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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 June 2002 (06.06.2002)

PCT

(10) International Publication Number
WO 02/44992 A2

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/US01/26887

(22) International Filing Date: 29 August 2001 (29.08.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/723,410 28 November 2000 (28.11.2000) US

(71) Applicant: **PHYSIOME SCIENCES, INC.** [US/US]; Suite 300, 150 College Road West, Princeton, NJ 08540-6604 (US).

(72) Inventors: **JIM, Kam-Chuen**; 6514 Town Court North, Lawrenceville, NJ 08648 (US). **LETT, Gregory, Scott**; 409 South Main Street, Hightstown, NJ 08520 (US). **PESTANO, Gary, Anthony**; 47-11 Fox Run Drive, Plainsboro, NJ 08536 (US). **ZUO, Zhuang**; 4231 Bayberry Court, Monmouth Junction, NJ 08852 (US).

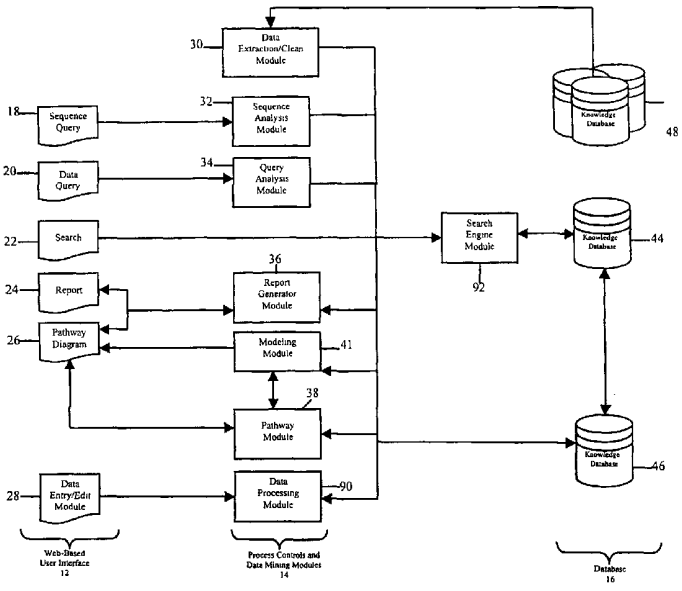
(74) Agent: **RESTAINO, Leslie, Gladstone**; Brown Raysman Millstein Felder & Steiner LLP, 4th floor, 55 Madison Avenue, Morristown, NJ 07960 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

*[Continued on next page]*

(54) Title: SYSTEM FOR MODELING BIOLOGICAL PATHWAYS




(57) Abstract: A computerized system for modeling biological pathways is provided. The system comprises a database having a knowledge database for storing at least one attribute of at least one entity and a pathway database for storing at least one pathway diagram; a user interface for creating, querying, manipulating and viewing data from the database; a process system having a modeling module for simulating or analyzing the behavior of an attribute on a pathway diagram and a pathway editor module for retrieving, editing and saving at least one pathway diagram. The invention may be implemented on a web-based system that provides end users with the ability to create new pathway models, edit or manipulate existing pathway models, perform sequence searches and data queries, and generate reports.

WO 02/44992 A2

**Declarations under Rule 4.17:**

- *as to the identity of the inventor (Rule 4.17(i)) for all designations*
- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for all designations*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii)) for all designations*

**Published:**

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

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

# SYSTEM FOR MODELING BIOLOGICAL PATHWAYS

## CROSS-REFERENCE TO RELATED APPLICATIONS

This is a continuation-in-part of U.S. Serial No. 09/499,575, filed February 7, 2000, which is a continuation-in-part of U.S. Serial No. 09/295,503, filed April 21, 1999, which claims the benefit of U.S. Serial No. 60/083,295, filed April 28, 1998.

## BACKGROUND OF THE INVENTION

### Field of the Invention:

The present invention relates to a computer-implemented system for storing, retrieving and utilizing biological data; and more particularly to a data processing system for modeling biological pathways.

### Description of the Related Art:

Cell biologists face a major challenge distilling the vast quantity of new data that is being generated at heretofore unprecedented rates. At present, hundreds of biological databases are listed in DBCAT, the INFOBIOGEN biological database catalog accessible from the World Wide Web (http://www.infobiogen.fr/services/dbcat/) and available publicly through the National Center for BioTechnology Information (http://www.ncbi nlm.nih.gov). This information explosion has been driven by the continuous development of information technology such as the Internet as well as the development of powerful new technologies for automatically collecting and storing data such as in gene sequencing and gene expression profiling. These databases contain genomic, proteomic, biochemical, chemical and molecular biology data as well as structural data comprising geometric and anatomical information from the subcellular to the whole organism level. Some of these data are organized by data type including, for example, the International Nucleic Acid Sequence Data Library (a.k.a. GenBank) and NAD for nucleic acid sequences; SWISS-PROT for protein sequences; RCDB, PROW and Pfam for protein structures and the like. Other databases are organism specific and include GDB and OMIM for human; MGD for mouse, PigBASE

for pig; ATDB for *Arabidopsis;* ECDC for *E. Coli,* SGD for yeast, and many others. Still other databases contain information on particular areas of interest, such as specific databases for individual genes, databases about specific protein families, and databases of transcription factors. Biochemical databases contain information regarding coupled biochemical reactions and feedback signals that take place within the cell. Additionally, proprietary databases such as the availability of entire genomic sequences due to improved high throughput gene sequencing, available from the large data production houses, have been created and are expanding with technology. Substantial work is underway to integrate data from these diverse databases. *See e.g.,* Macauley et al., A Model System for Studying the Integration of Molecular Biology Databases, 14 *Bioinformatics* 575-582 (1998).

Efforts to organize and analyze the vast amount of genomic data have stimulated the development of a new field of computational science known as bioinformatics -- the science of using computers and software to store, extract, organize, analyze, interpret and utilize gene sequence data to identify new genes and gene function -- in order to understand the genetic basis of disease and to further gene-based drug discovery and development. This approach typically uses a one-dimensional computational analysis to study explicit information about the genome such as percentage of gene sequence similarity across species, homology of sequence motifs across species, protein expression levels in various tissue types, secondary structure correlations, etc.

Although the acquisition of genomic information is clearly essential, there is growing recognition that conventional methods are insufficient for correlating that information with the functional role of genes and gene products. Rather, in all cells, genetic expression produces self-organizing networks controlling cell functions, including developmental pathways, progression through cell cycle, metabolism, intracellular signaling, cell excitability and motility, and feedback loops regulating gene expression. At present, bioinformatics is unable to simulate these complex, highly nonlinear dynamic interactions that occur between each gene or gene product, and other components of the network they are a part of. Thus, bioinformatics

researchers do not, at present, have the necessary tools to obtain a complete representation of subcellular and cellular processes, as well as the effect of these processes on tissues and organs.

One approach to dealing with these complex and highly nonlinear interactions has focused on computational modeling. There is an extensive forty-year history of such modeling that includes simple models with a few state equations that describe processes within cells to highly complex models of organ systems that must be implemented on high performance multiprocessor computers (Rall W., Burke R.E., Holmes W.R., Jack J.J., Redman S.J., Segev I. (1992) *Physiol. Rev.* 72(4 Suppl) 5159-86; Rall W. (1967) *J. NeuroPhysiol* 30(5): 1169-93, Segev I. and Rall W. (1998) *Trends Neurosci* 21(11): 453-60; Koch C., Poggio T., and Torre V. (1982) *Philos. Trans. Roy. Soc. Lond. B.* 298(1090):227-63, Chay T.R. and Rinzel J. (1985) *Biophys. J.* 47(3): 357-66; Smolen P., Rinzel J., Sherman A. (1993) *Biophys J.* 64(6): 1668-80, Shepherd G.M. et al. (1998) *Trends Neurosci* 21(11): 460-68). This approach provides a means to link experimental data regarding specific biological processes to cell function.

The culmination of this forty-year history can be seen in several efforts such as the nationally funded efforts, The Human Brain Project and the Virtual Cell Project. The Human Brain Project is a multi-agency funded multi-site effort to organize and utilize diverse data about the brain and behavior. The Virtual Cell project has developed a framework for organizing, modeling, simulating, and visualizing cell structure and physiology. However, these projects lack an overall ability to link to existing genetic, protein and structural databases. In addition, these projects have not defined procedures for modeling biological systems using information stored in local or distributed databases.

Another approach to understanding the complex interactions of biological systems simulates system kinetics. One such approach is found in the PATHDB™ and GEPASI™ programs available from the National Center for Genome Resources (http://www.ncgr.org). PATHDB™ is a JAVA™ language general metabolic pathway database intended to represent current knowledge of metabolism. The main data types represented by PATHDB™

are compounds, reactions, enzymes and other metabolic proteins and pathways. Similar metabolic pathway databases containing gene sequence data and other biochemical information include EMP and MPW, which are both available from the Argonne National Laboratory Computational Biology Group, http://wit.mcs.anl.gov/EMP and http://wit.mcs.anl.gov/MPW. GEPASI™ is a biochemical kinetics simulator of the dynamics of metabolic pathways. Similar software programs for simulating metabolic pathways include MetaModel (Cornish-Bowden, A. and Hofmeyr, J.H. (1991), Comput. Appl. Biosci., 7, 89-93); SCAMP (Sauro, H.M. (1993), Comput. Appl. Biosci., 9, 441-50); SIMFIT (Holzhutter, H.G. and Colosimo, A. (1990), Comput. Appl. Biosci., 6, 23-28); MIST (Ehlde, M. and Zacchi, G. (1995), Comput. Appl. Biosci., 11, 201-07); and Dbsolve (Goryanin, I. et al. (1999), Bioinformatics, 15, 749-58). While there has been some mention of the desirability of integrating the PATHDB™ and GEPASI™ programs, major modifications, heretofore unachievable, would be required to accommodate such a combination.

