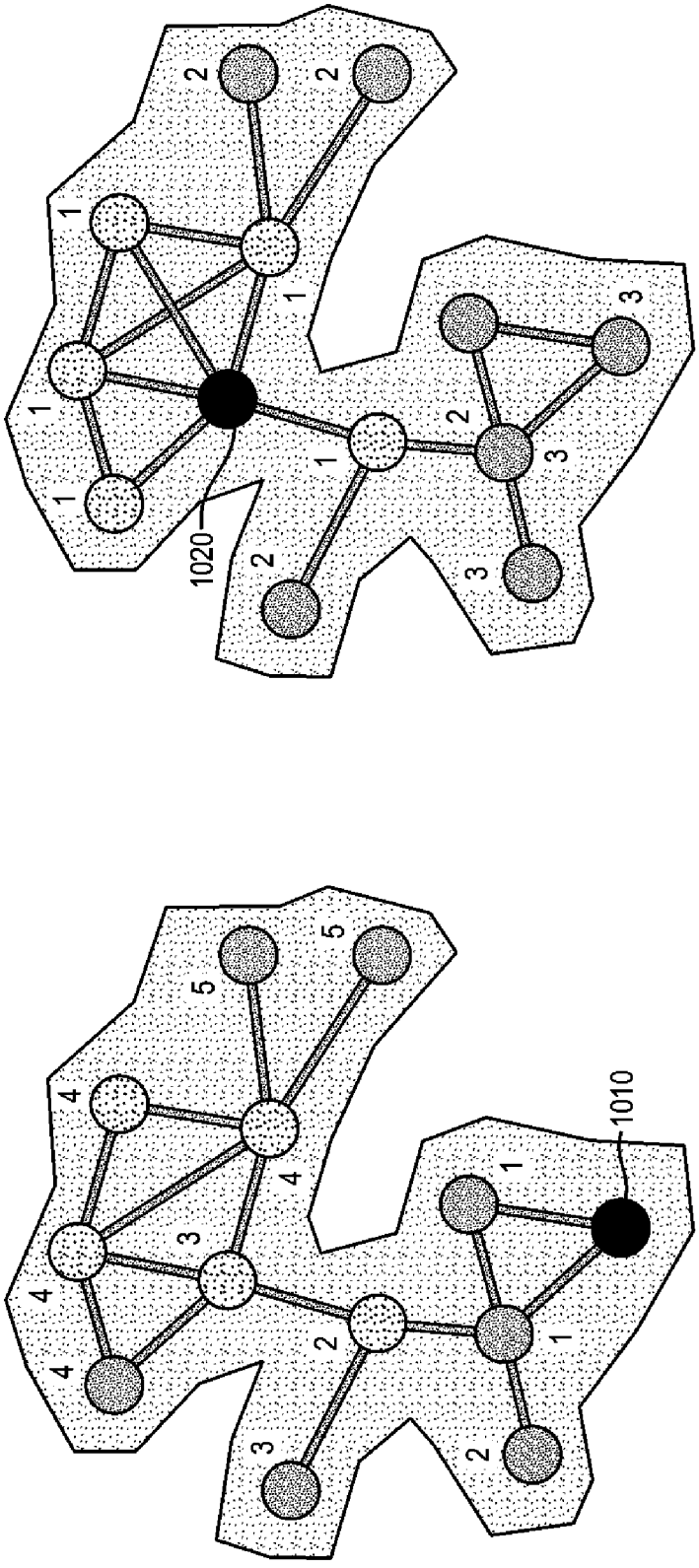

CLOSENESS CENTRALITY
1020
1010
INVERSE OF THE AVERAGE
MIN-PATH LENGTH

FIG. 10

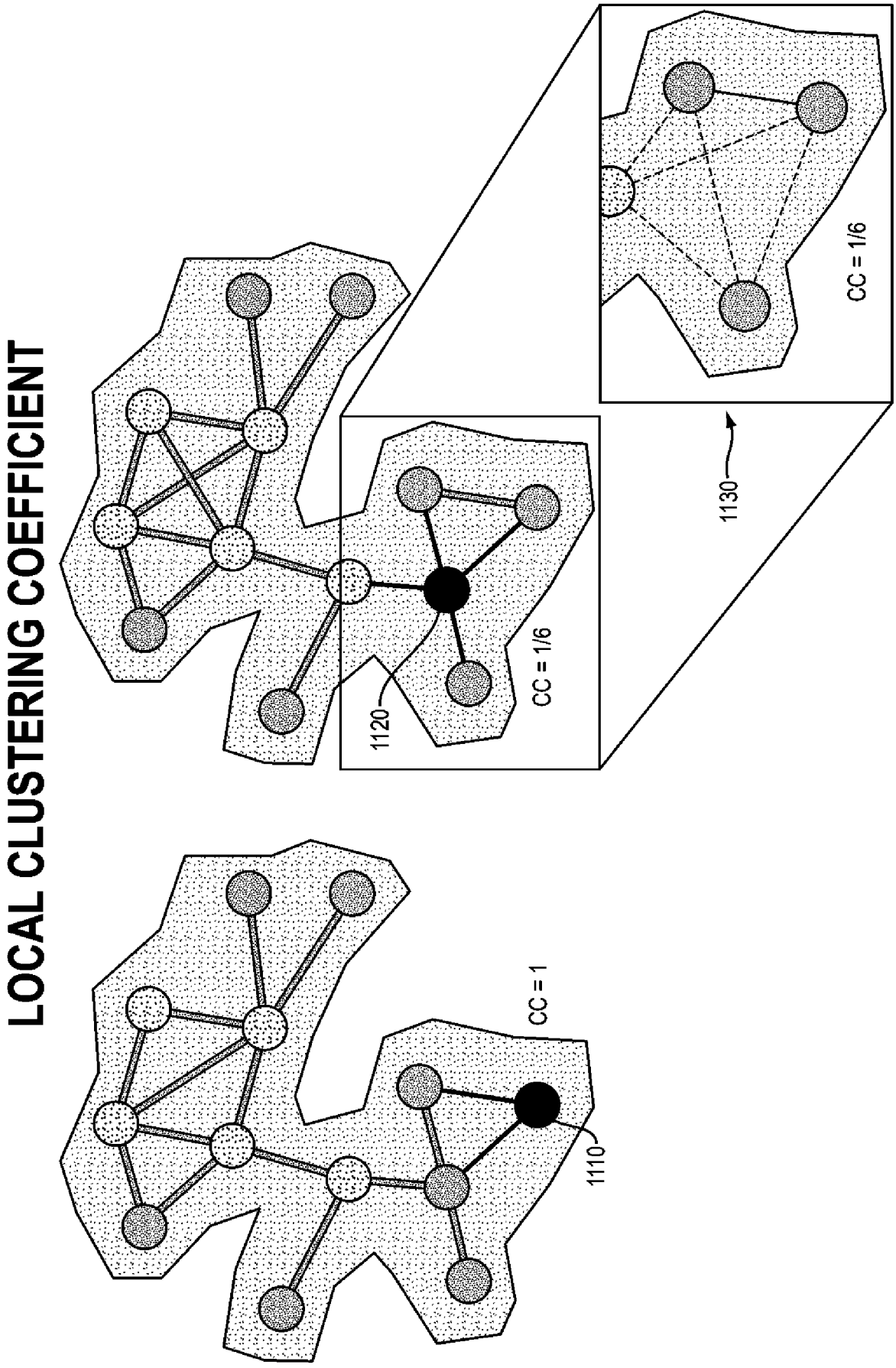


FIG. 11

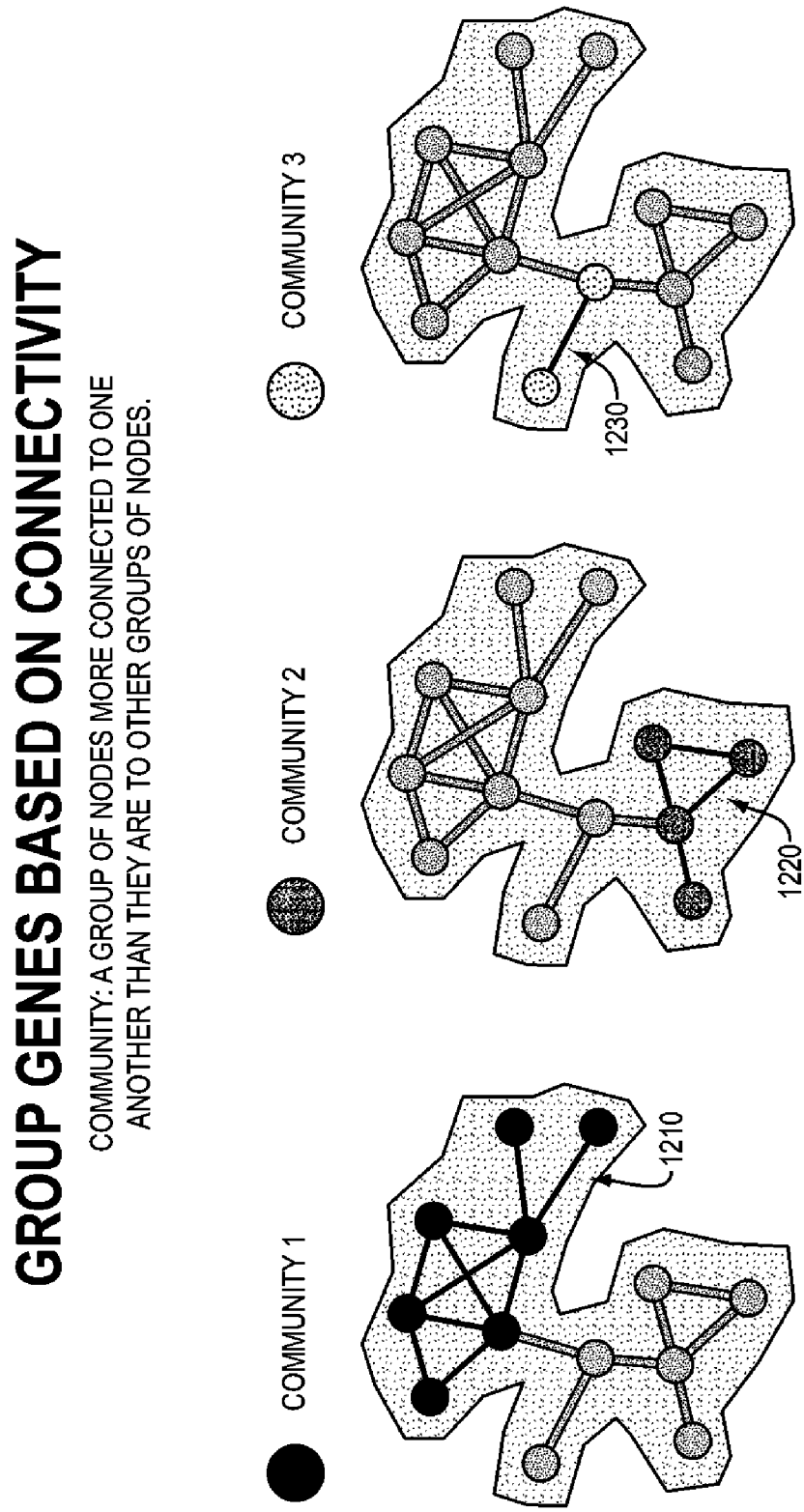

GROUP GENES BASED ON CONNECTIVITY
COMMUNITY: A GROUP OF NODES MORE CONNECTED TO ONE ANOTHER THAN THEY ARE TO OTHER GROUPS OF NODES.
COMMUNITY 1
COMMUNITY 2
COMMUNITY 3
1210
1220
1230

FIG. 12

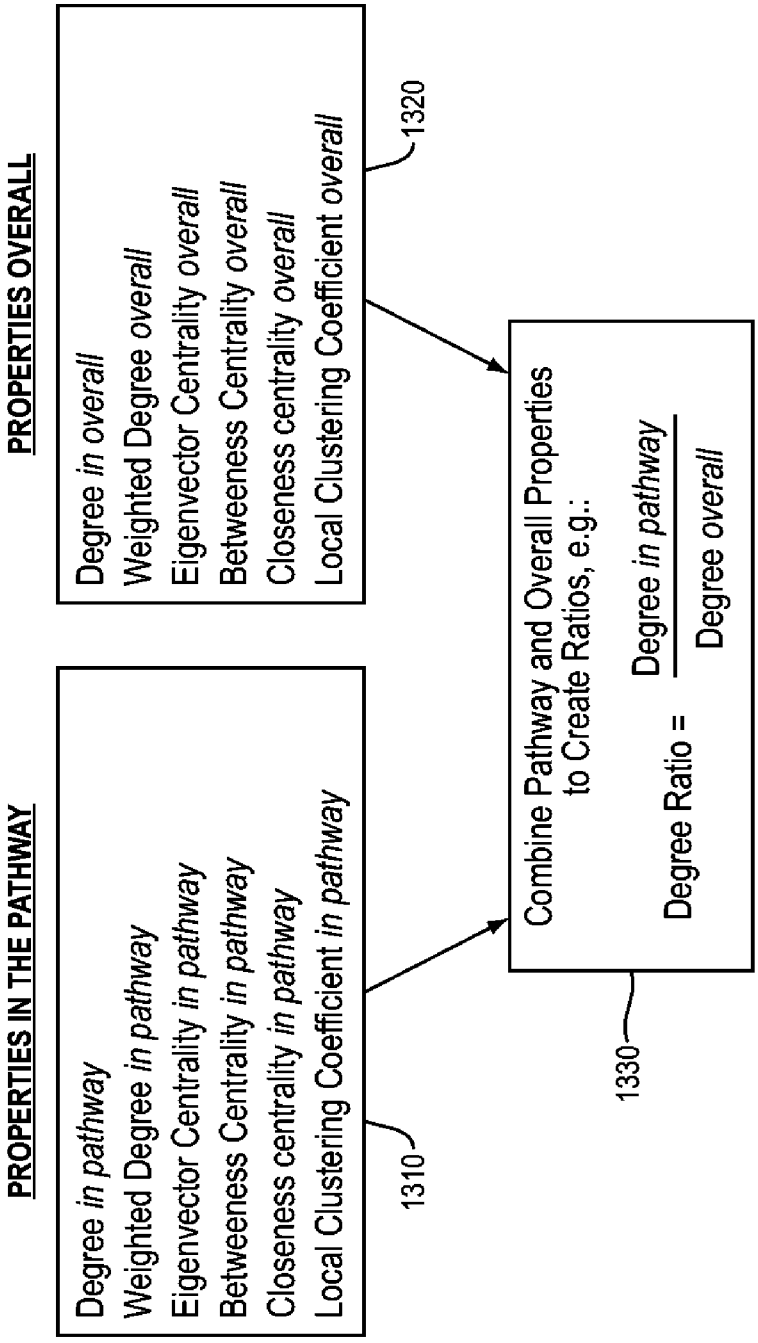

PROPERTIES IN THE PATHWAY
Degree in pathway
Weighted Degree in pathway
Eigenvector Centrality in pathway
Betweeness Centrality in pathway
Closeness centrality in pathway
Local Clustering Coefficient in pathway
1310
PROPERTIES OVERALL
Degree in overall
Weighted Degree overall
Eigenvector Centrality overall
Betweeness Centrality overall
Closeness centrality overall
Local Clustering Coefficient overall
1320
Combine Pathway and Overall Properties to Create Ratios, e.g.:
Degree Ratio = Degree in pathway / Degree overall
1330