What is needed therefore are new computer-based tools for modeling biological pathways and ascertaining their effect on intercellular systems. Such tools will provide a means for linking information at the level of the gene to functional properties of intercellular systems in health and disease, will further the understanding of disease processes, and aid in drug target identification and screening.

## SUMMARY OF THE INVENTION

In accordance with the present invention, there is provided a computerized system for modeling biological pathway, comprising a database having a knowledge database for storing at least one attribute of at least one entity and a pathway database for storing at least one pathway diagram; a user interface for creating, querying, manipulating and viewing data from the database; and a processing system having a modeling module for simulating or analyzing the behavior of at least one attribute in a pathway diagram and a pathway editor module for retrieving, editing, saving and rendering at least one pathway diagram. Advantageously, the system is capable of processing

biological information from relational databases using an object-oriented approach. Accordingly, the present invention provides new tools that enable the transformation of unlimited descriptive biological information into the properties of a definite number of biological objects. This affords system users the ability to work with all relevant biological information, and thereby study and understand complex biological systems.

In another aspect of the invention, there is provided a computer-readable program product comprising database access means for accessing a database having a knowledge database for storing at least one attribute of at least one entity and a pathway database for storing at least one pathway diagram; user interface access for creating, querying, manipulating and viewing of data from the database; at least one machine learning algorithm for generating or selecting a pathway model or parameters in a pathway model; and a pathway editor for retrieving and saving at least one pathway diagram.

In yet another aspect of the invention, there is provided a web-based server system which provides data and applications for modeling biological pathways comprising a database server having access to a database having a knowledge database for storing at least one attribute of at least one entity and a pathway database for storing at least one pathway diagram; a user interface for creating, querying, manipulating and viewing data from the database; and an application server having access to a processing system having a modeling module for simulating or analyzing the behavior of at least one attribute in a pathway diagram and a pathway editor module for retrieving, editing, saving and rendering at least one pathway diagram.

The present invention also provides a computer-implemented method for modeling biological pathways comprising the steps of accessing at least one attribute of an entity from a knowledge database; accessing at least one pathway diagram from a pathwaydatabase; and utilizing a machine learning algorithm for generating or selecting a pathway model to be simulated. Advantageously, the method provided herein may also be used to streamline drug development.

## BRIEF DESCRIPTION OF THE DRAWINGS

The invention will be more fully understood and further advantages will become apparent when reference is made to the following detailed description and the accompanying drawings in which:

FIG. 1 is a block diagram illustrating operation of the present invention;

FIG. 2 is a block diagram detailing a specific operation flow of the present invention;

FIG. 3 is a logic diagram illustrating pathway opening and attribute simulation;

FIG. 4 illustrates data flow for keyword search;

FIG. 5 illustrates data flow for a sequence query;

FIG. 6 illustrates data flow for a data query;

FIG. 7 illustrates data flow for a data entry and edit process;

FIG. 8 illustrates data flow for report generation;

FIG. 9 illustrates data normalization and storage from external resources;

FIG. 10 is a pathway diagram of a biophysical process; and

FIG. 11 is a pathway diagram of T-cell differentiation.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention provides a web-based intelligent database processing system for modeling the highly complex network of biological pathways. The system includes a fully interactive web-based user interface, a database for knowledge storage, and computerized systems and machine learning tools for data analysis, simulation and dynamic graphical representation of the pathways. Some of the underlying features for implementing the present invention are described in detail in parent patent application, U.S. Serial No. 09/499,575, filed February 7, 2000, which is a continuation-in-part of U.S. Serial No. 09/295,503, filed April 21, 1999, both of which are expressly incorporated herein by reference.

In order to understand the logic commands governing implementation of the present invention, a brief description of the salient terms is provided.

A biological entity ("entity") is a particular or discrete unit that is part of, plays a role in, or affects a biological system. Biological entities include any components of a biological system or any objects, elements or molecules that affect biological function. For example, a biological entity may comprise a protein, a cell organelle, or any variable affecting a biological system. As used herein, a variable refers to anything which defines interdependencies in cell processes – for example, elements or ions important to cell function such as $K^+$, $Na^+$, $Ca^+$, $H^+$, organic or inorganic compounds such as ATP, ADP, $P_i$, gases such as $O_2$ and $CO_2$, or any abstracted quantity describing the state of a biochemical or biophysical process, and which relates to organ, tissue, cellular, subcellular, molecular, or genetic function. Entities may also comprise state variables – that is, a set of parameters that allow the calculation of the behavior of the system at a point in time. Each entity may be associated with one or more attributes or properties, such as pH, concentration, temperature, activity or membrane potential.

Pathway diagrams ("pathways") are graphical representations of relationships between and among biological entities or compositions of biological entities. Typically, each node or vertex of a pathway represents a biological entity, and the edges or sides represent a state transition, reaction or causal relationship between the nodes/vertices connected by that edge. For example, as shown in FIG. 11, a simple pathway might represent a simple two-state closed-open model of a cardiac ion channel, thus modeling a biophysical relationship. In this instance, entity $C_1$ corresponds the closed state of an ion channel (a variable), $C_2$ corresponds to the open state of the ion channel (a variable), and additional third and fourth entities would be identical and equal to membrane potential V (variables). The functional dependence of the transition rate constants K12 and K21 on quantities such as temperature, pH, membrane potential, and, in general, variables and/or

proteins as defined previously, on membrane potential may or may not be specified, but the fact that a dependence exists would be.

A more complex pathway may comprise information regarding signal transduction in T-cell differentiation, as shown in FIG. 11. Other complex pathways may detail information pertaining to protein expression in different phases of a cell's existence (e.g., naive (quiescent), activated, and inhibited (for the naive or activated state), viable or apoptotic).

As described in more detail below, the present invention preferably provides at least two logical databases: a knowledge database for storing information relating to properties of the entities (i.e., attributes) in the pathways, and a pathway database for storing information about the pathways in pathway diagrams. Advantageously, the knowledge and pathway information are stored in separate logical databases in order to facilitate easier manipulation and management of such data. The present invention can, however, be implemented using a single logical database to store both the knowledge and pathway information.

Details of the pathway diagrams may be stored in the pathway database in an XML (extensible mark-up language) format. Each pathway diagram is given a unique name or identifier, which allows the pathway to be saved or retrieved from the pathway database. For example, a pathway diagram showing T-cell differentiation may be stored as "T_cell_path1." Detailed examples and description of an XML format for representing pathways are found in the incorporated parent patent applications, U.S. Serial Nos. 09/499,575 and 09/295,503 under the subheading "Computer System."

The knowledge database stores the values of properties or attributes of entities relating to pathways stored in the pathway database. For example, attributes associated with a protein might describe the organism in which the protein is found, the specific cell in which the protein is found, the specific gene coding for the protein, the sequence of the gene coding for the gene and so forth. The attributes may be defined and hierarchically arranged by the user by means of a graphical user interface (GUI).

Each attribute may thus comprise a pointer to a specific portion of the knowledge database where the specific information associated with that

attribute is located. By way of example, the attributes associated with a given protein could be arranged in the following hierarchy: Organism:Organ:Cell:Gene:State:Sequence:Structure:Location:Model. In this example, the attribute "Organism" is a pointer to an organism or organism-type having cells which produce that protein. The attribute "Organ" is a pointer to an organ or tissue-type having cells that produce that protein. The attribute "Cell" points to the specific cell type in which the protein is expressed. The attribute "Gene" is a pointer to the specific gene that codes for the protein. The attribute "State" identifies the state of the Organism:Organ:Cell:Gene system and may be anything that might affect expression of the protein such as an age-related parameter, the presence of a particular disease in the organism, a particular time in the progression of a disease, or the like. The attribute "State" is a pointer identifying a particular subset of the Organism:Organ:Cell:Gene database. The attribute "Sequence" is a pointer to sequence data in the structure of the gene coding for the protein. The attribute "Structure" is a pointer to the three-dimensional structure of the protein coded by that gene, if known. The attribute "Model" is a pointer to the portion of the database in which functional models of the protein coded by that gene are stored. Although reference has been made to protein-related attributes, any information regarding biological entities is within the scope of the present invention. Moreover, the attribute information stored in the knowledge database need not be stored in a hierarchical manner.

This knowledge database is preferably a relational database, thereby providing for advanced searching, data mining, data manipulation and linking capabilities. A more detailed description of a system for storing entity attributes is also found in the incorporated parent patent applications, U.S. Serial Nos. 09/499,575 and 09/295,503. The data to be stored in the knowledge database may be gathered or derived from various sources (including proprietary or public domain sources), such as published literature, public databases, experimental data, and user-provided information and expert opinion.

The system extracts data from the knowledge database and integrates the data with a selected pathway using a computer program referred to as the modeling module. Any number of attributes may be incorporated into a pathway diagram, thus the system may be used to explore the effect of different entities on pathway structure and function. As used herein, modeling is defined in the broadest sense. Models integrate information so as to simulate the function of complete systems. Accordingly, modeling and its associated simulations embody the principle of cause and effect and are based on a simultaneous system of differential equations and associated algebraic equations that define the state variables and rate laws for a particular biological system. Preferably, modeling is effectuated via an object-oriented computer program.

Turning to FIG. 1, the overall architecture of system 10 is illustrated. Preferably, system 10 is implemented in a web-based environment accessible to user 19 via browser 21. However, system 10 may be implemented in any fashion, such as a stand-alone system or in other client-server environments (e.g., local area network, (LAN) intranet, etc.) In this regard, the data storage and other processes employed by system 10 may reside in a single physical location or may be distributed in any manner.

System 10 comprises a web-based graphical user interface system or Web GUI 12. Web GUI 12 comprises web pages that can be dynamically generated and downloaded to user 19 via web server 13. Web server 13 preferably supports server side applets, such as active server pages running on Microsoft Internet Information Server (IIS) (Microsoft Corp., Redmond, WA). This ensures cross-platform compatibility and usability as a browser based Internet application. GUI 12 provides an interface for creating, querying, manipulating and viewing data from database 16. Moreover, the web-based interface provides means for a user to link to third party tools such as data analysis tools (e.g., MATLAB (Mathworks, Inc., Natick, MA)), spreadsheets, plotting tools, graphics programs, artificial neural networks and the like. Advantageously, the data from database 16 may be exported to these third party tools. Similarly, data from third party tools may be integrated into the system platform.

The graphical user interface system 12 provides an intuitive means for inputting, modifying and manipulating data stored in database 16 and for displaying pathway diagrams and simulation results. Pathway diagrams are represented visually as a set of nodes representing biological entities and directed arcs connecting the nodes to each other; an arc pointing from a first node to a second node indicates that the entity represented by the first node influences or affects the entity represented by the second node in some manner. In addition, in a preferred embodiment, an arc may connect a first node to an arc connecting two or more other nodes; this indicates that the entity represented by the first node influences or affects the relationship between the other two nodes. The user is able to construct a pathway diagram using intuitive point-and-click actions and selections from a menu; and the pathway editor module will generate the underlying mathematical relationships between and among various biological entities automatically. Moreover, the user may, by clicking and dragging items (e.g., nodes, arcs) in the pathway diagram, redefine relationships between biological entities and automatically modify the underlying mathematical equations to be simulated.

Simulation results can also displayed by the GUI 12 in an intuitive and easy to understand manner. For example, the color (or some other visual characteristic, such as the size) of a node can correspond to a quantitative measure of a variable associated with that node (e.g., the absolute level or amount of that variable, the rate of change of that variable, the ratio of that variable to some reference value). Different flow rates between nodes can be represented or displayed as different colored arcs connecting the nodes. In simulating the evolution of a pathway over time, the GUI can display the changing colors of the various nodes to provide a simple visualization of the pathway's transient behavior. A color key displayed with the pathway diagram can inform the user of the meaning of the various colors and color changes. In addition, a slider bar can be used to allow the user to replay the time-series simulation or to select a specific point in time to view and/or analyze the values of the pathway diagram nodes.

The graphical representation of a pathway diagram created by the user can be converted into a directed graph ("digraph") representation, which can be stored in the pathway database or in dynamic memory. In such digraph representations, the digraph nodes represent biological entities, and the directed arcs

connecting one node to another indicates that the entity represented by the first node influences or affects the entity represented by the second node in a relationship defined by the pathway model. The digraph representation of a pathway model facilitates the use of various graph-theoretic or network analysis techniques, including determination of articulation points (i.e., nodes or entities that if removed will result in splitting the pathway into two or more disconnected parts); shortest path analysis (i.e., determining the shortest path between two nodes); and parallel/alternate path analysis (i.e., determining alternative paths connecting two nodes). The results of such network analyses can be displayed by highlighting the node or path thereby identified. Digraph representations of pathway models also facilitate the ability to merge two or more pathways sharing one or more common entities or attributes into a single pathway model.

System 10 further comprises processing system 14, which includes computer programs run by application server 15. Processing system 14 provides various operational program modules and machine learning tools that facilitate the access and manipulation of data in database 16. More specifically, processing system 14 includes a modeling module for simulating the effect of at least one attribute of a biological pathway. Processing system 14 further includes a pathway editor module for retrieving and saving pathway diagrams. Preferably, processing system 14 comprises control and data mining modules, namely, object-oriented machine-learning programs written in JAVA™ (a programming language available from Sun Microsystems (Palo Alto, CA)) and/or C++. Advantageously, implementing the system in an object-oriented fashion allows data from database 16 to be retrieved and processed as properties of objects, and provided to web server 13 to generate dynamic web pages for an end user via Web GUI 12.

Database 16 may comprise any type of data storage, and may be implemented by a database server, such as Oracle 8i™ (Oracle Corp., Redwood Shores, CA), or Microsoft SQL™ (Microsoft Corp., Redwood, WA) server. In general, database 16 stores both qualitative and quantitative information about pathway diagrams and modeled biological pathways. The term "simulate" or "simulation," as used herein, is meant to encompass all forms of quantitative or mathematical modeling of a biological system,

including traditional techniques and methods such as time-series analysis and/or calculation of the steady-state values of the variables of interest, as well as the network analysis techniques hereinabove described and more sophisticated quantitative or semi-quantitative modeling methods such as metabolic control analysis and clustering analysis. Information in database 16 may be collected from published or unpublished experimental data, cleaned (i.e., to eliminate duplicate and/or false data) and normalized (i.e., grouped into smaller tables for optimization of query performance), and stored into a relational database format. Pathway models generated by the system may also be stored in database 16.

Database 16 includes at least two database types: a knowledge database and a pathway database. The knowledge database stores properties of entities found in pathways. The pathway database stores information about the pathways, or pathway diagrams. In the knowledge database, each entity may be assigned a unique identification designation that allows the entity to be tracked, or linked to, by information and other databases. Advantageously, the present invention can link to external resources such as compound, metabolic, sequence and structural databases. These may be accessed via the web interface discussed above, and the data gleaned may be associated with each entity and stored in the knowledge database.

It is understood that the various devices, mechanisms and systems described herein may be realized in hardware, software, or a combination of hardware and software. They may be implemented using any type of computer system - or other apparatus adapted for carrying out the methods described herein. A typical combination of hardware and software could be a general-purpose computer system with a computer program that, when loaded and executed, controls the computer system such that it carries out the methods described herein. Alternatively, a specific-use computer, containing specialized hardware for carrying out one or more of the functional tasks of the invention could be utilized. The present invention can also be embedded in a computer program product, which comprises all the features enabling the implementation of the methods and functions described herein, and which, when loaded in a computer system, is able to carry out

these methods and functions. The terms, computer program, software program, program, module, program product, or software, in the present context, mean any expression, in any language, code or notation, of a set of instructions intended to cause a system having an information processing capability to perform a particular function either directly or after one or both of the following: (a) conversion to another language, code or notation; and/or (b) reproduction in a different material form.

FIG. 2 provides a representative operational diagram of system 10. At the front end, web-based user interface 12 provides various functional interfaces that allow for interaction between an end user and system 10. These interfaces include a sequence query 18 which provides sequence analysis; a data query 20 which queries the databases for a specific biological effect; a search interface 22 which browses or searches the databases for a specific biological entity in a pathway; a pathway diagram facility 26 for viewing and creating pathways; a data entry/edit module 28 which allows an end user to annotate the database; and a report facility 24 which provides reports.

Processing system 14 comprises various systems that can access and manipulate data in databases 44 and 46. Representative systems include a data extraction/cleaning module 30 for receiving data from external sources 48 and normalizing the data to a form that is compatible with other data in the databases, a sequence analysis module 32 for obtaining sequence queries and passing them to a search engine module 42, and a query analysis module 34 for receiving data queries and likewise passing them to search engine module 42. Search engine module 42 also includes a module for receiving searches directly from search interface 22. A report generator module 36 is provided for generating reports. Processing system 14 further comprises a modeling module 41 having at least one machine learning mathematical algorithm stored therein for simulating the effect of at least one attribute on a pathway diagram. A pathway editor module 38 is also provided for retrieving and saving at least one pathway diagram. Data processing module 40 is provided to facilitate the process of entering and editing data and creating new pathway diagrams from information gleaned from databases 44 and 46 and external

resources 48. While these systems have been described as residing within processing system 14, they may reside in external systems without affecting system 10 functionality.

Search engine module 42 may be implemented with any known means, including commercial database front-ends with SQL queries, web-based solutions such as Perl scripts and Java-based tools for accessing remote databases, as well as any of a number of cross-platform software tools available, including, for example, tools from Genomica Corp. (Boulder, CO), DoubleTwist, Inc. (Oakland, CA) and NetGenics Inc. (Cleveland, OH).