FIG. 13

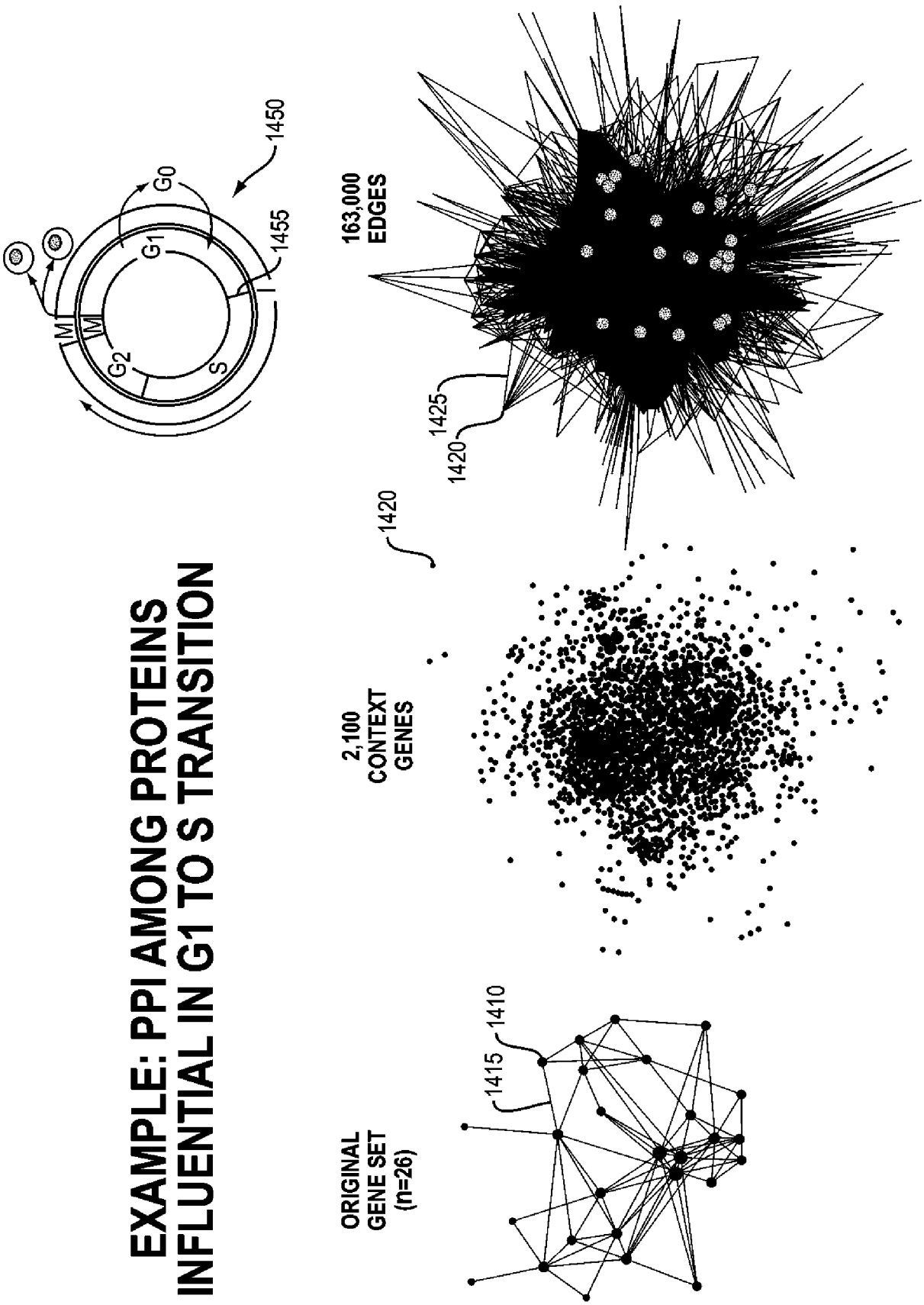
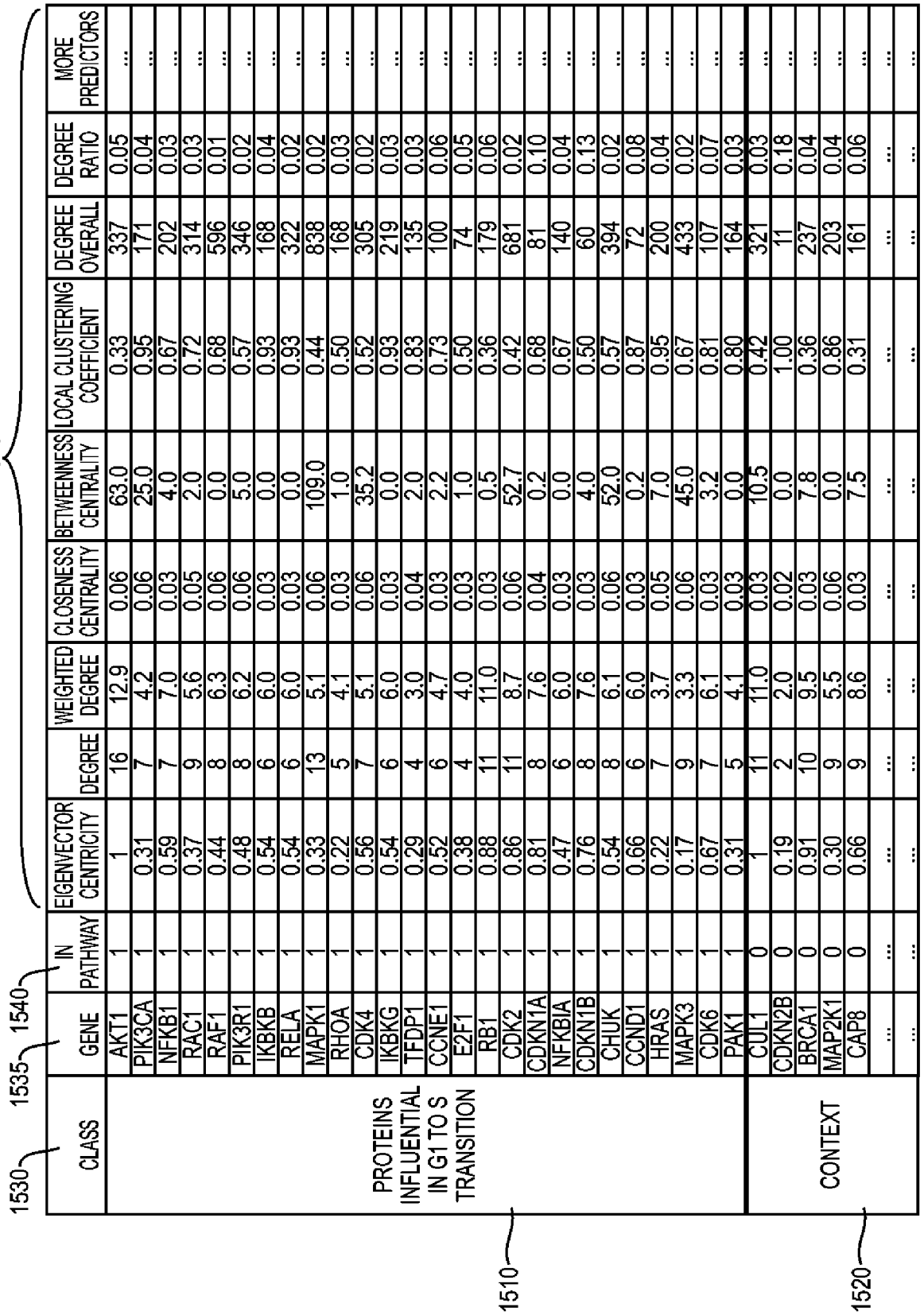

EXAMPLE: PPI AMONG PROTEINS INFLUENTIAL IN G1 TO S TRANSITION
1450
1455
G0
G1
S
G2
M
M
163,000 EDGES
1425
1420
1420
2,100 CONTEXT GENES
ORIGINAL GENE SET (n=26)
1415
1410

FIG. 14



CLASS	GENE	IN PATHWAY	EIGENVECTOR CENTRICITY	DEGREE	WEIGHTED DEGREE	CLOSENESS CENTRALITY	BETWEENNESS CENTRALITY	LOCAL CLUSTERING COEFFICIENT	DEGREE OVERALL	DEGREE RATIO	MORE PREDICTORS
PROTEINS INFLUENTIAL IN G1 TO S TRANSITION	AKT1	1	1	16	12.9	0.06	63.0	0.33	337	0.05	...
	PIK3CA	1	0.31	7	4.2	0.06	25.0	0.95	171	0.04	...
	NFKB1	1	0.59	7	7.0	0.03	4.0	0.67	202	0.03	...
	RAC1	1	0.37	9	5.6	0.05	2.0	0.72	314	0.03	...
	RAF1	1	0.44	8	6.3	0.06	0.0	0.68	596	0.01	...
	PIK3R1	1	0.48	8	6.2	0.06	5.0	0.57	346	0.02	...
	IKBKB	1	0.54	6	6.0	0.03	0.0	0.93	168	0.04	...
	RELA	1	0.54	6	6.0	0.03	0.0	0.93	322	0.02	...
	MAPK1	1	0.33	13	5.1	0.06	109.0	0.44	838	0.02	...
	RHOA	1	0.22	5	4.1	0.03	1.0	0.50	168	0.03	...
	CDK4	1	0.56	7	5.1	0.06	35.2	0.52	305	0.02	...
	IKBKG	1	0.54	6	6.0	0.03	0.0	0.93	219	0.03	...
	TFDP1	1	0.29	4	3.0	0.04	2.0	0.83	135	0.03	...
	CCNE1	1	0.52	6	4.7	0.03	2.2	0.73	100	0.06	...
	E2F1	1	0.38	4	4.0	0.03	1.0	0.50	74	0.05	...
	RB1	1	0.88	11	11.0	0.03	0.5	0.36	179	0.06	...
	CDK2	1	0.86	11	8.7	0.06	52.7	0.42	681	0.02	...
	CDKN1A	1	0.81	8	7.6	0.04	0.2	0.68	81	0.10	...
	NFKBIA	1	0.47	6	6.0	0.03	0.0	0.67	140	0.04	...
	CDKN1B	1	0.76	8	7.6	0.03	4.0	0.50	60	0.13	...
	CHUK	1	0.54	8	6.1	0.06	52.0	0.57	394	0.02	...
	CCND1	1	0.66	6	6.0	0.03	0.2	0.87	72	0.08	...
	HRAS	1	0.22	7	3.7	0.05	7.0	0.95	200	0.04	...
	MAPK3	1	0.17	9	3.3	0.06	45.0	0.67	433	0.02	...
	CDK6	1	0.67	7	6.1	0.03	3.2	0.81	107	0.07	...
	PAK1	1	0.31	5	4.1	0.03	0.0	0.80	164	0.03	...
CONTEXT	CUL1	0	1	11	11.0	0.03	10.5	0.42	321	0.03	...
	CDKN2B	0	0.19	2	2.0	0.02	0.0	1.00	11	0.18	...
	BRCA1	0	0.91	10	9.5	0.03	7.8	0.36	237	0.04	...
	MAP2K1	0	0.30	9	5.5	0.06	0.0	0.86	203	0.04	...
	CAP8	0	0.66	9	8.6	0.03	7.5	0.31	161	0.06	...
	...	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...	...

FIG. 15

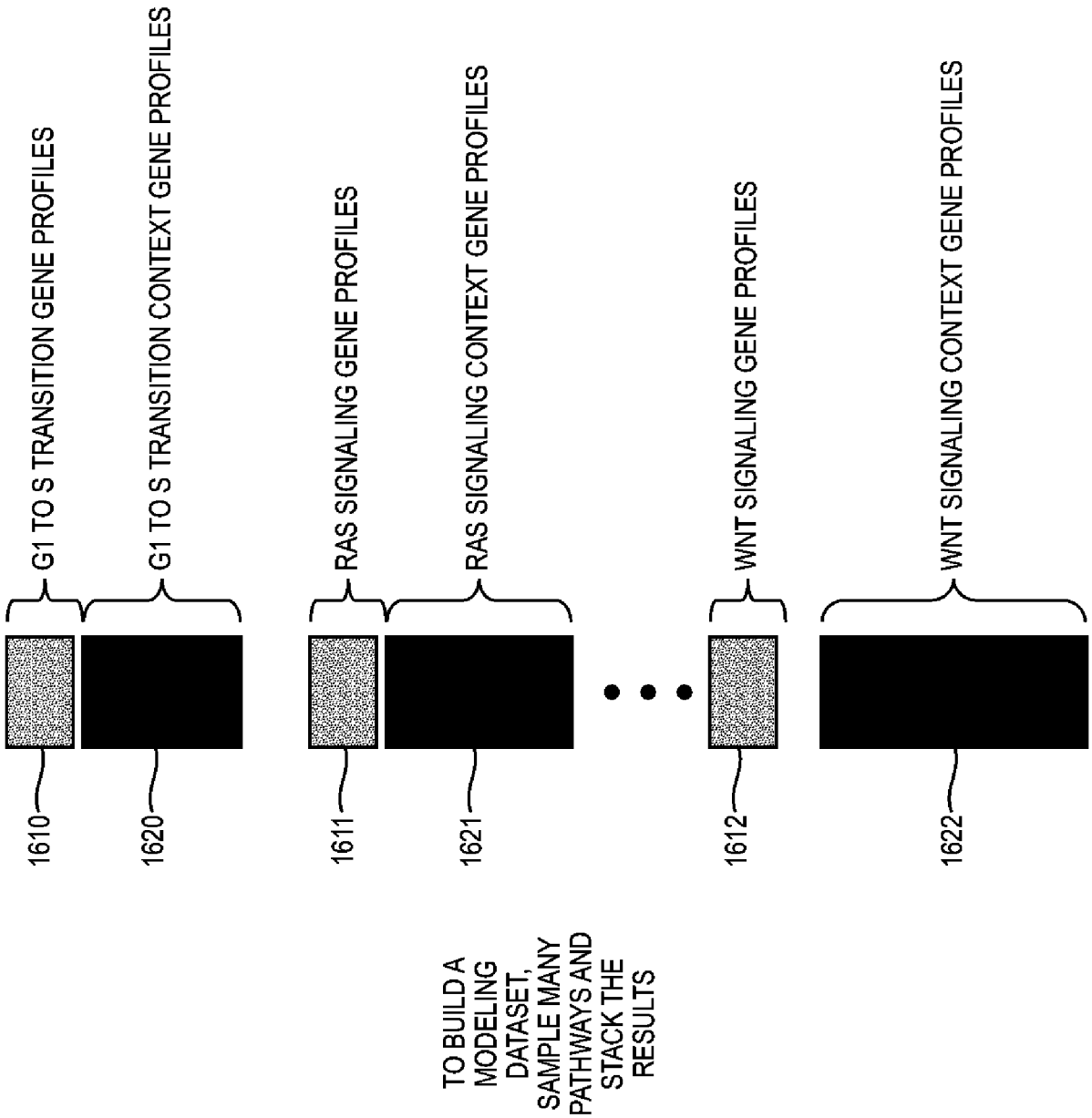

G1 TO S TRANSITION GENE PROFILES
G1 TO S TRANSITION CONTEXT GENE PROFILES
RAS SIGNALING GENE PROFILES
RAS SIGNALING CONTEXT GENE PROFILES
WNT SIGNALING GENE PROFILES
WNT SIGNALING CONTEXT GENE PROFILES
1610
1620
1611
1621
1612
1622
TO BUILD A MODELING DATASET, SAMPLE MANY PATHWAYS AND STACK THE RESULTS