Turning now to FIG. 3, there is shown a logic flow diagram illustrating the process for opening existing pathways or creating new pathways, and simulating the effect of an attribute thereon. This presents system processing after a user enters the system by conventional means (general login, network login, etc.). More particularly, logic begins at block 37 and branches to blocks 42 and 38 where an existing pathway diagram 26 is accessed, or a new pathway diagram 13 is created in block 48. In the latter case, the system invokes pathway editor module 38, which presents the user with a blank working area that can be used to create a new pathway. The user can access knowledge database 44 for any information useful for new pathway creation. Where the former option is selected, the system presents the user with a series of pathway names from the diagrams stored in pathway database 46. Based upon user input, search engine module 42 retrieves the selected pathway diagram from pathway database 46. The pathway diagram (stored as XML code in pathway database 46) is forwarded to pathway editor module 38, which reconstructs the diagram and displays the pathway to the user.

Once a pathway diagram 26, 48 has been defined, modeling module 51 retrieves at least one attribute of interest from knowledge database 44 via a search engine module (not shown). This data is integrated into the pathway diagram 26, 48 and the appropriate machine learning algorithm(s) invoked to model the biological behavior of interest. Pathway editor module 38 can retrieve a pathway diagram and allow a user to annotate the pathway diagram, access the knowledge database 44 in order to revise the attributes

for simulation, and, where desired, generate a new pathway diagram. This process may be repeated any number of times.

Newly created pathways 48 and pathways generated via simulations 53 may be forwarded for storage in pathway database 46. Where pathway diagrams are created or edited, a file location is allocated in pathway database 46 for revised pathway parameters. These include pathway name, associated entities and pathway function. The user concurrently updates knowledge database 44 with new entities added as a result of the creation or revision of a pathway diagram. The new entities are added into the general entity population in knowledge database 44.

Likewise, the attributes in the knowledge database may be dynamically updated in accordance with the system described herein. In addition, the system can generate a report of each attribute(s) of interest based upon the current information in the knowledge database.

Referring back to the system diagram provided in FIG. 2 for illustration, attention is directed to FIGS. 4-6 where processing logic for locating relevant pathway diagrams in accordance with specific attributes is presented.

FIG. 4 shows the logic associated with a keyword search. After a user enters a keyword search string at keyword search page 50, assemble search query mechanism 52 processes the search string or strings inputted by the user and converts the search string into a format understood by search engine module 42. Search engine module 42 then searches knowledge database 44 for the queried information. Where a positive response is obtained (i.e., the location of the relevant information is identified in the knowledge database), pathway database 46 loads the appropriate pathway 43 to the GUI, and the pathway nodes containing the queried information is highlighted 47.

FIG. 5 provides a logic diagram for querying the system with a sequence, such as a DNA sequence or protein sequence. Here, a user submits a sequence query 18 via a sequence query web page. The query is forwarded to sequence analysis module 32, which parses it into a format understandable by search engine module 42. Search engine module 42 then searches

knowledge database 44 for the sequence, and if located (block 45), the appropriate pathway containing the sequence is loaded to the GUI and the relevant pathway nodes are highlighted 49.

FIG. 6 illustrates the logic flow associated with a query for functional data. In contrast to sequence data (such as the sequence of DNA nucleotides in a gene or the sequence of amino acids in a protein), functional data relate to various biological functions (such as, for example, cell growth, apoptosis or promoter/inhibitor ability), and can take many forms. In accordance with FIG. 6, a user first submits a data query 20 via a data query web page. Query analysis module 34 then converts the query into a format understandable by search engine module 42, which then searches knowledge database 44 for this information. If found (block 47), the appropriate pathway is identified, loaded, and the nodes are highlighted 61.

Continuing now with FIG. 7, a data entry and edit process is illustrated. In this case, a user may enter data via data entry block 54 or pathway diagram 26. Likewise, a user may process data via data entry block 54 or pathway diagram 26. In either instance, data processing module 40 interacts with editing module 56 to provide for editing of any of the entered data. This information is forwarded to search engine module 42, which interfaces with either the knowledge database or, where appropriate, the pathway database (not shown) to retrieve or store the edited information.

In accordance with the present invention, reports may be generated. FIG. 8 illustrates such a process. After a pathway diagram 26 is generated, as shown for example in FIGS. 3-6, it is forwarded to report generator module 36, which compiles the appropriate data and generates report 24.

As previously discussed, system 10 allows users to create links to outside resources. Where this feature is invoked, the Web GUI presents the user with the option to navigate the Internet for additional information pertaining to biological entities. FIG. 9 illustrates that data extraction/cleaning module 30 functions to normalize such data to a compatible format and store it in knowledge database 44.

Advantageously, the present invention provides end users with the ability to build proteins, genes and cellular states into a network of biological pathways; obtain information about the role of a particular protein, gene or biological state in a pathway network; predict and analyze functions of newly discovered proteins; predict the functional consequences of gene knock out/in; find optimal target genes for drug development; and systematically analyze gene chip data.

The system further includes at least one, and preferably a plurality, of machine learning algorithms that automatically adjust the mathematical relational properties between biological entities in the pathways when validated experimental data is provided. This automatically updates related records (i.e., other components in a given biological system) in the database and optimizes the system.

As previously discussed, the machine learning algorithms discussed herein allow the system to ascertain the parameter values that provide the best fit for specific data, determine the model and its parameter values that provide a best fit for specific data, and create new models. Parameter values are readily determined by using non-linear regression techniques such as Gauss-Newton, Lavenberg-Marquardt and Simplex, all of which are well known to the skilled artisan. In addition, evolutionary computation (e.g., genetic algorithms) or miscellaneous stochastic methods (e.g., simulated annealing, general perturbation methods) may be useful when the gradient for an optimization problem cannot be easily computed, or when the surface to be optimized has many local minima, respectively. Models are selected by conducting parameter estimations on a pre-selected group of models and determining which model provides the best data fit. Advantageously, the system can create models and associated parameter values that provide the best fit for the data. In particular, the system creates models via neural networks and genetic programming, both of which are readily known to a skilled artisan.

Neural networks have the desirable property of being "universal approximators" of all reasonable functions and can be used to construct a "black box" model of the data. A neural network can be trained to represent

the entire model. However, even if a neural network is used as a black box, it may be useful as a substitute for an explicit model if the neural network is computationally less expensive than the model. In this case, it is not always necessary to have actual measured data, as the neural network can be trained from data generated by the explicit model that it is trying to mimic.

Alternatively, a neural network can be combined with an existing model in a way such that the neural network is used as a "correction factor" for that model. In this manner, the neural network is used to represent any unknown information or errors not accommodated by the original model. Oftentimes the equations for a model are simplified representations of the known phenomena for computational reasons, so the neural network correction factor can represent errors introduced by these simplifications (e.g. many single cell models do not account for single or multiple binding sites of channel permeation).

Genetic programming evolves a set of equations and their parameters. Genetic programming works directly with predefined operators and operands and evolves programs using these operators/operands. The genetic program output can be readily interpreted in a meaningful way.

Moreover, genetic programming can accommodate *a priori* information in several ways. First, one could start with an initial population of individual programs that consists of equations from existing models. It is also possible to tag certain parts of a program as unchangeable, thereby constraining the solution based on real world knowledge that a subset of equations applies. Finally, one could also specify relationships between operands (e.g., "Equation X always goes with Equation Y").

Since the system, in a preferred embodiment, utilizes XML to represent the components and equations, genetic programming is capable of directly manipulating the XML nodes and attributes during the evolutionary process, with the resulting output XML programs having syntactically correct XML components and equations. Normally, it is difficult and time-consuming to create the code that for translating the problem into a form that can be manipulated by genetic programming. Advantageously, the use of a genetic programming implementation that directly manipulates XML

significantly reduces the amount of code that must be translated, thereby providing a faster, more streamlined process.

By way of example, the following machine learning algorithms may be used to develop a cell model. An important step in the creation of such a model is the determination of the membrane current relationships. The following is a generic equation for a channel current.

$$I_{channel} = p_{open} * I_{open\_channel}$$

where $I_{channel}$ is the channel current, $p_{open}$ is the open probability, and $I_{open\_channel}$ is the open channel current.

In one approach, machine learning can be used to model $I_{channel}$ directly without specifying any of the underlying governing equations. This would be an example of a pure, unconstrained "black box" approach. Alternatively, the form of the underlying equations may be specified, and the values of the relevant variables that determine the values of $p_{open}$ and $I_{open\_channel}$ can be "learned."

Explicitly specifying the open probability and open channel current models is useful when there is *a priori* information about the gating mechanisms of the channel. For example, if the gating mechanisms of a channel are known, one can directly use the $p_{open}$ equation that identifies all the opening and closing rate parameters of the gates. In this manner, the learned rate parameters will be meaningful, and one is constraining the search space with this information (thereby decreasing the computational cost of the search).

There can be varying degrees of specificity or stringency in constraining the open probability -- from the case of using a floating point parameter to represent $p_{open}$ to using a Hodgkin-Huxley model to using a general hidden Markov model (HMM) with numerous transition rate parameters.

Conversely, modeling $I_{channel}$ directly can be useful: (1) when little or nothing is known about the gating mechanisms; (2) when one is not interested in the gating mechanisms; and/or (3) when one wishes to allow the

learning algorithm more freedom to return an empirical solution (e.g., it may come back with a solution that has no clearly identifiable "open probability" parameter).

The information gleaned from modeling individual membrane currents can be used to optimize the parameters of all current and ion concentration equations in a single-cell model to best fit a set of data; return one model with optimized parameters from a set of candidate single-cell models; or return a new model with optimized parameters combining features of several single-cell models selected from a set of initial single-cell models.

In addition, the machine learning algorithms utilized by the present invention provide a means for automatically creating new classifications of objects using some similarity measure based on experimental results or simulated time-series data associated with these objects. This is also useful to finding new relationships that would otherwise not be apparent (e.g., realizing that one object is similar to another). Advantageously, this provides a tool for automatically adding new object types to a database ontology. Use of known approaches such as hidden Markov models (HMMs); stochastic grammars; feedforward and recurrent neural networks; evolutionary computation; and hybrid approaches (e.g., combining neural networks with HMMs to overcome the first order Markov property of HMMs) can be used to create classifiers which can classify new objects into an existing category, or suggest that a new category be added.

The machine learning algorithms discussed herein also provide means for selecting an appropriate cell model from a set of specified cell models. That is, given a set of data, the system can predict what type of model (from a library of models) would be most suitable for the data set (i.e., "best fit" the data set). Such a pattern recognition system can be implemented by first generating an output data set from each model and then computing the correlation between the generated curve and the input data curve.

Having thus described the invention in rather full detail, it will be understood that such detail need not be strictly adhered to, but that various changes and modifications may suggest themselves to one skilled in

the art, all falling within the scope of the present invention as defined by the subjoined claims.

## CLAIMS

What is claimed is:

1. A computer system for modeling biological pathways comprising:

a database having a knowledge database for storing at least one attribute of at least one entity and a pathway database for storing at least one pathway diagram;

a user interface for creating, querying, manipulating and viewing data from the database; and

a processing system having a modeling module for simulating the effect of an attribute on a pathway diagram and a pathway editor module for retrieving, editing and saving at least one pathway diagram.

2. A computer system as recited by claim 1, wherein the database is updateable by the user.

3. A computer system as recited by claim 1, wherein the processing system is capable of performing a time series simulation.

4. A computer system as recited by claim 1, wherein the simulation generated by the processing system includes network analysis.

5. A computer system as recited by claim 1, wherein the simulation generated by the processing system includes metabolic control analysis.

6. A computer system as recited by claim 1, further comprising links to external data sources.

7. A computer system as recited by claim 1, wherein the pathway database and knowledge database are stored in separate logical databases.

8. A computer system as recited by claim 1, wherein the user interface includes a report generation system.

9. A computer system as recited by claim 1, further comprising a data extraction and cleaning module for converting data from the external sources into a compatible data format.

10. A computer system as recited by claim 1, wherein the processing system includes at least one machine learning algorithm for determining the mathematical relationship between at least two biological entities.

11. A computer system as recited by claim 10, wherein related records in the database are updated in accordance with the mathematical relationship determined by the machine learning algorithm.

12. A computer system as recited by claim 1, wherein the processing system includes at least one machine learning algorithm for ascertaining the parameter values that provide the best fit with respect to specific data.

13. A computer system as recited by claim 1, wherein the simulation generated by the modeling module is stored in the pathway database.

14. A computer system as recited by claim 1, wherein the pathway diagram is stored in the pathway database in an extensible markup language.

15. A computer system as recited by claim 1, wherein the pathway editor includes an object-oriented program that handles each attribute as an object.

16. A computer system as recited by claim 1, wherein the user interface includes a system for creating a new pathway model; and

wherein the processing system includes a system for separating a pathway model into pathway diagram details suitable for storage in the pathway database and attributes suitable for in the knowledge database.

17. A computer-readable program product comprising:

a means for accessing a knowledge database storing at least one attribute of at least one entity;

a means for accessing a pathway database storing at least one pathway diagram;

user interface access for creating, querying, manipulating and viewing of data from the database;

at least one machine learning algorithm for generating or selecting a pathway model based upon data stored in the knowledge and pathway databases; and

a pathway editor for retrieving and saving at least pathway diagram.

18. A web-based server system which provides data and applications for modeling biological pathways comprising:

a database server having access to a database having a knowledge database for storing at least one attribute of at least one entity and a pathway database for storing at least one pathway diagram;

a user interface for creating, querying, manipulating and viewing data from the database; and

an application server having access to a processing system having a modeling module for simulating or analyzing the behavior of an attribute in a pathway diagram and a pathway editor module for retrieving and saving at least one simulation generated by the modeling module.

19. A web-based server system as recited by claim 18, further comprising a system for providing active server pages for viewing the simulation generated by the modeling module.

20. A web-based server system as recited by claim 18, further comprising a system for providing active server pages for modifying the simulation generated by the modeling module.

21. A web-based server system as recited by claim 18, further comprising a system for providing active server pages for creating new simulations.

22. A computer-implemented method for modeling biological pathways, comprising the steps of:

accessing at least one attribute of an entity from a knowledge database;

accessing at least one pathway diagram from a pathway diagram; and

utilizing a machine learning algorithm to generate or select a pathway model to be simulated.

23. A computer-implemented method for drug development, comprising the steps of:

accessing at least one attribute of an entity from a knowledge database;

accessing at least one pathway diagram from a pathway diagram; and

utilizing a machine learning algorithm to generate or select a pathway model to be simulated.

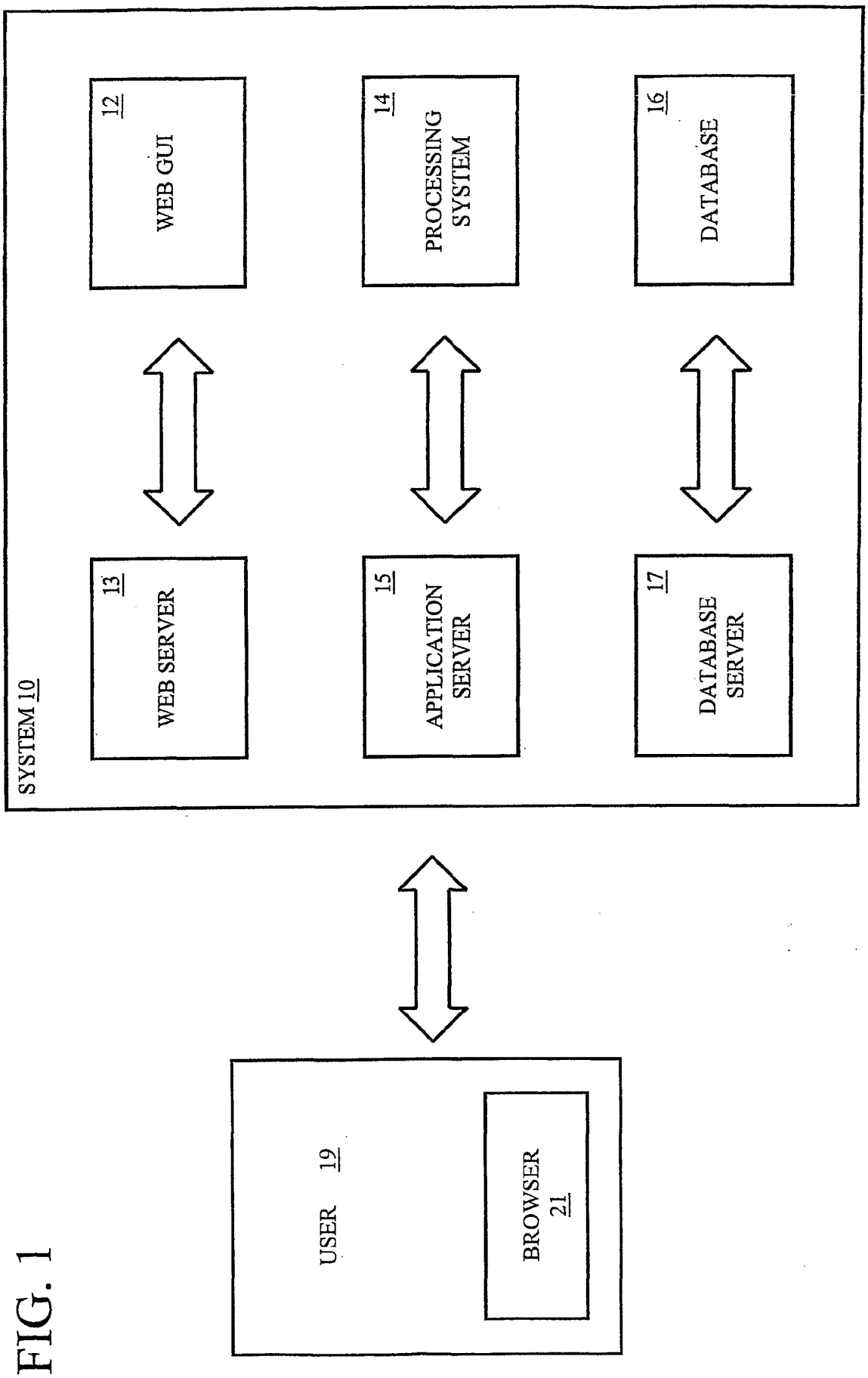

SYSTEM 10
12
WEB GUI
14
PROCESSING SYSTEM
16
DATABASE
13
WEB SERVER
15
APPLICATION SERVER
17
DATABASE SERVER
USER 19
BROWSER 21