FIG. 16

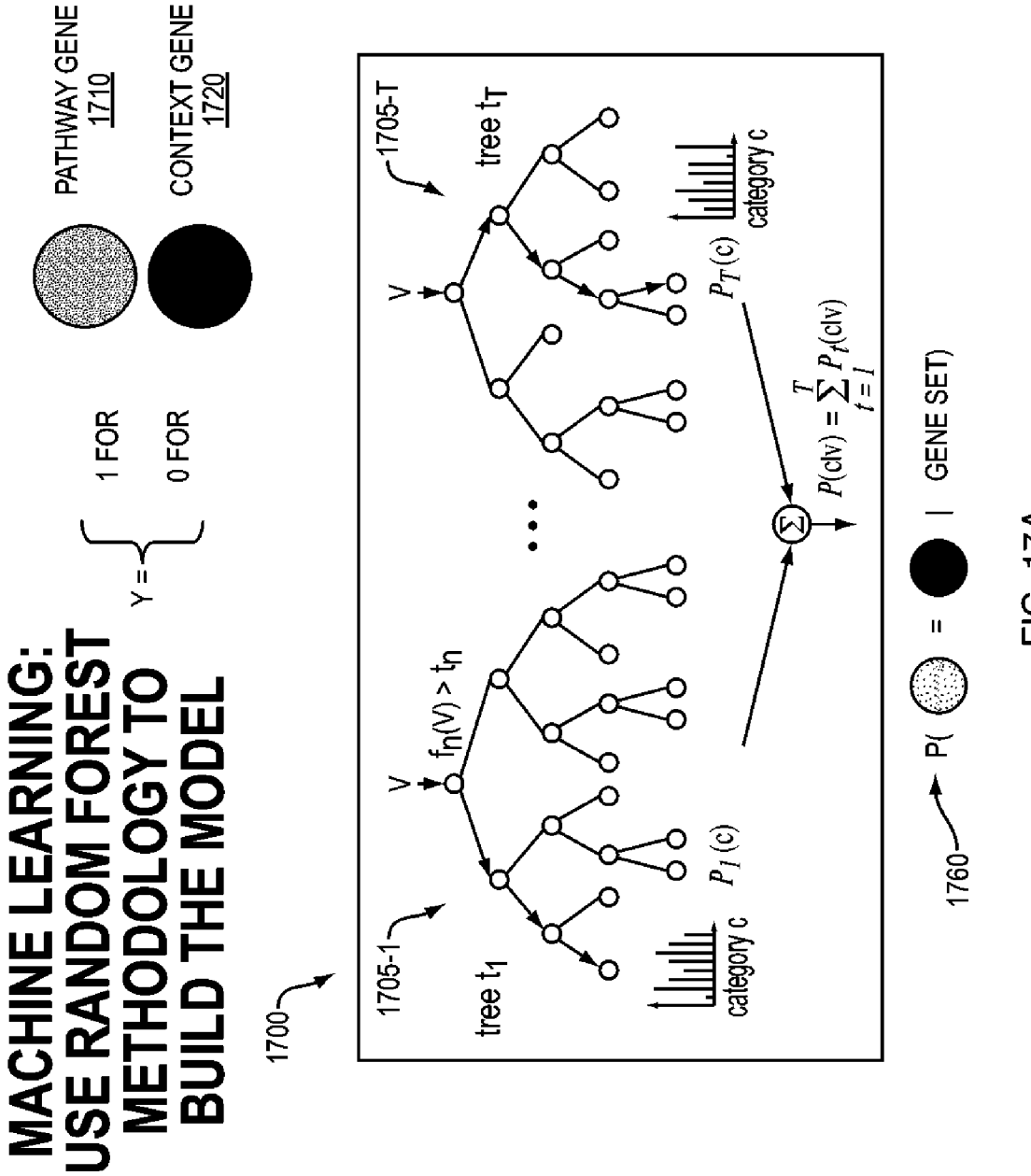

MACHINE LEARNING:
USE RANDOM FOREST
METHODOLOGY TO
BUILD THE MODEL
PATHWAY GENE
1710
CONTEXT GENE
1720
Y =
1 FOR
0 FOR
1700
1705-1
tree t_1
V
$f_n(V) > t_n$
category c
$P_1(c)$
1705-T
tree t_T
V
$P_T(c)$
category c
$P(c|v) = \sum_{t=1}^{T} P_t(c|v)$
1760
P(= | GENE SET)

FIG. 17A

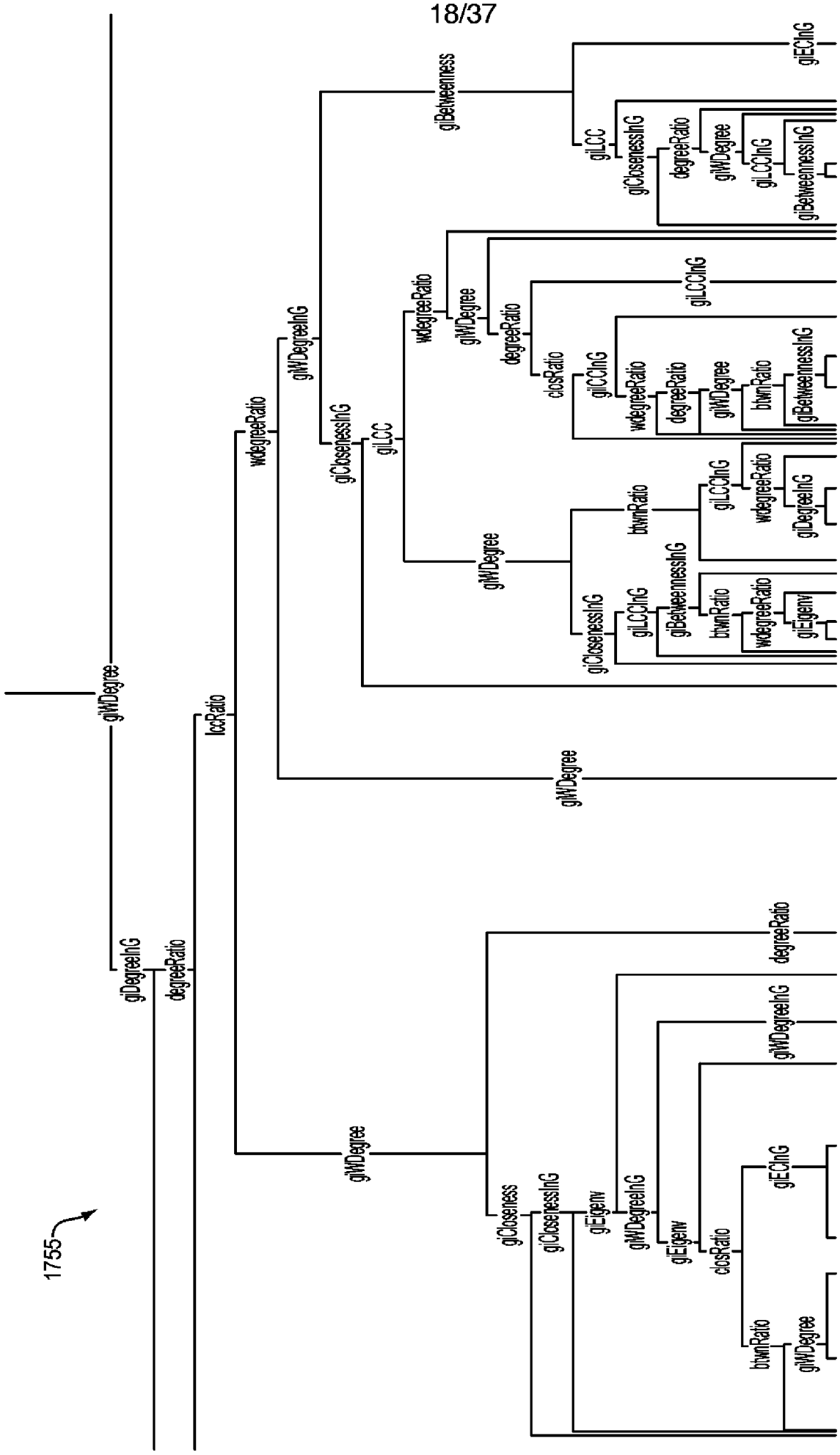

1755
gWDegree
giDegreeInG
degreeRatio
lccRatio
wdegreeRatio
gWDegreeInG
giClosenessInG
giLCC
giBetweenness
giECInG
giLCCInG
gWDegree
giCloseness
giEigenv
giBetweennessInG
btwnRatio
closRatio

FIG. 17B

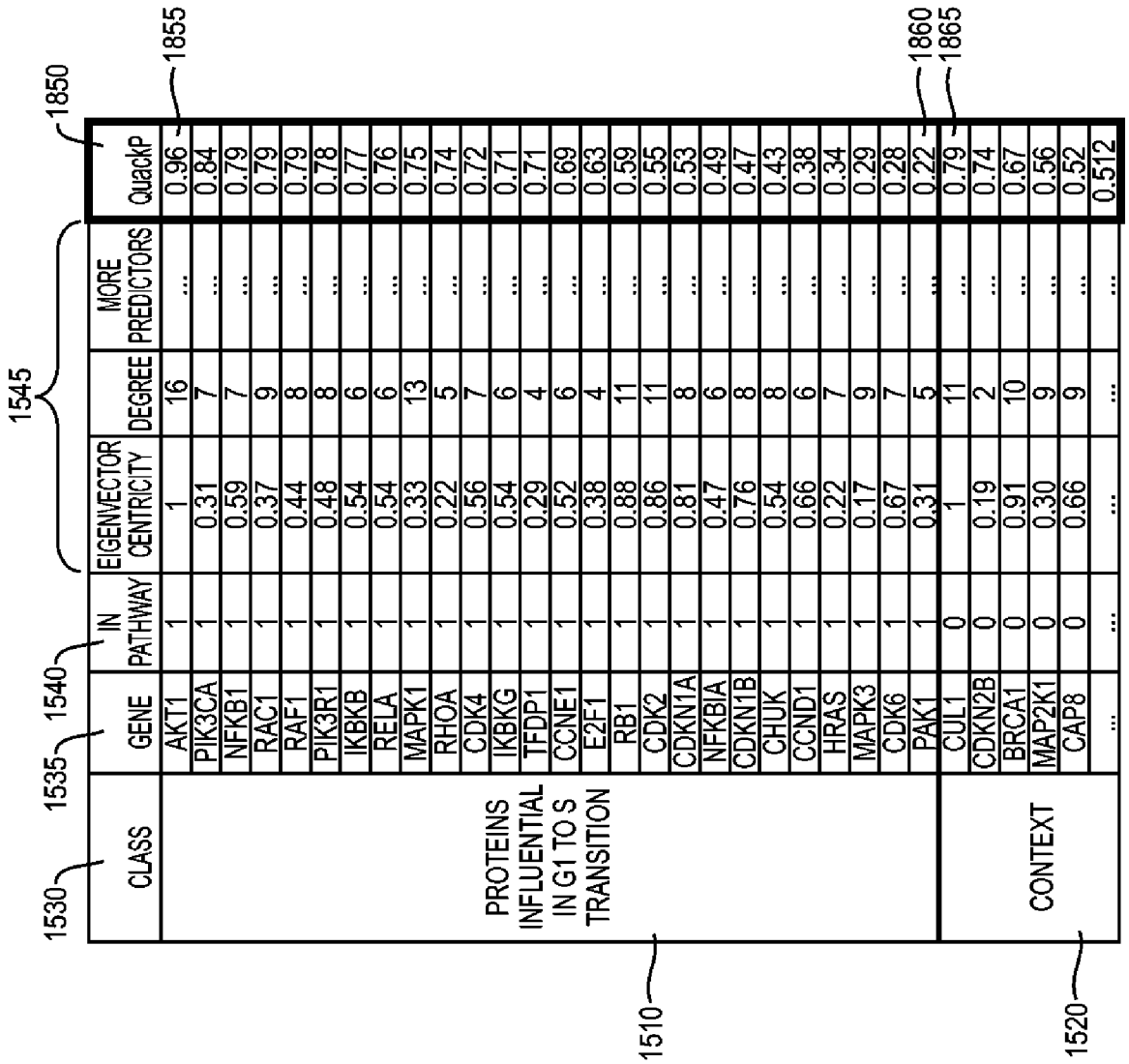

1530
1535
1540
1545
1850
1855
1860
1865
1510
1520
CLASS
GENE
IN PATHWAY
EIGENVECTOR CENTRICITY
DEGREE
MORE PREDICTORS
QuackP
PROTEINS INFLUENTIAL IN G1 TO S TRANSITION
AKT1 1 1 16 ... 0.96
PIK3CA 1 0.31 7 ... 0.84
NFKB1 1 0.59 7 ... 0.79
RAC1 1 0.37 9 ... 0.79
RAF1 1 0.44 8 ... 0.79
PIK3R1 1 0.48 8 ... 0.78
IKBKB 1 0.54 6 ... 0.77
RELA 1 0.54 6 ... 0.76
MAPK1 1 0.33 13 ... 0.75
RHOA 1 0.22 5 ... 0.74
CDK4 1 0.56 7 ... 0.72
IKBKG 1 0.54 6 ... 0.71
TFDP1 1 0.29 4 ... 0.71
CCNE1 1 0.52 6 ... 0.69
E2F1 1 0.38 4 ... 0.63
RB1 1 0.88 11 ... 0.59
CDK2 1 0.86 11 ... 0.55
CDKN1A 1 0.81 8 ... 0.53
NFKBIA 1 0.47 6 ... 0.49
CDKN1B 1 0.76 8 ... 0.47
CHUK 1 0.54 8 ... 0.43
CCND1 1 0.66 6 ... 0.38
HRAS 1 0.22 7 ... 0.34
MAPK3 1 0.17 9 ... 0.29
CDK6 1 0.67 7 ... 0.28
PAK1 1 0.31 5 ... 0.22
CONTEXT
CUL1 0 1 11 ... 0.79
CDKN2B 0 0.19 2 ... 0.74
BRCA1 0 0.91 10 ... 0.67
MAP2K1 0 0.30 9 ... 0.56
CAP8 0 0.66 9 ... 0.52
... ... 0.512