FIG. 1

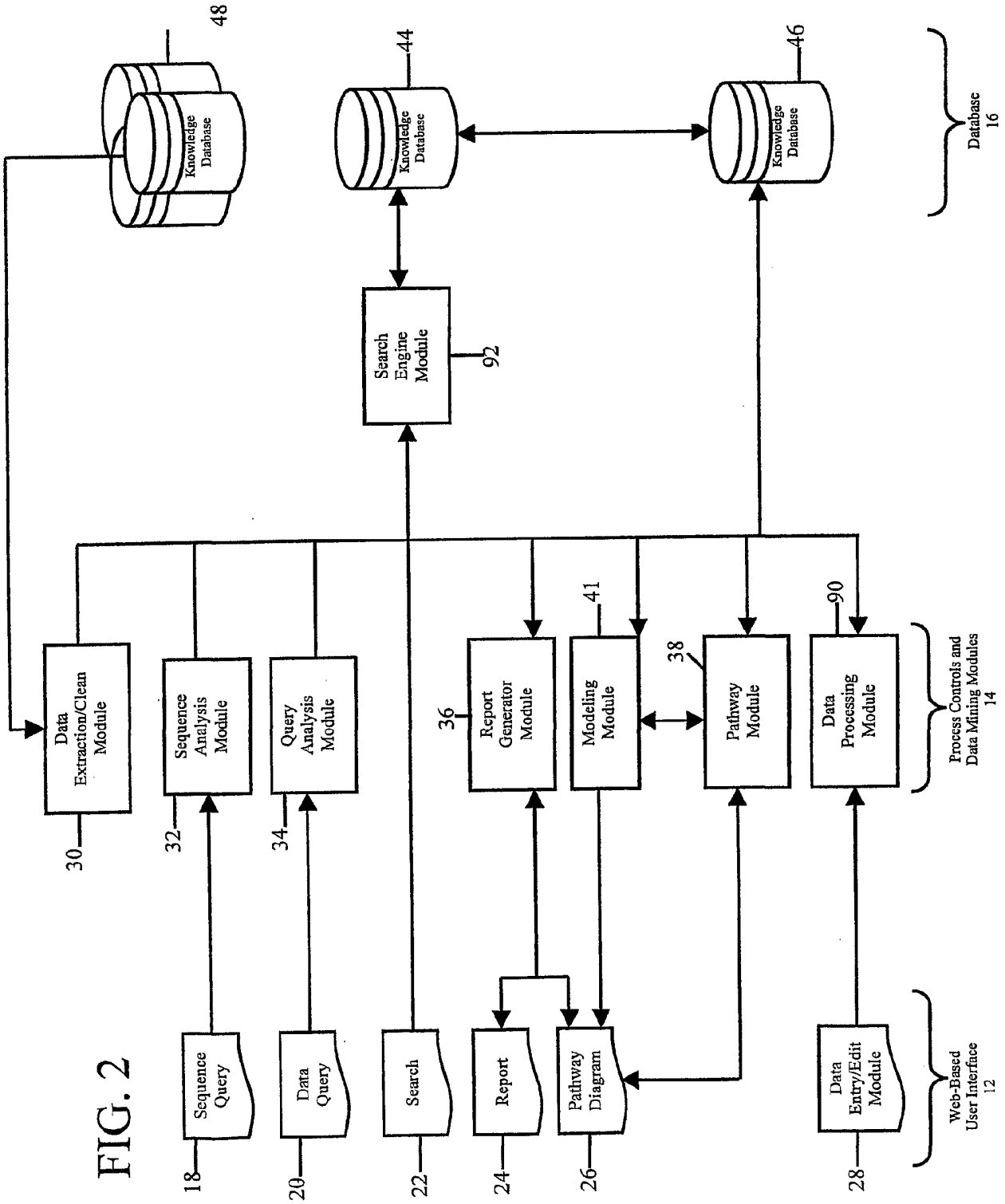
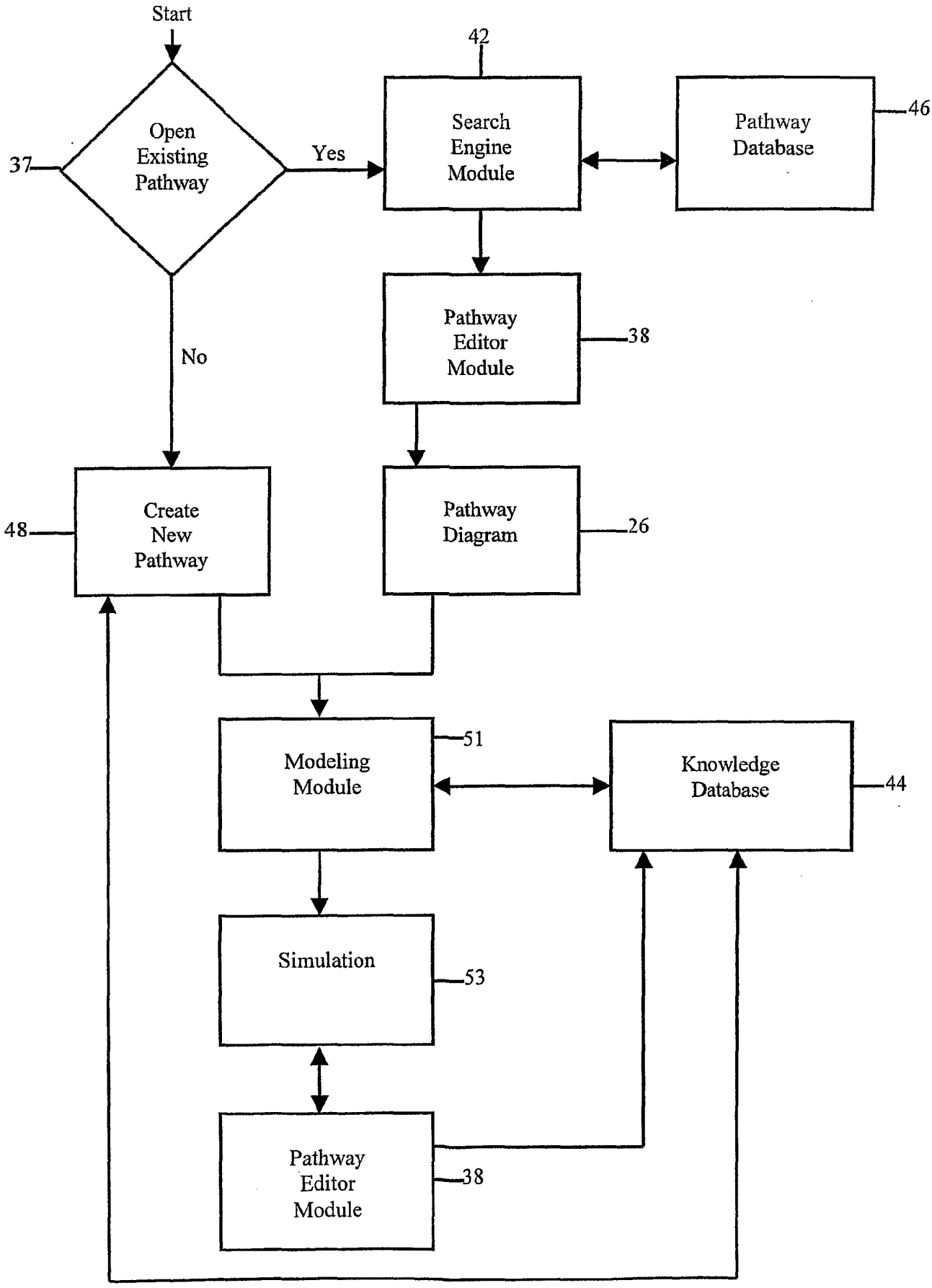

48
Knowledge Database
44
Knowledge Database
46
Knowledge Database
Database 16
Search Engine Module
92
Data Extraction/Clean Module
30
Sequence Analysis Module
32
Query Analysis Module
34
Report Generator Module
36
Modeling Module
41
Pathway Module
38
Data Processing Module
90
Process Controls and Data Mining Modules 14
Sequence Query
18
Data Query
20
Search
22
Report
24
Pathway Diagram
26
Data Entry/Edit Module
28
Web-Based User Interface 12

FIG. 2

FIG. 3


Start
Open Existing Pathway
37
Yes
No
42
Search Engine Module
Pathway Database
46
Pathway Editor Module
38
Create New Pathway
48
Pathway Diagram
26
Modeling Module
51
Knowledge Database
44
Simulation
53
Pathway Editor Module
38