FIG. 18

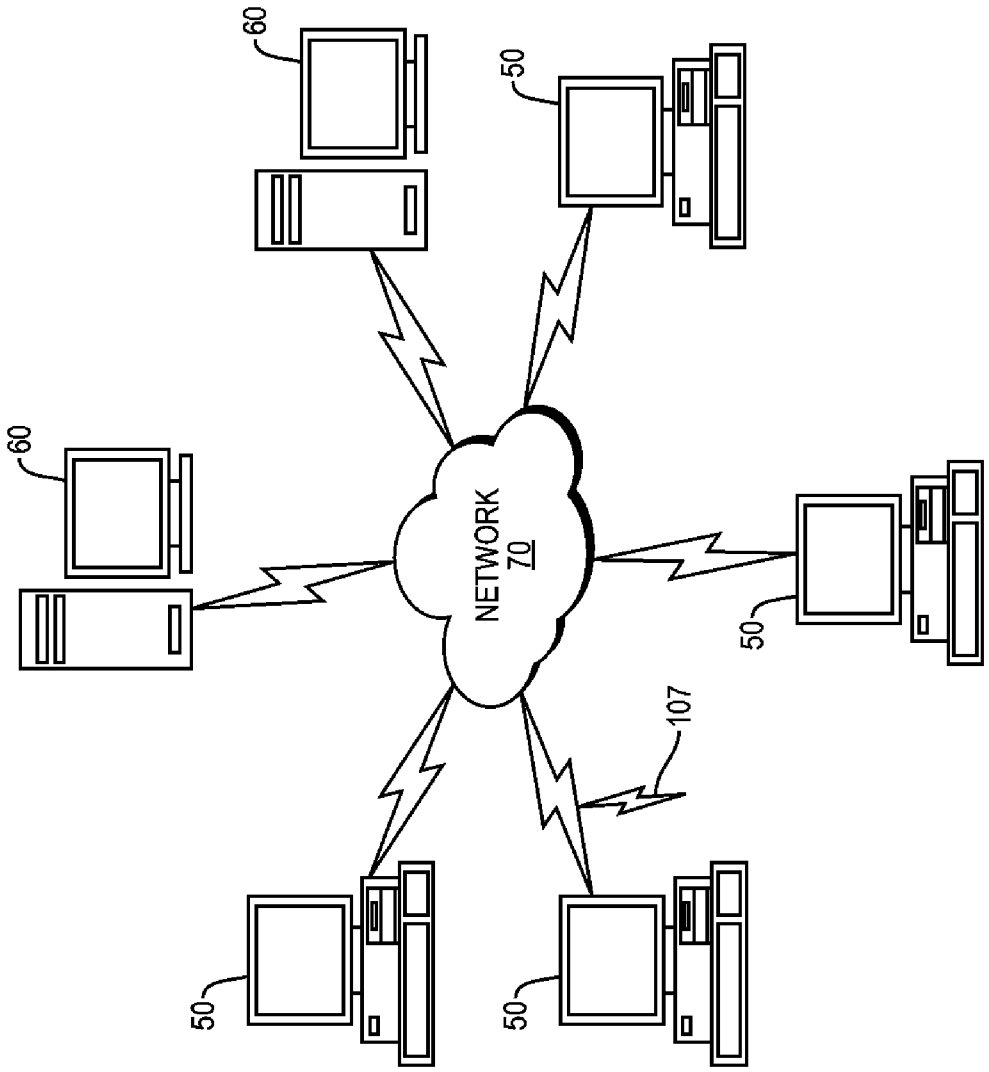


FIG. 19

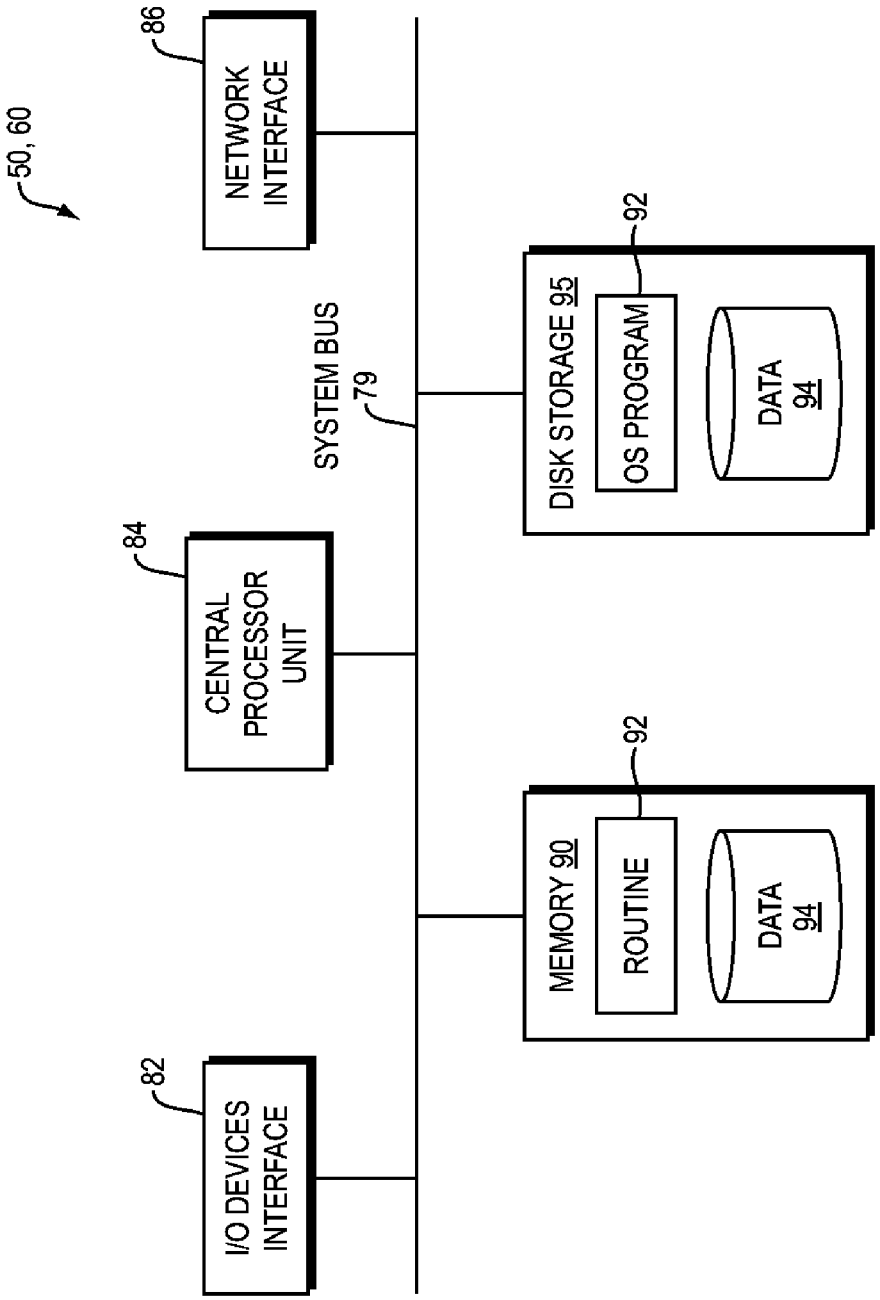

50, 60
86
NETWORK INTERFACE
SYSTEM BUS
79
92
DISK STORAGE 95
OS PROGRAM
DATA 94
84
CENTRAL PROCESSOR UNIT
92
MEMORY 90
ROUTINE
DATA 94
82
I/O DEVICES INTERFACE

FIG. 20

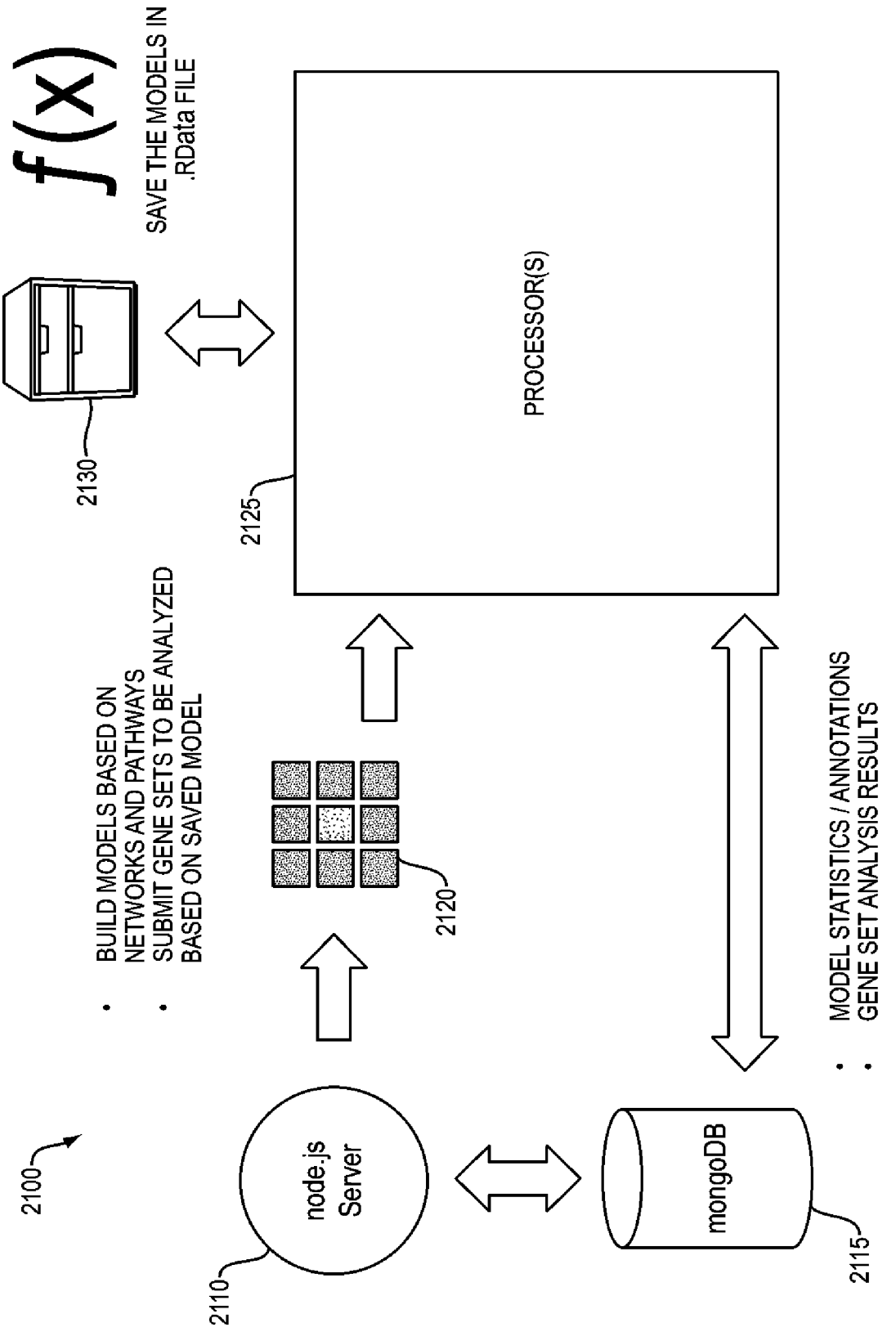

2100
node.js Server
2110
mongoDB
2115
2120
2125
PROCESSOR(S)
2130
f(x)
SAVE THE MODELS IN .RData FILE
BUILD MODELS BASED ON NETWORKS AND PATHWAYS
SUBMIT GENE SETS TO BE ANALYZED BASED ON SAVED MODEL
MODEL STATISTICS / ANNOTATIONS
GENE SET ANALYSIS RESULTS

FIG. 21

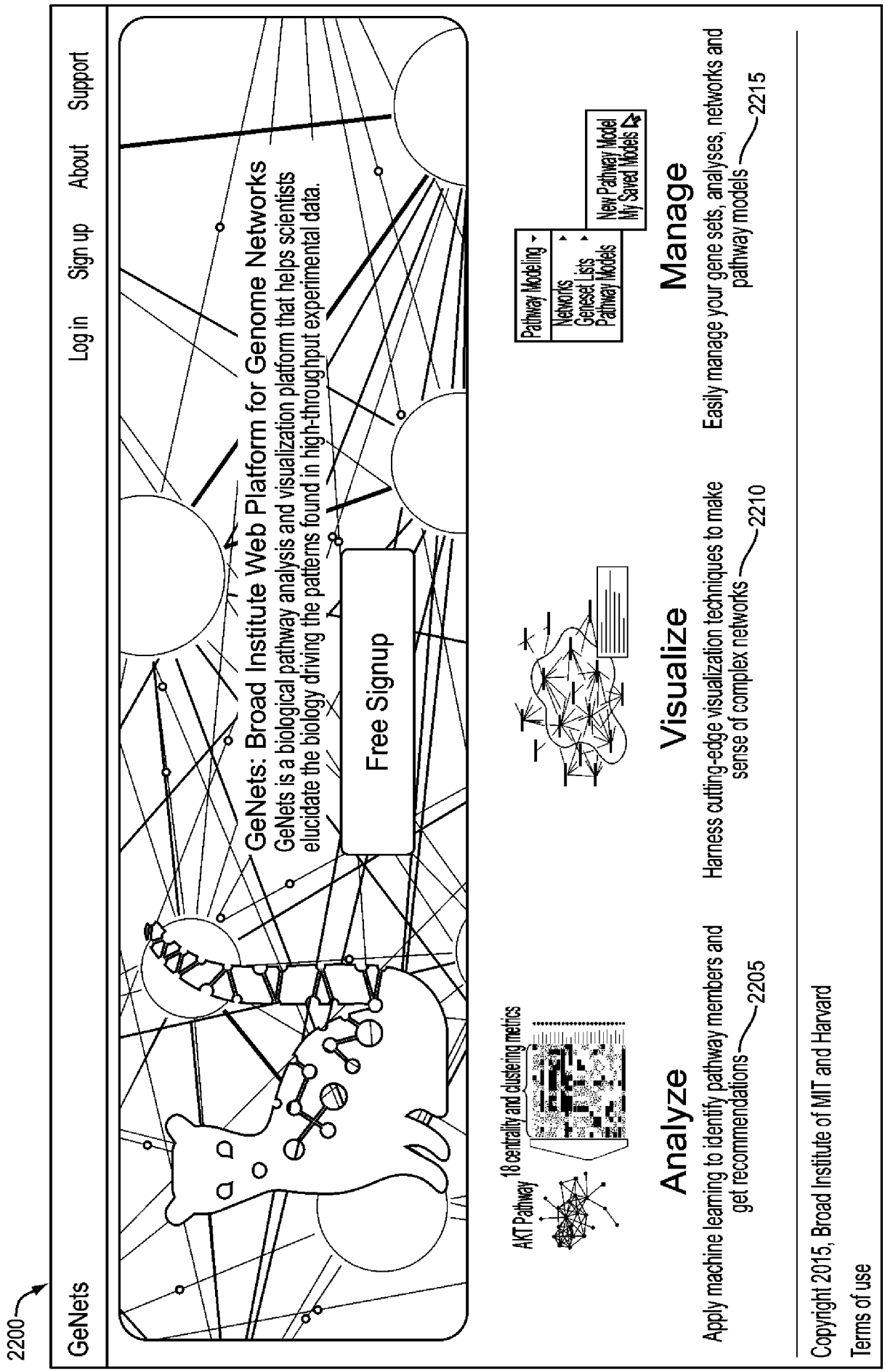

2200
GeNets
Log in
Sign up
About
Support
GeNets: Broad Institute Web Platform for Genome Networks
GeNets is a biological pathway analysis and visualization platform that helps scientists elucidate the biology driving the patterns found in high-throughput experimental data.
Free Signup
Pathway Modeling
Networks
Geneset Lists
Pathway Models
New Pathway Model
My Saved Models
Manage
Easily manage your gene sets, analyses, networks and pathway models
2215
Visualize
Harness cutting-edge visualization techniques to make sense of complex networks
2210
Analyze
Apply machine learning to identify pathway members and get recommendations
2205
AKT Pathway
18 centrality and clustering metrics
Copyright 2015, Broad Institute of MIT and Harvard
Terms of use

FIG. 22A

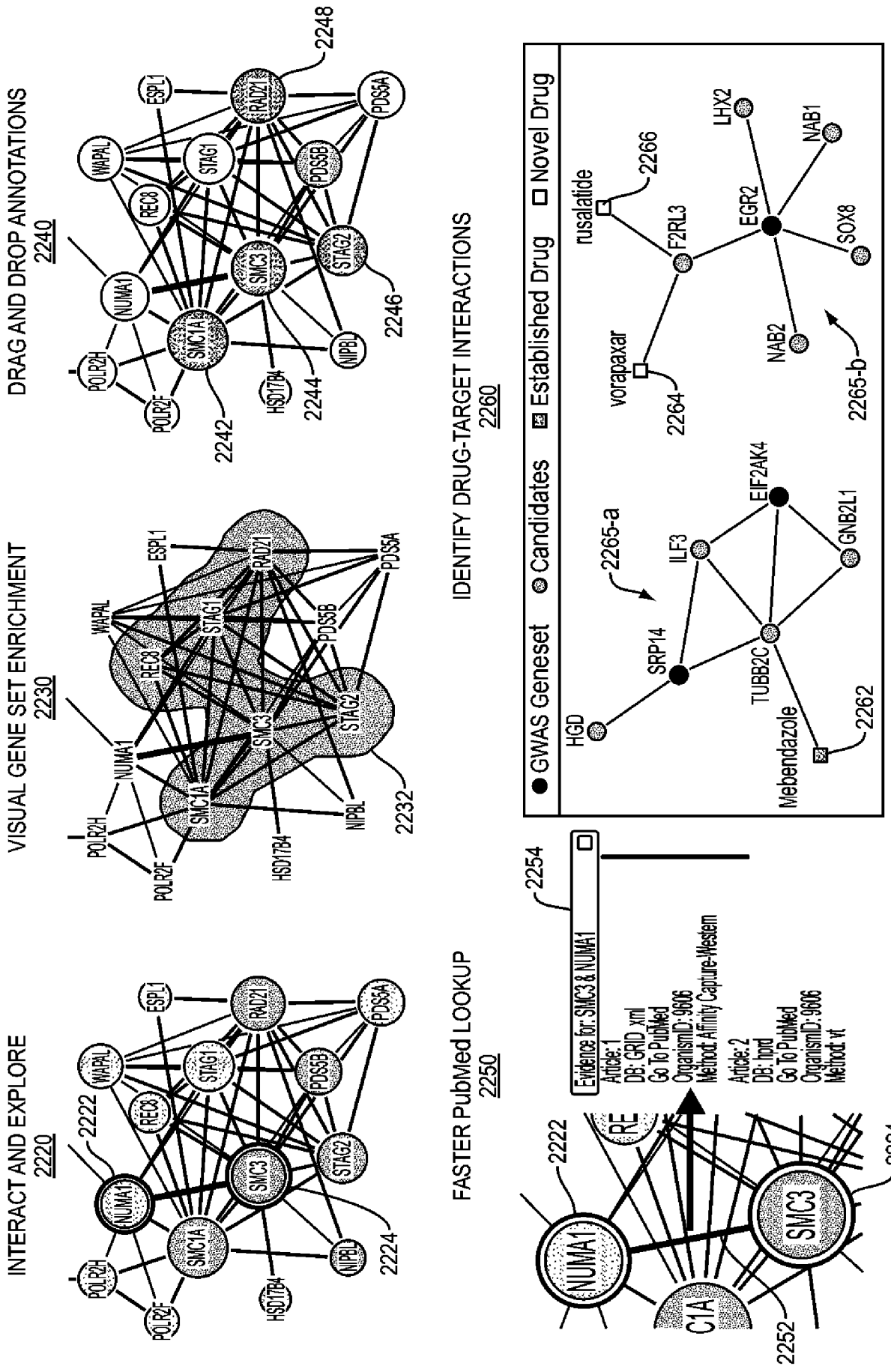

INTERACT AND EXPLORE
2220
2222
2224
VISUAL GENE SET ENRICHMENT
2230
2232
DRAG AND DROP ANNOTATIONS
2240
2242
2244
2246
2248
IDENTIFY DRUG-TARGET INTERACTIONS
2260
GWAS Geneset
Candidates
Established Drug
Novel Drug
HGD
SRP14
2265-a
ILF3
EIF2AK4
GNB2L1
TUBB2C
Mebendazole
2262
vorapaxar
2264
F2RL3
rusalatide
2266
EGR2
LHX2
NAB1
SOX8
NAB2
2265-b
FASTER PubMed LOOKUP
2250
2254
Evidence for: SMC3 & NUMA1
Article: 1
DB: GRID_xml
Go To PubMed
OrganismID: 9606
Method: Affinity Capture-Western
Article: 2
DB: hprd
Go To PubMed
OrganismID: 9606
Method: vt
2252
NUMA1
SMC3
SMC1A
STAG1
STAG2
RAD21
ESPL1
WAPAL
REC8
PDS5A
PDS5B
NIPBL
HSD17B4
POLR2H
POLR2F

FIG. 22B

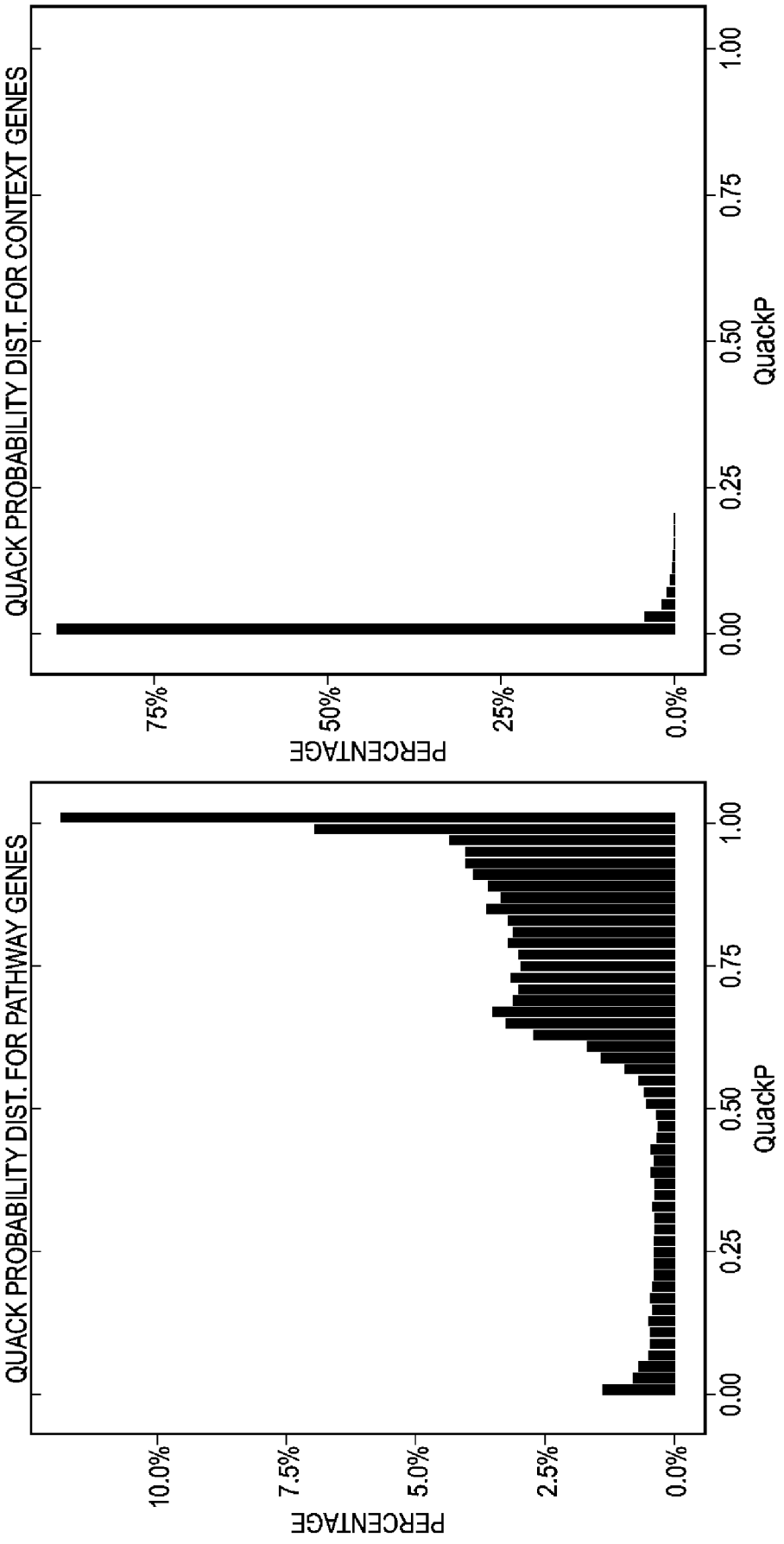

QUACK PROBABILITY DIST. FOR CONTEXT GENES
PERCENTAGE
75%
50%
25%
0.0%
0.00
0.25
0.50
0.75
1.00
QuackP
QUACK PROBABILITY DIST. FOR PATHWAY GENES
PERCENTAGE
10.0%
7.5%
5.0%
2.5%
0.0%
0.00
0.25
0.50
0.75
1.00
QuackP

FIG. 23

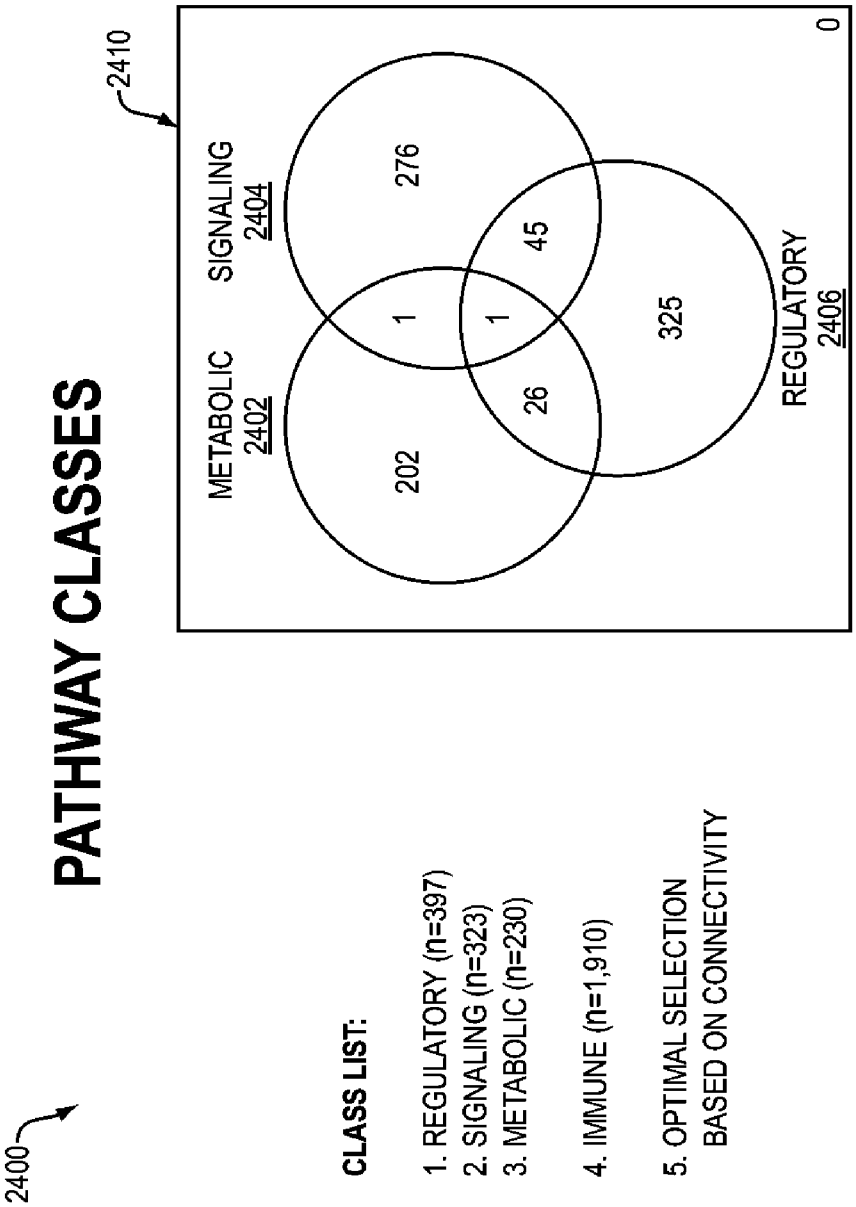

2400
PATHWAY CLASSES
2410
METABOLIC
2402
SIGNALING
2404
REGULATORY
2406
202
1
276
1
45
26
325
0
CLASS LIST:
1. REGULATORY (n=397)
2. SIGNALING (n=323)
3. METABOLIC (n=230)
4. IMMUNE (n=1,910)
5. OPTIMAL SELECTION
BASED ON CONNECTIVITY

FIG. 24

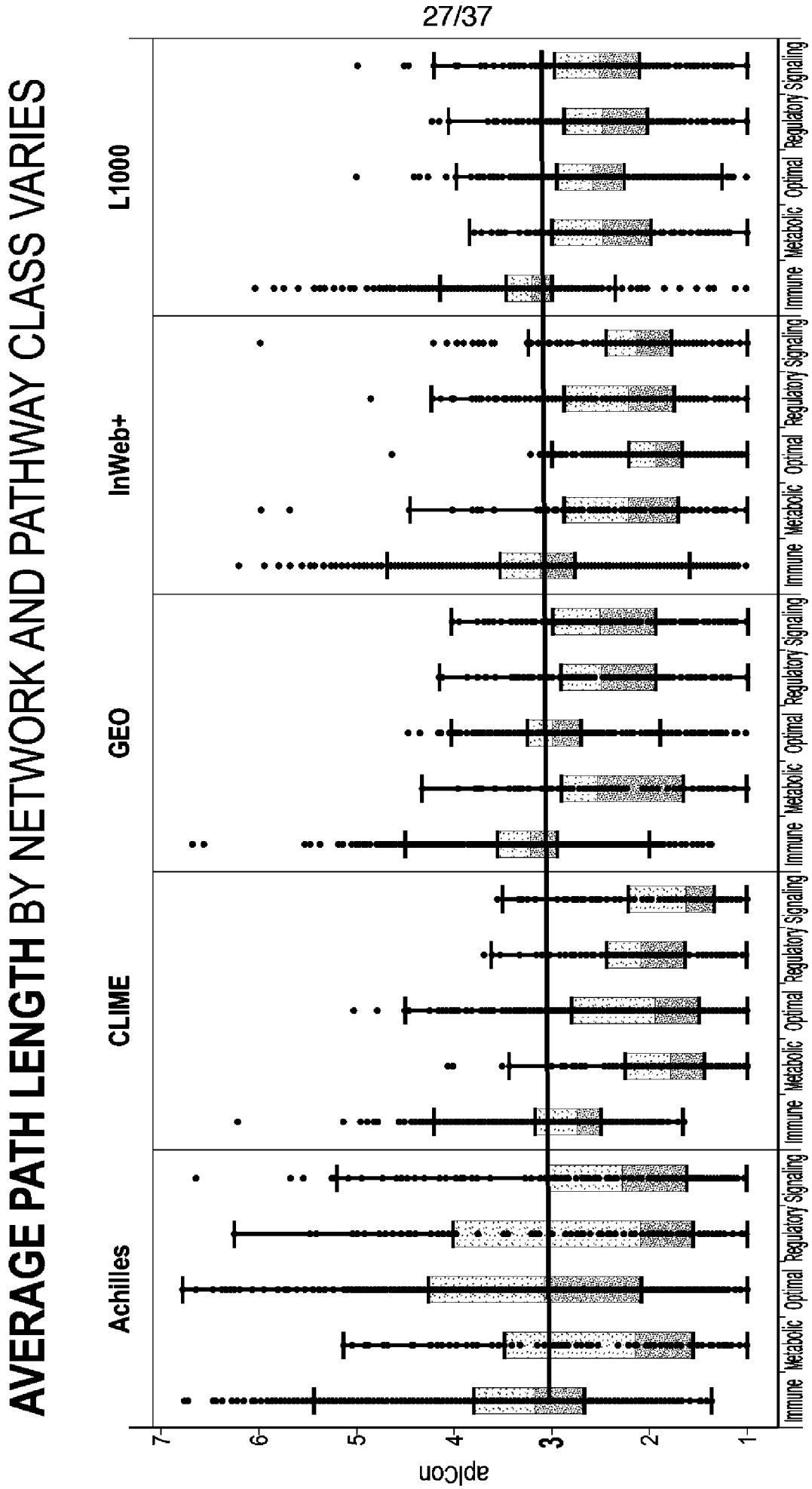

AVERAGE PATH LENGTH BY NETWORK AND PATHWAY CLASS VARIES
Achilles
CLIME
GEO
InWeb+
L1000
apl\con
7
6
5
4
3
2
1
Immune
Metabolic
Optimal
Regulatory
Signaling
I M O R S