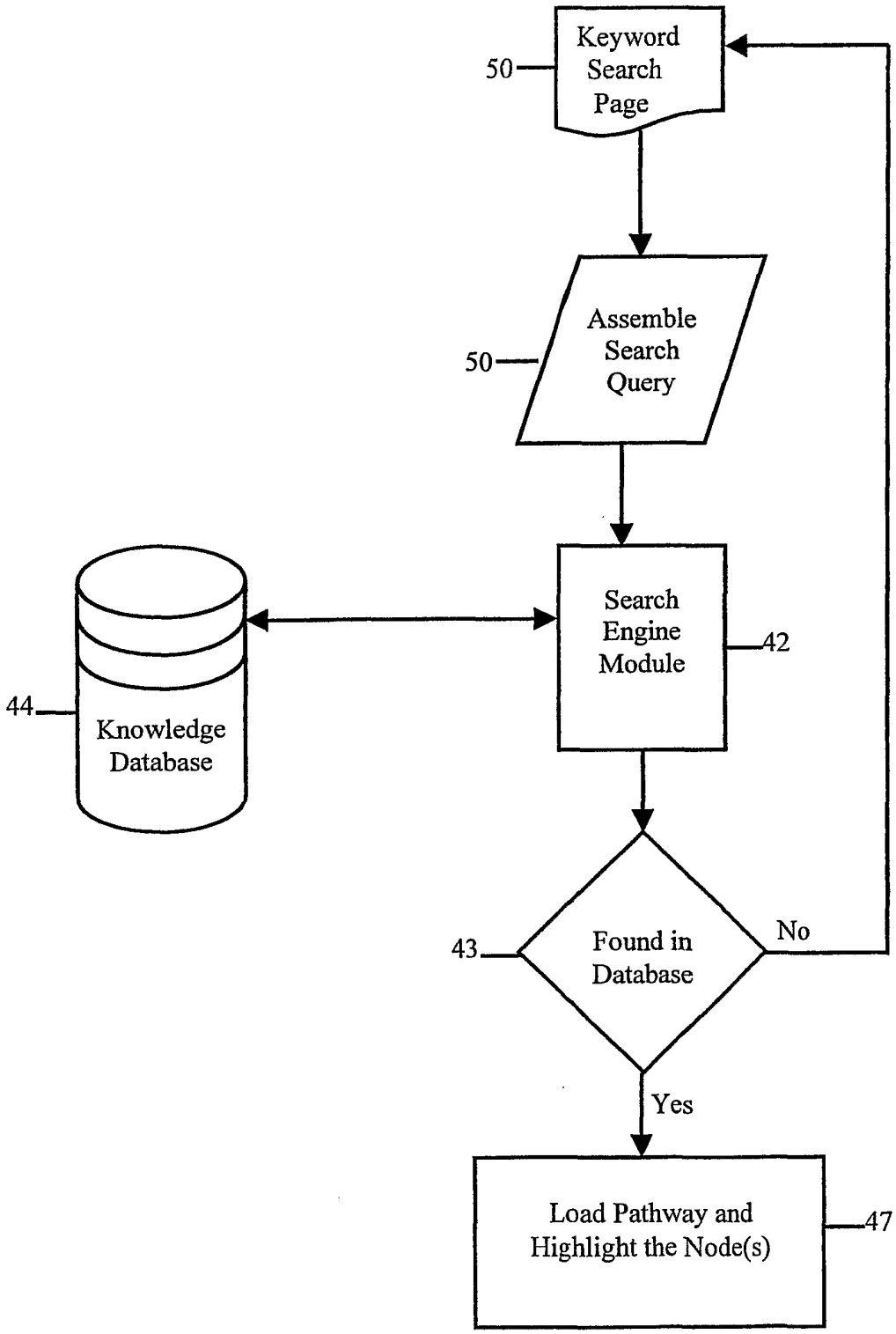

Keyword Search Page
50
Assemble Search Query
50
Search Engine Module
42
Knowledge Database
44
Found in Database
43
No
Yes
Load Pathway and Highlight the Node(s)
47

FIG. 4

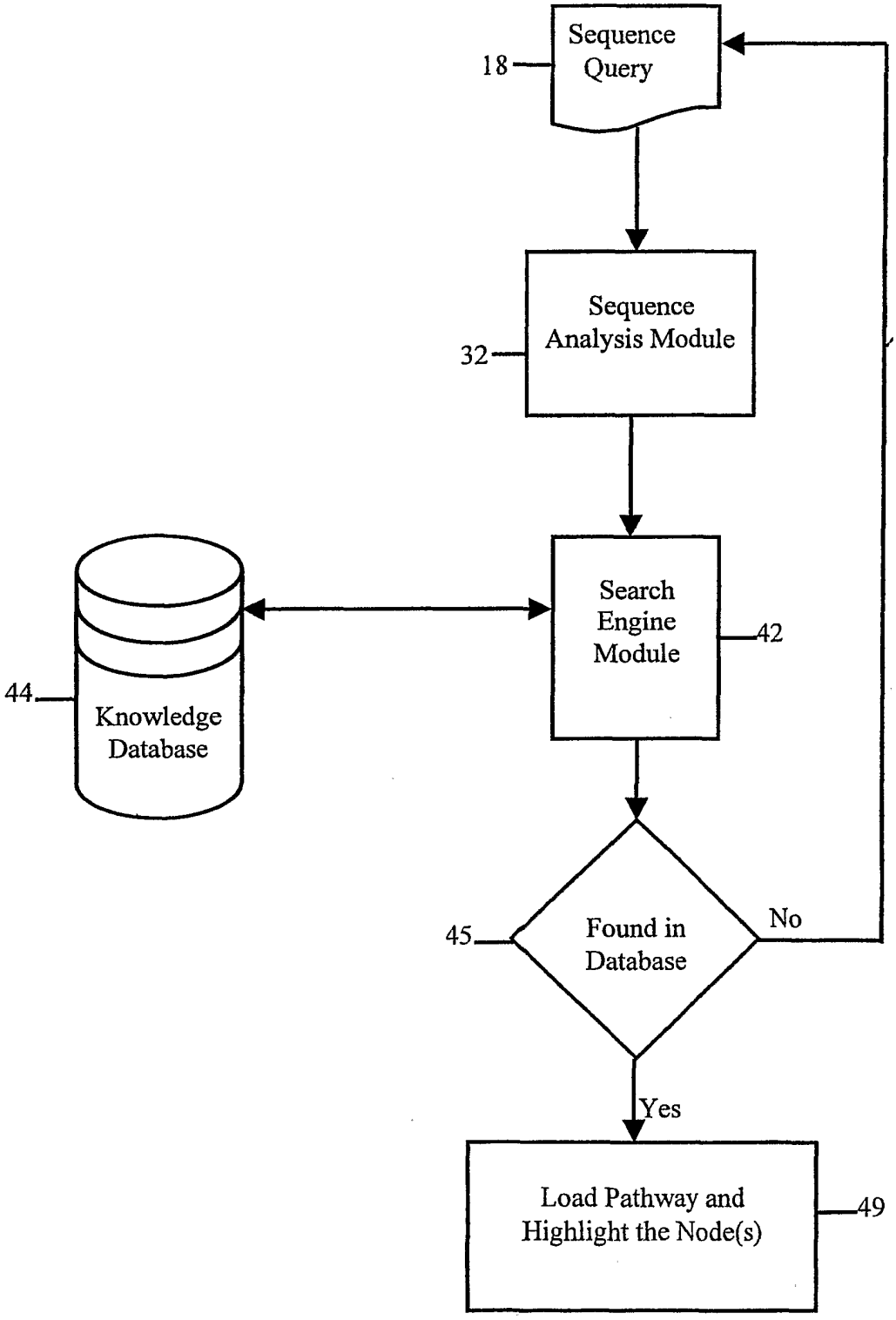

Sequence Query
18
Sequence Analysis Module
32
Search Engine Module
42
Knowledge Database
44
Found in Database
45
No
Yes
Load Pathway and Highlight the Node(s)
49