FIG. 25A

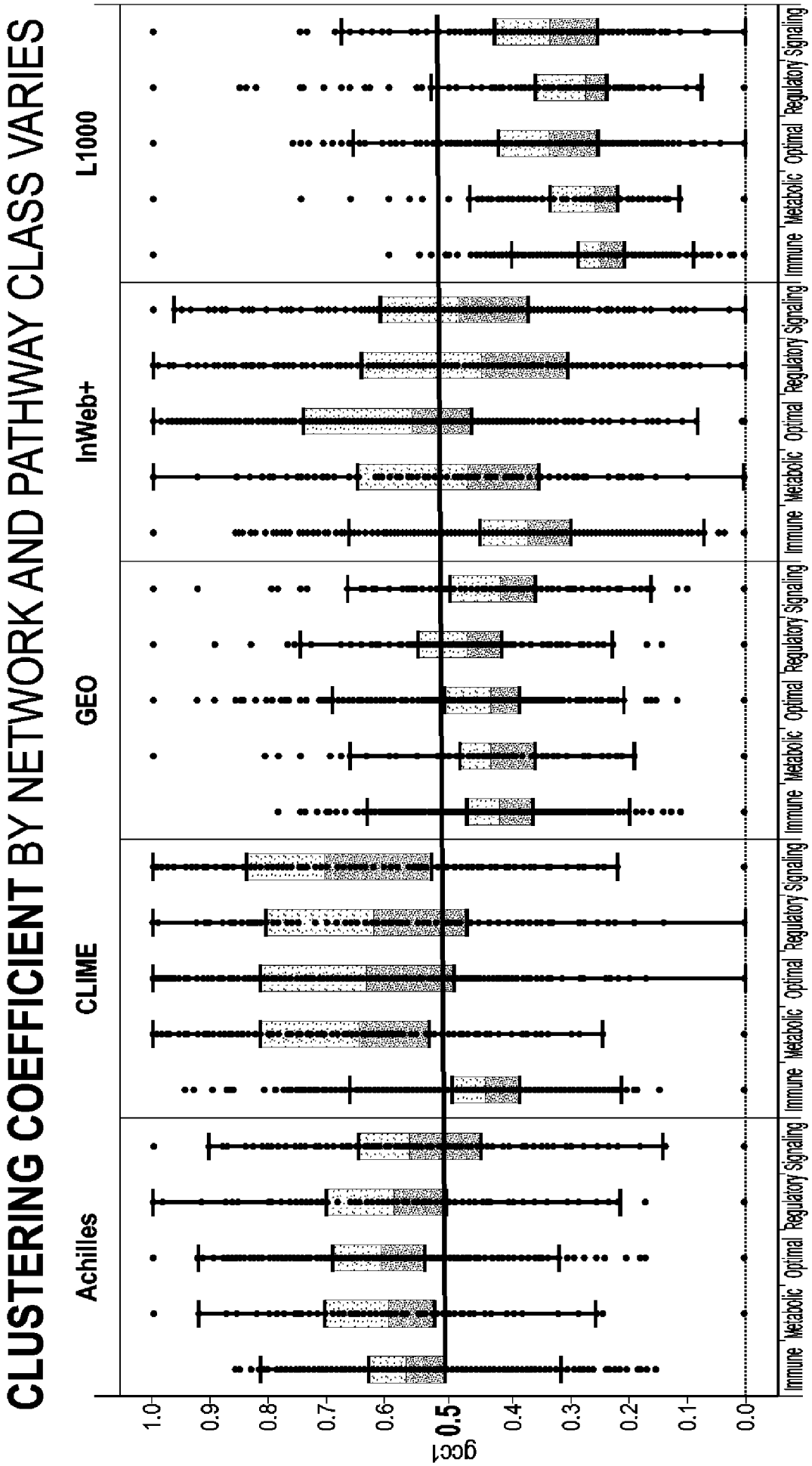

CLUSTERING COEFFICIENT BY NETWORK AND PATHWAY CLASS VARIES
Achilles
CLIME
GEO
InWeb+
L1000
gcc1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Immune
Metabolic
Optimal
Regulatory
Signaling
I M O R S

FIG. 25B

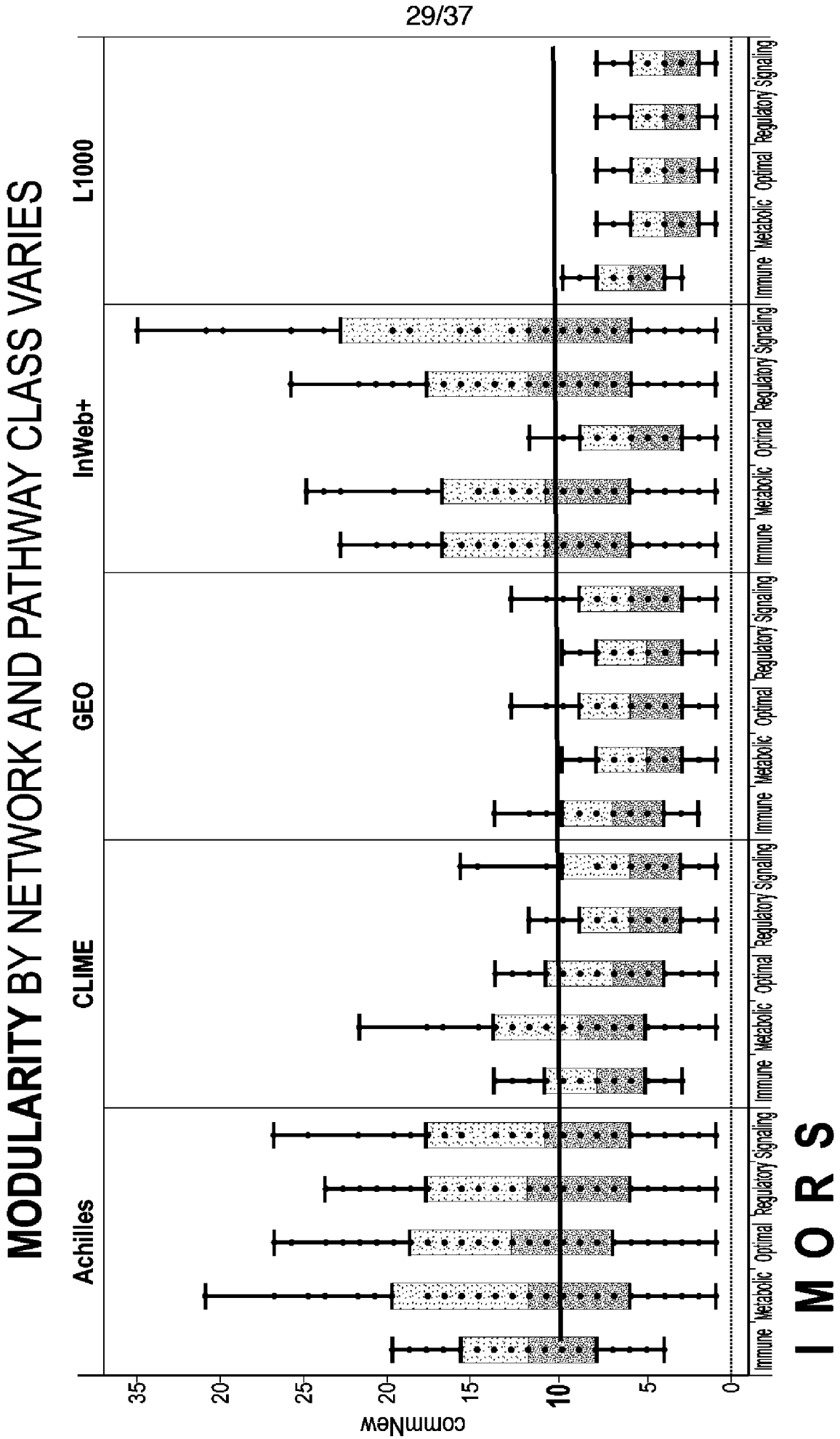

MODULARITY BY NETWORK AND PATHWAY CLASS VARIES
Achilles
CLIME
GEO
InWeb+
L1000
commNew
35
20
25
20
15
10
5
0
Immune
Metabolic
Optimal
Regulatory
Signaling
I M O R S

FIG. 25C

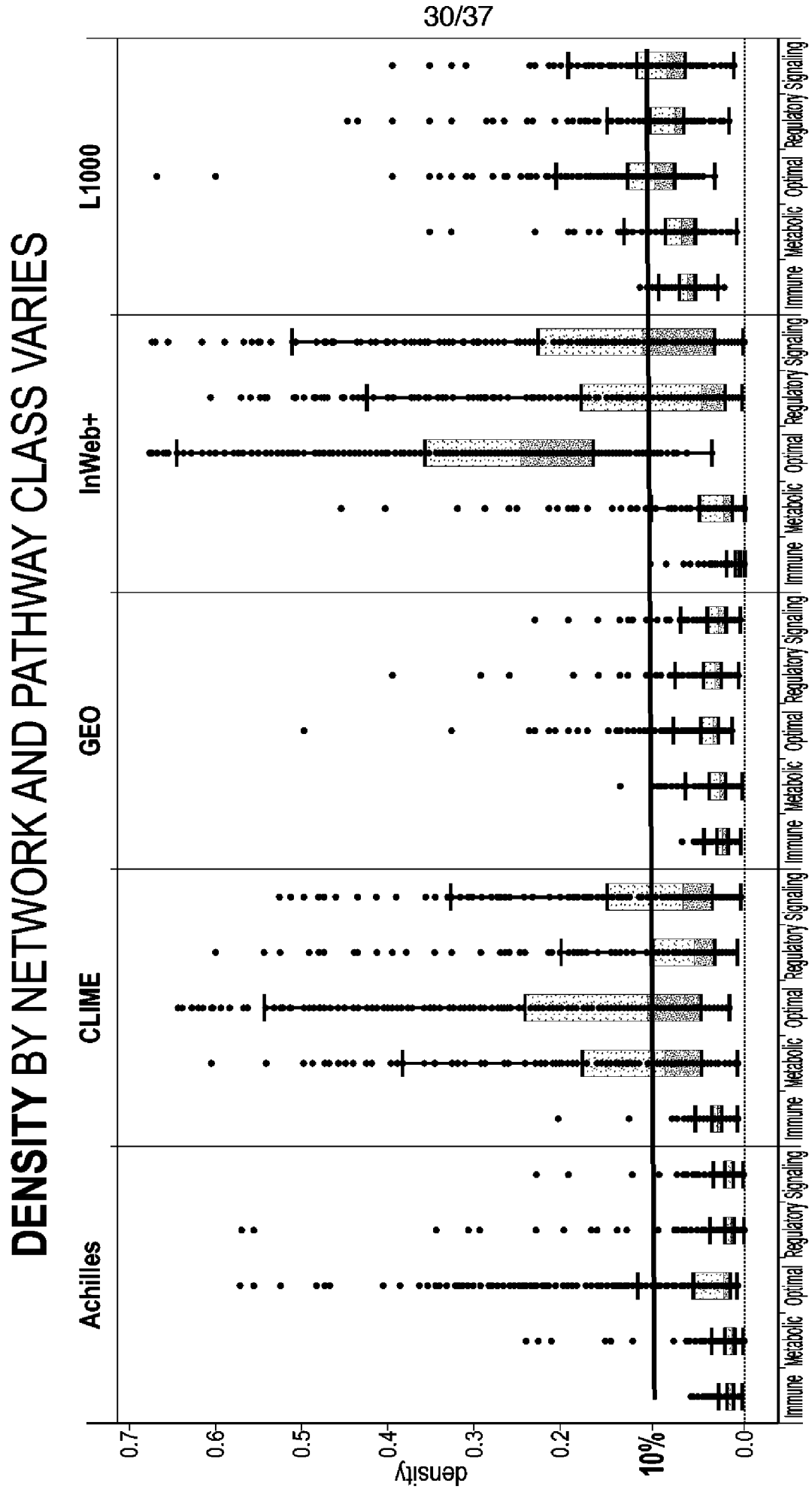

DENSITY BY NETWORK AND PATHWAY CLASS VARIES
Achilles
CLIME
GEO
InWeb+
L1000
density
0.7
0.6
0.5
0.4
0.3
0.2
10%
0.0
Immune
Metabolic
Optimal
Regulatory
Signaling
I M O R S

FIG. 25D

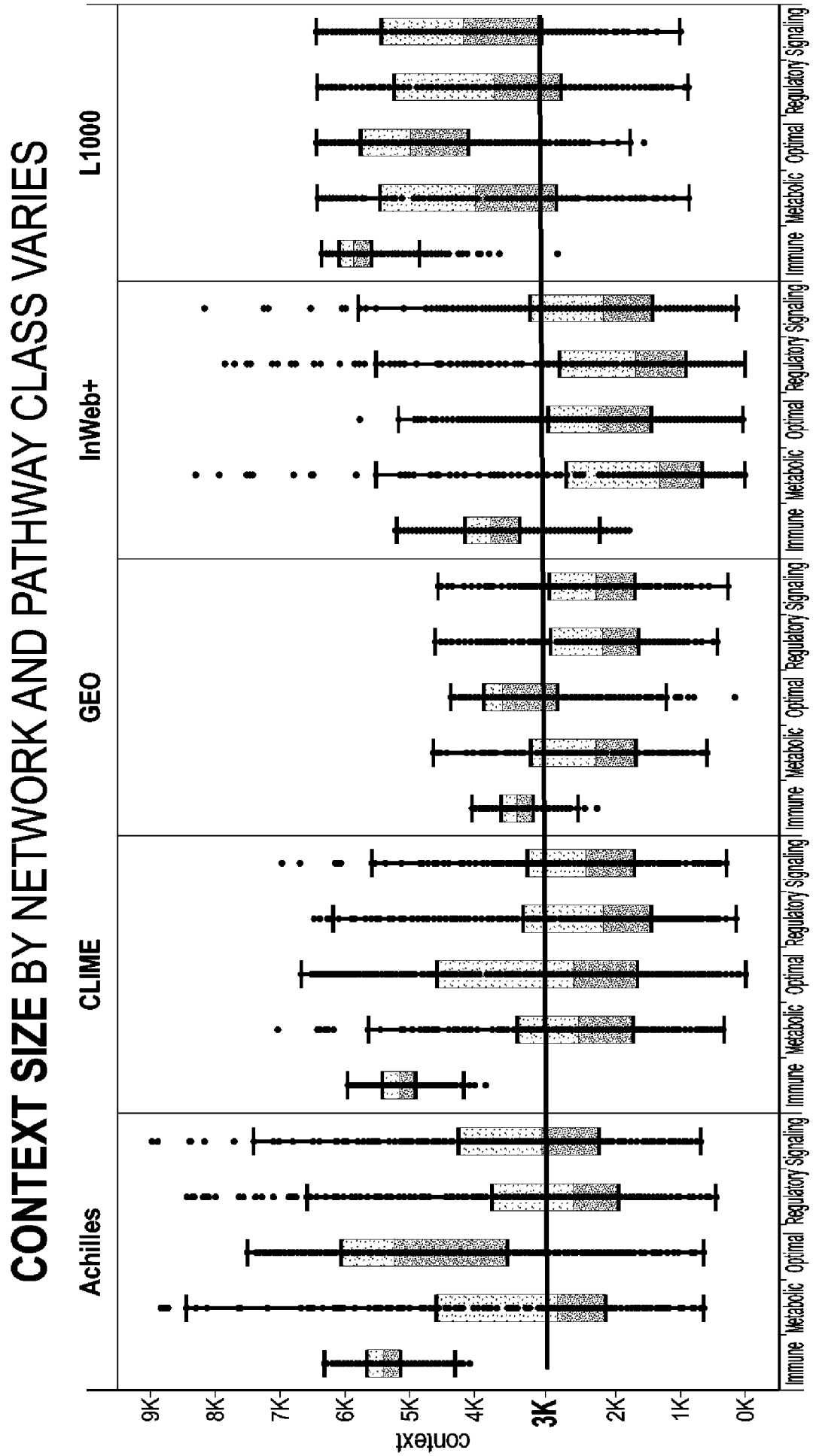

CONTEXT SIZE BY NETWORK AND PATHWAY CLASS VARIES
Achilles
CLIME
GEO
InWeb+
L1000
context
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
Immune
Metabolic
Optimal
Regulatory
Signaling
I M O R S

FIG. 25E

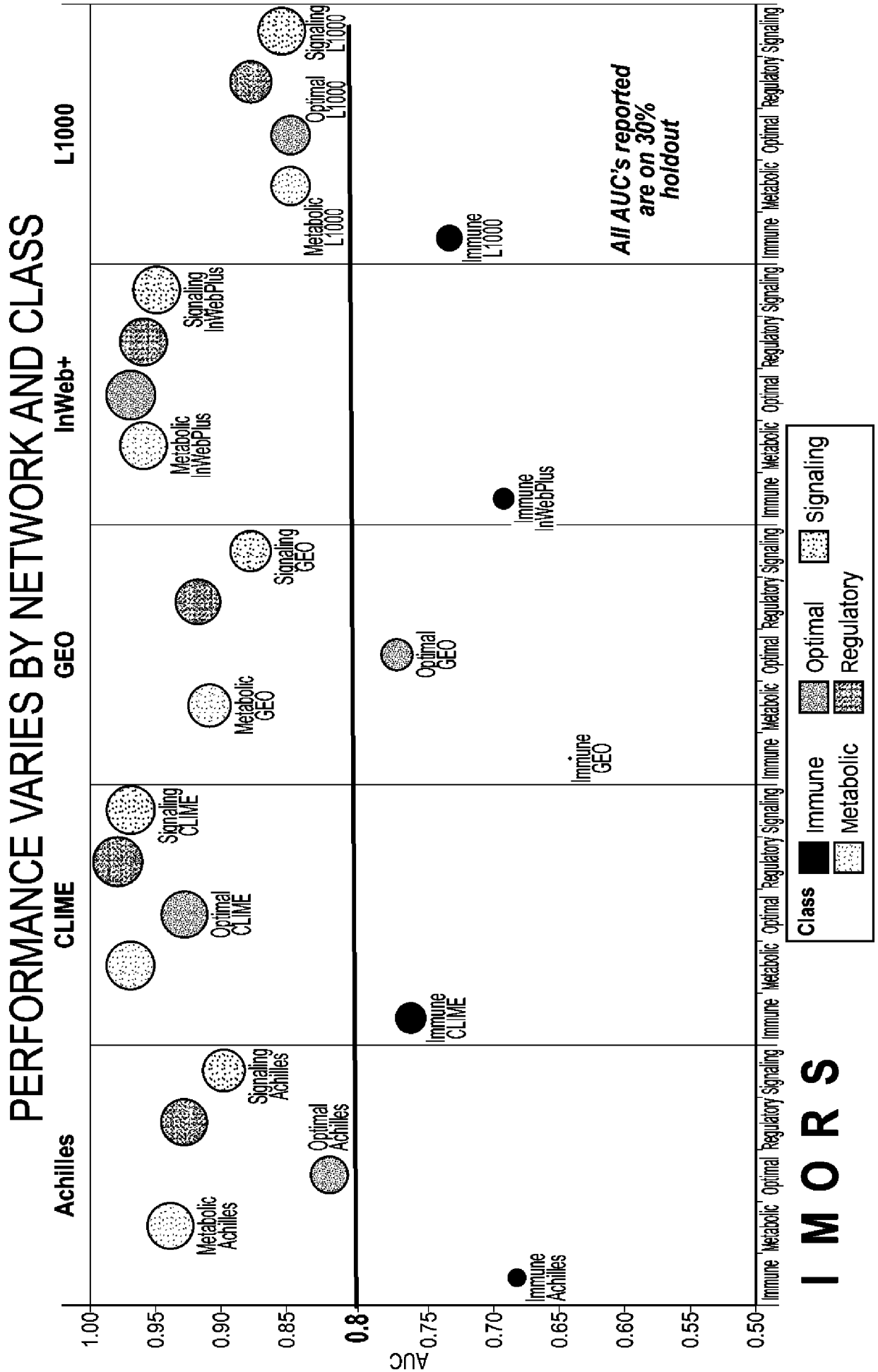

PERFORMANCE VARIES BY NETWORK AND CLASS
Achilles
CLIME
GEO
InWeb+
L1000
AUC
1.00
0.95
0.90
0.85
0.8
0.75
0.70
0.65
0.60
0.55
0.50
Metabolic Achilles
Optimal Achilles
Signaling Achilles
Immune Achilles
Optimal CLIME
Signaling CLIME
Immune CLIME
Metabolic GEO
Signaling GEO
Optimal GEO
Immune GEO
Metabolic InWebPlus
Signaling InWebPlus
Immune InWebPlus
Metabolic L1000
Optimal L1000
Signaling L1000
Immune L1000
All AUC's reported are on 30% holdout
Class
Immune
Metabolic
Optimal
Regulatory
Signaling
I M O R S

FIG. 26

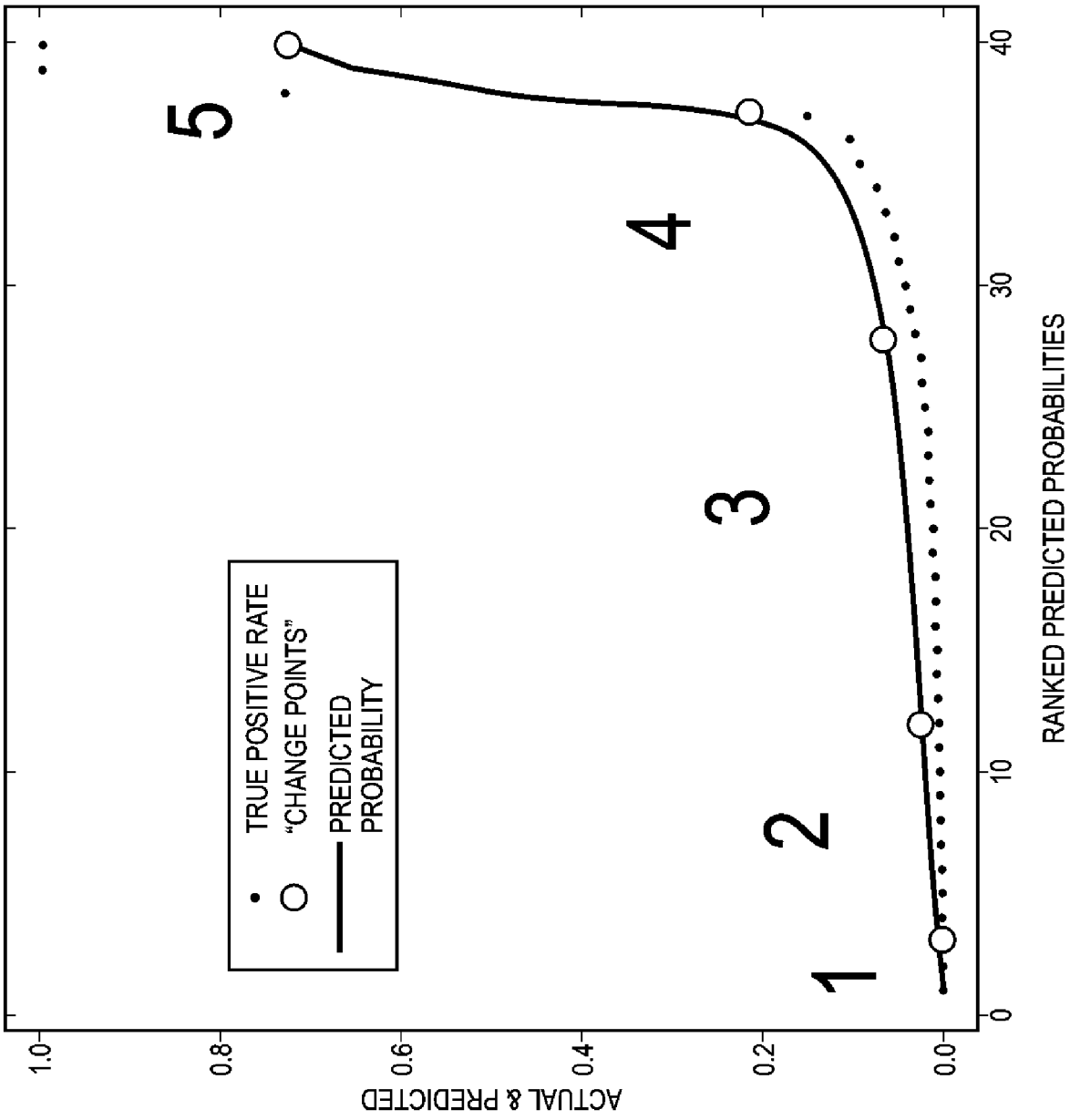

ACTUAL & PREDICTED
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
RANKED PREDICTED PROBABILITIES
TRUE POSITIVE RATE
"CHANGE POINTS"
PREDICTED PROBABILITY
1
2
3
4
5

FIG. 27

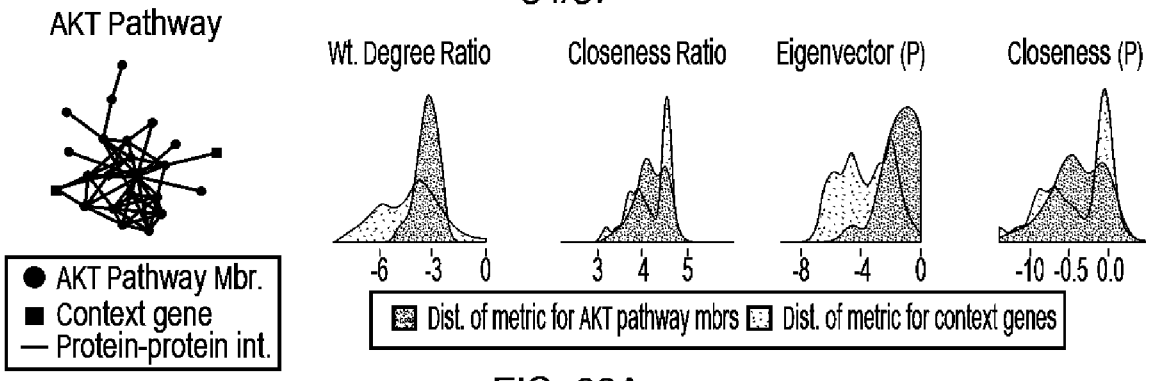


FIG. 28A

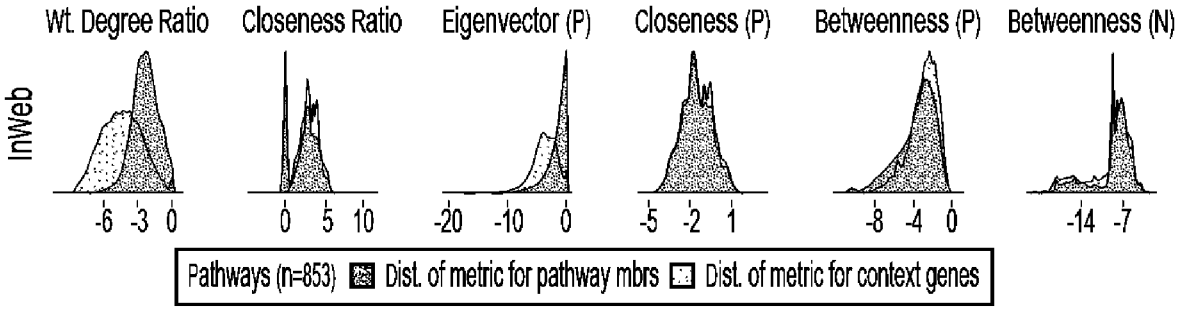


FIG. 28B

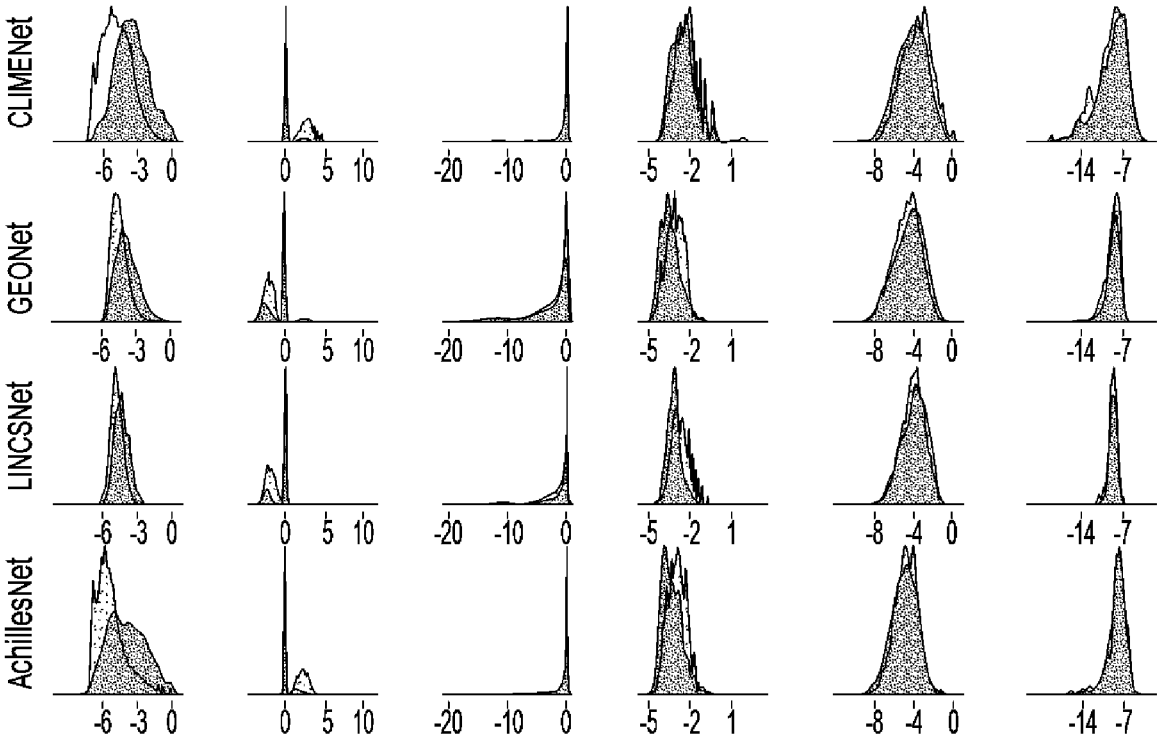


FIG. 28C

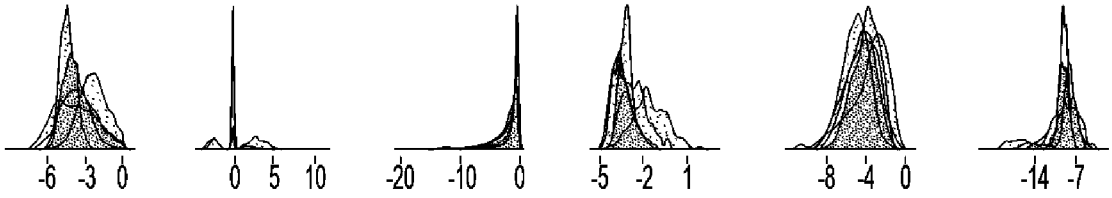


FIG. 28D

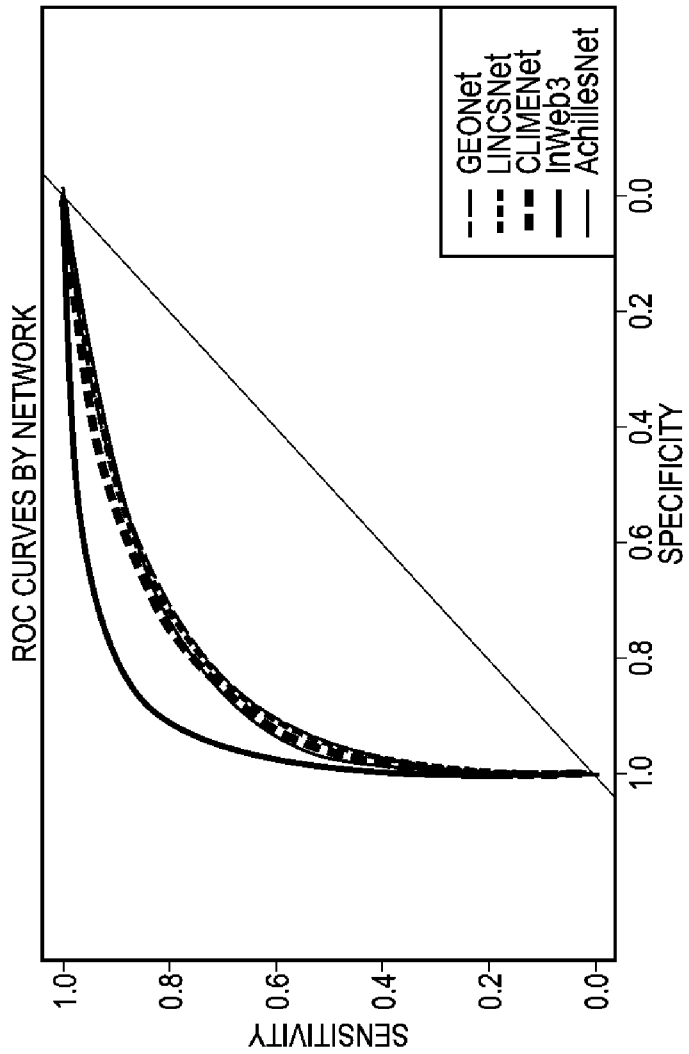


FIG. 29A

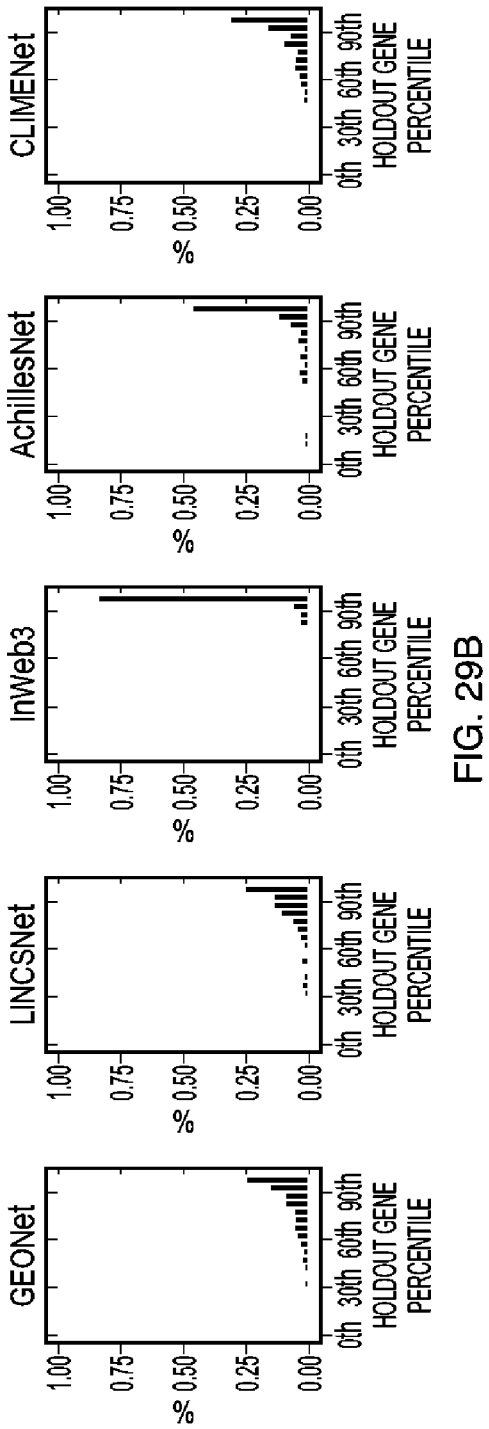


FIG. 29B

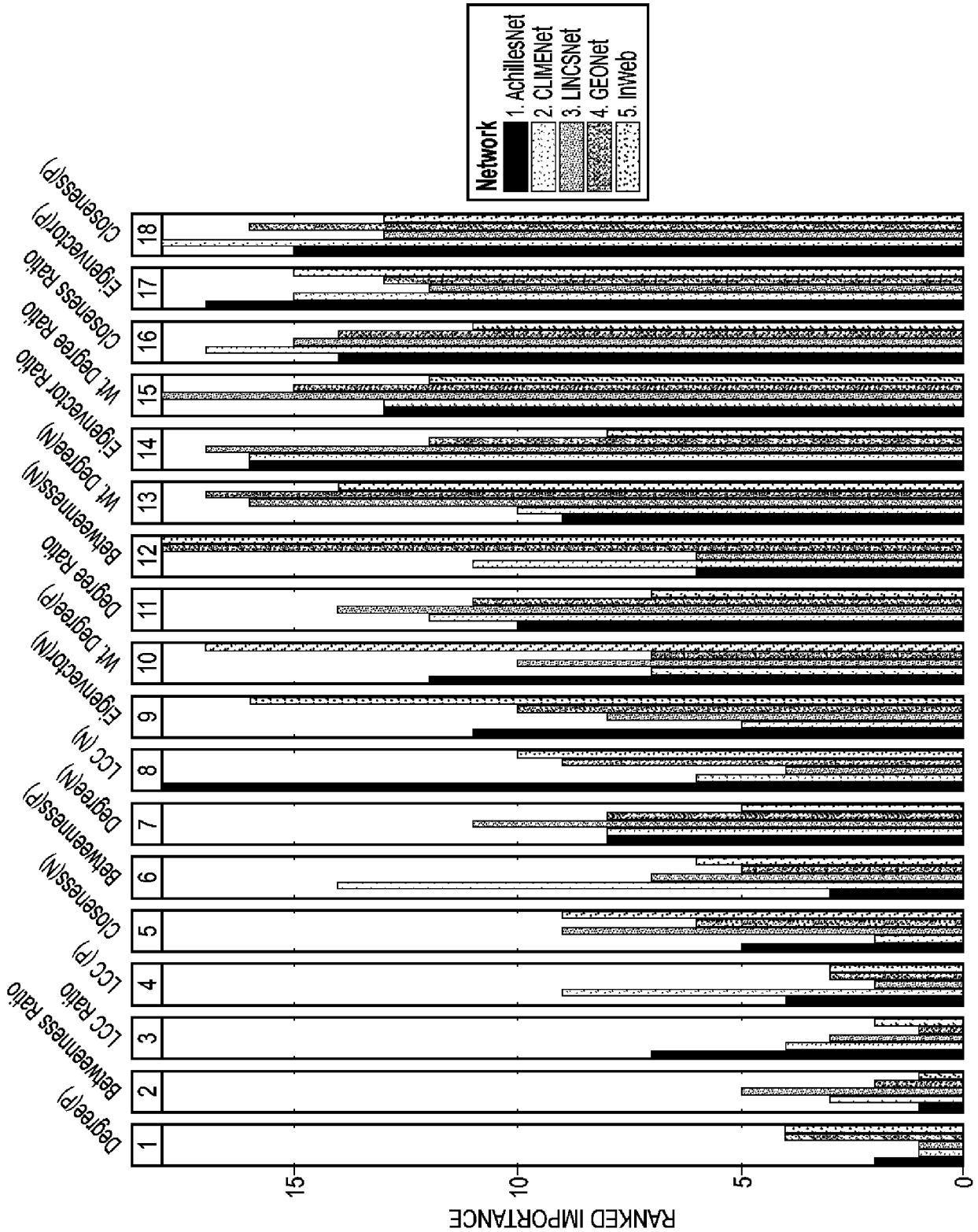

Network
1. AchillesNet
2. CLIMENet
3. LINCSNet
4. GEONet
5. InWeb
RANKED IMPORTANCE
0
5
10
15
1 Degree(P)
2 Betweenness Ratio
3 LCC Ratio
4 LCC (P)
5 Closeness(N)
6 Betweenness(P)
7 Degree(N)
8 LCC (N)
9 Eigenvector(N)
10 Wt. Degree(P)
11 Degree Ratio
12 Betweenness(N)
13 Wt. Degree(N)
14 Eigenvector Ratio
15 Wt. Degree Ratio
16 Closeness Ratio
17 Eigenvector(P)
18 Closeness(P)

FIG. 30

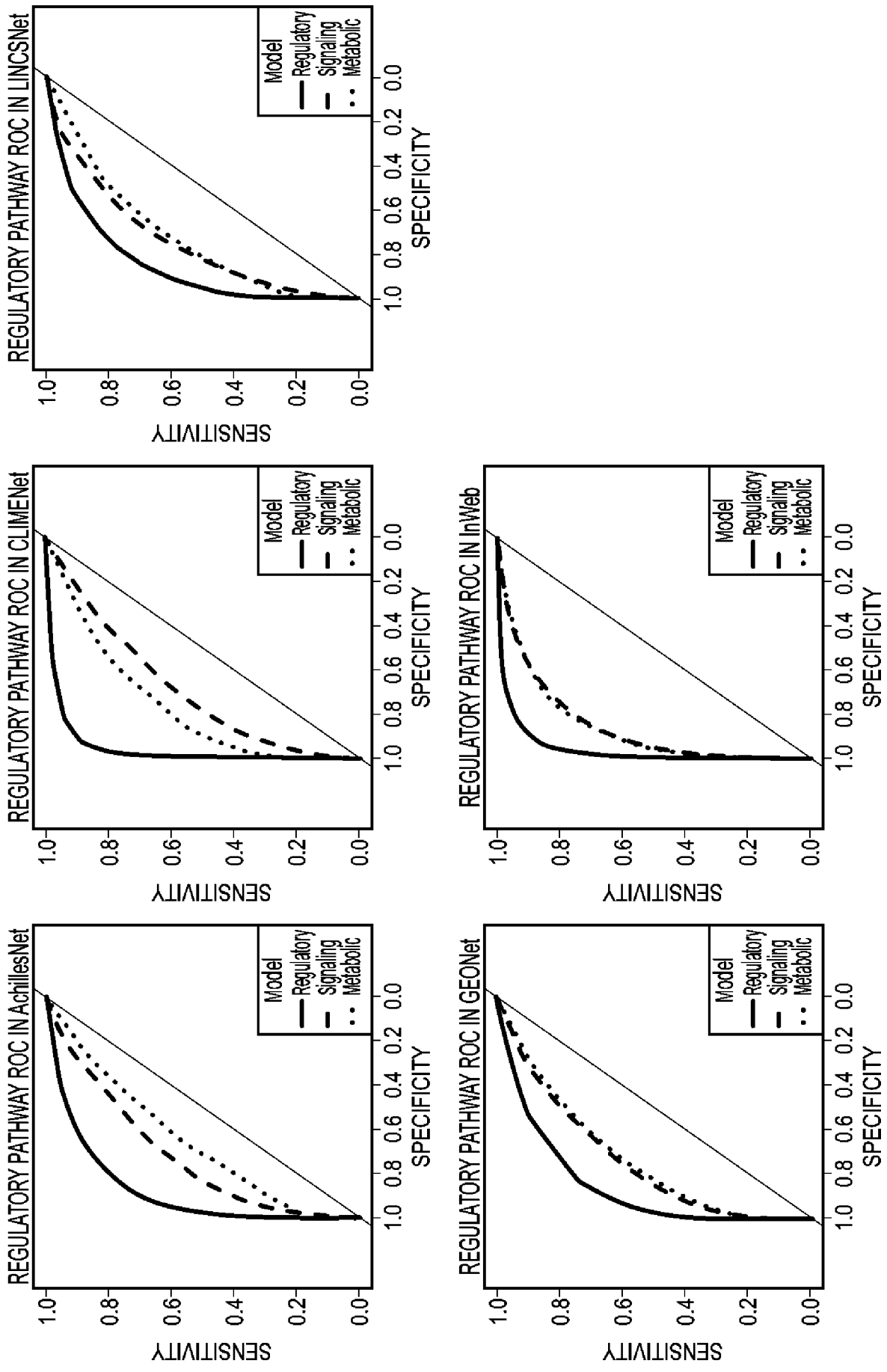

REGULATORY PATHWAY ROC IN LINCSNet
REGULATORY PATHWAY ROC IN CLIMENet
REGULATORY PATHWAY ROC IN InWeb
REGULATORY PATHWAY ROC IN AchillesNet
REGULATORY PATHWAY ROC IN GEONet
Model
Regulatory
Signaling
Metabolic
SENSITIVITY
SPECIFICITY
1.0
0.8
0.6
0.4
0.2
0.0

FIG. 31