FIG. 5

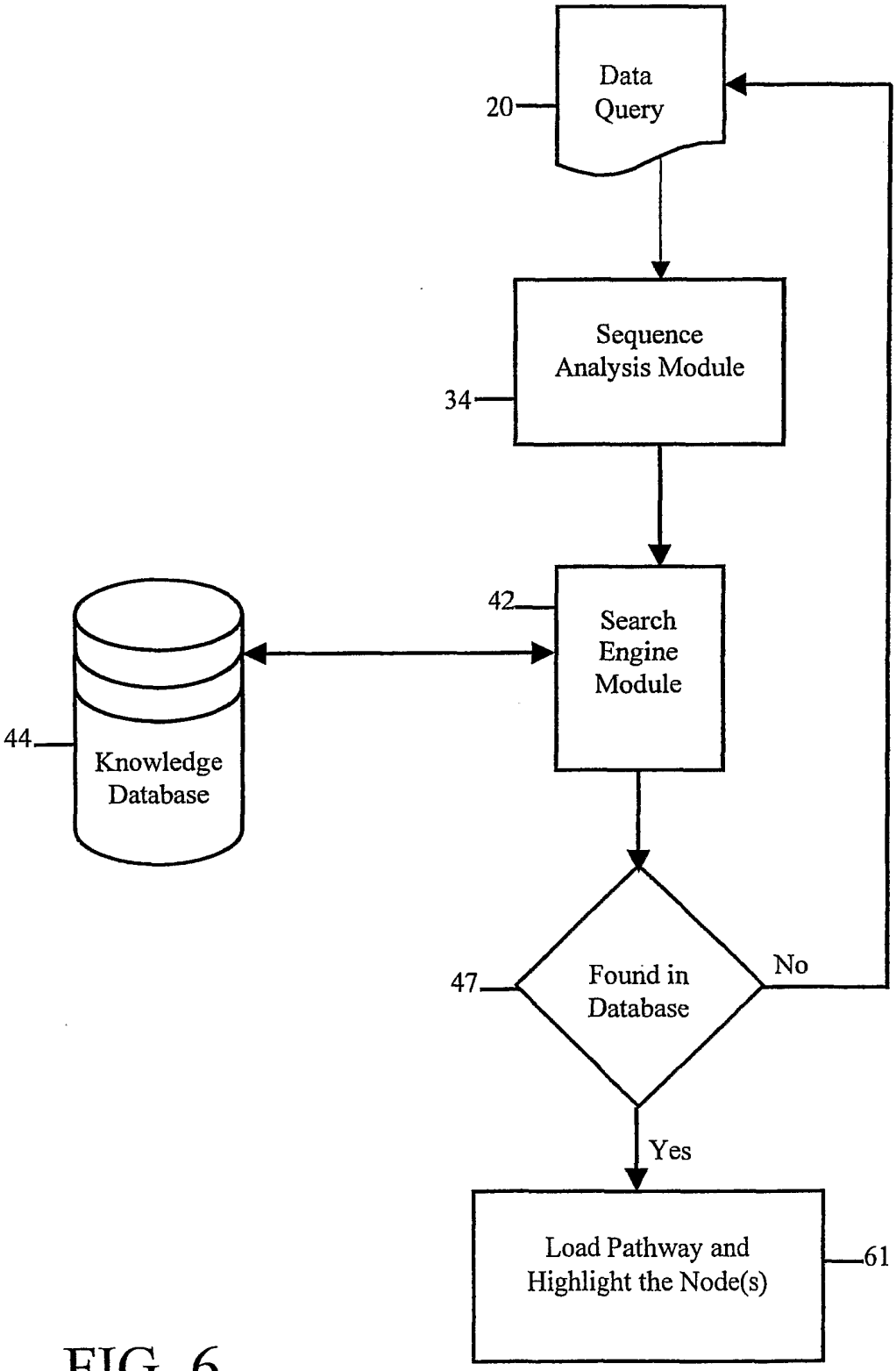

Data Query
20
Sequence Analysis Module
34
42
Search Engine Module
44
Knowledge Database
Found in Database
47
No
Yes
Load Pathway and Highlight the Node(s)
61

FIG. 6

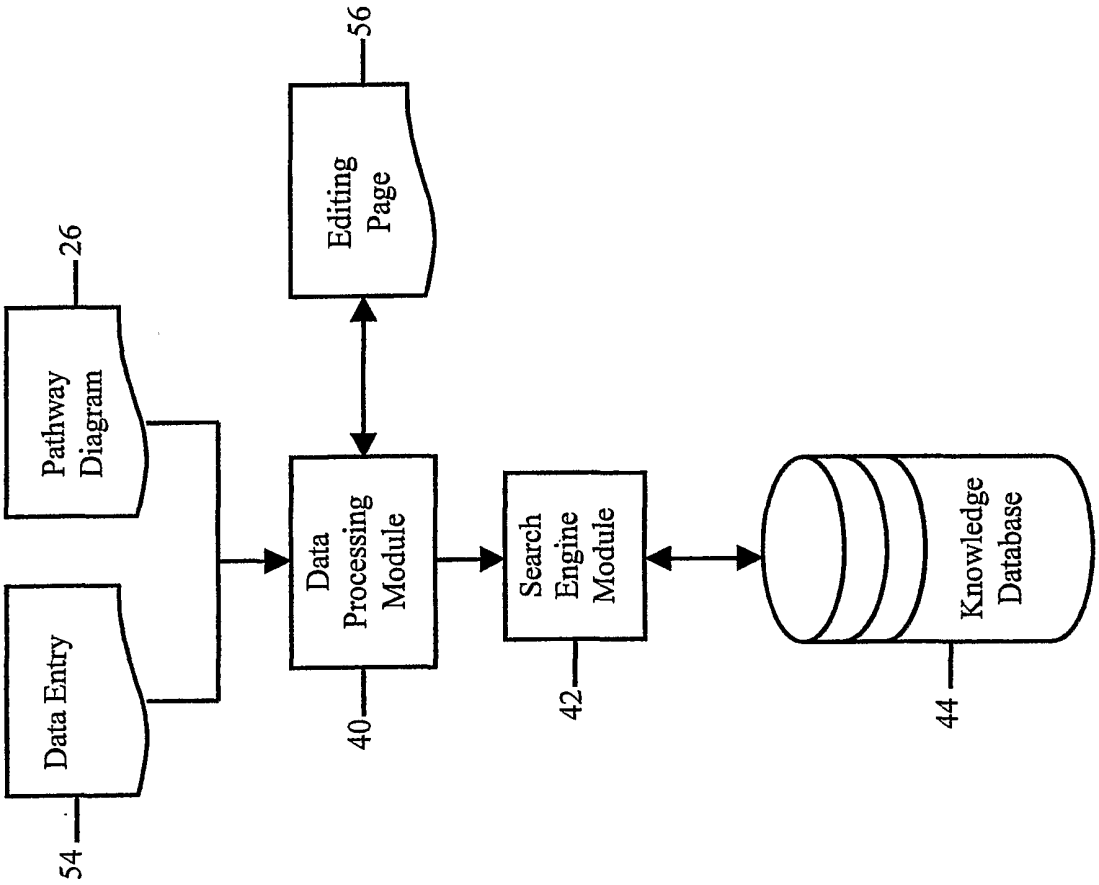

Data Entry
54
Pathway Diagram
26
Data Processing Module
40
Editing Page
56
Search Engine Module
42
Knowledge Database
44

FIG. 7

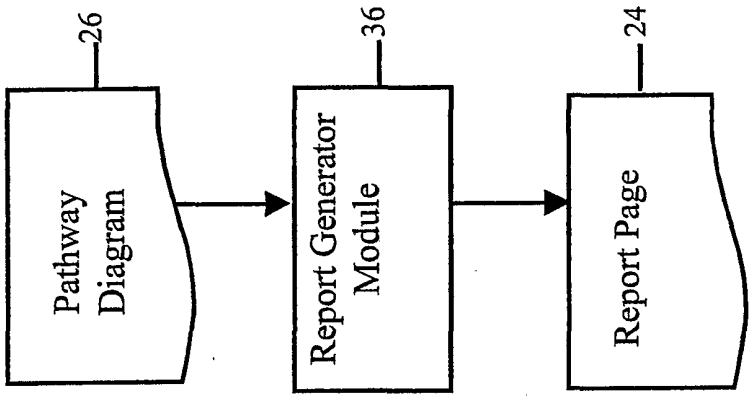

26
Pathway Diagram
36
Report Generator Module
24
Report Page

FIG. 8

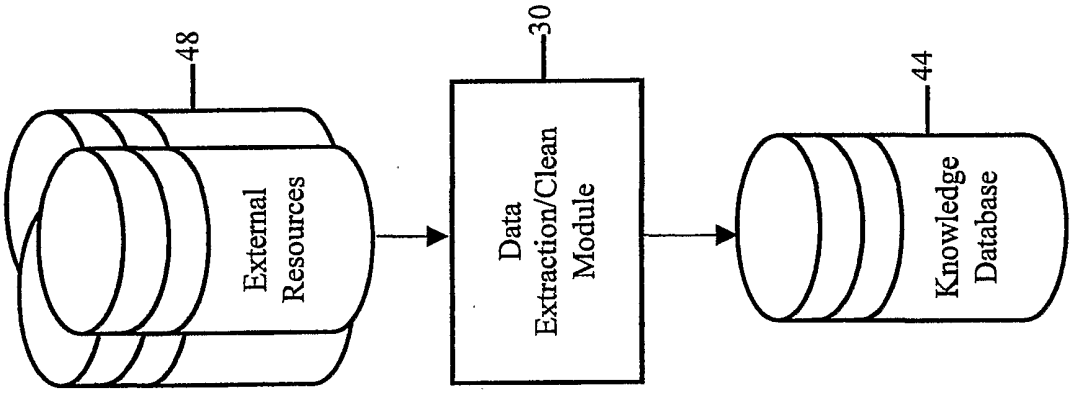
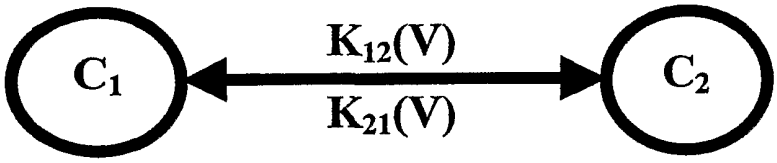
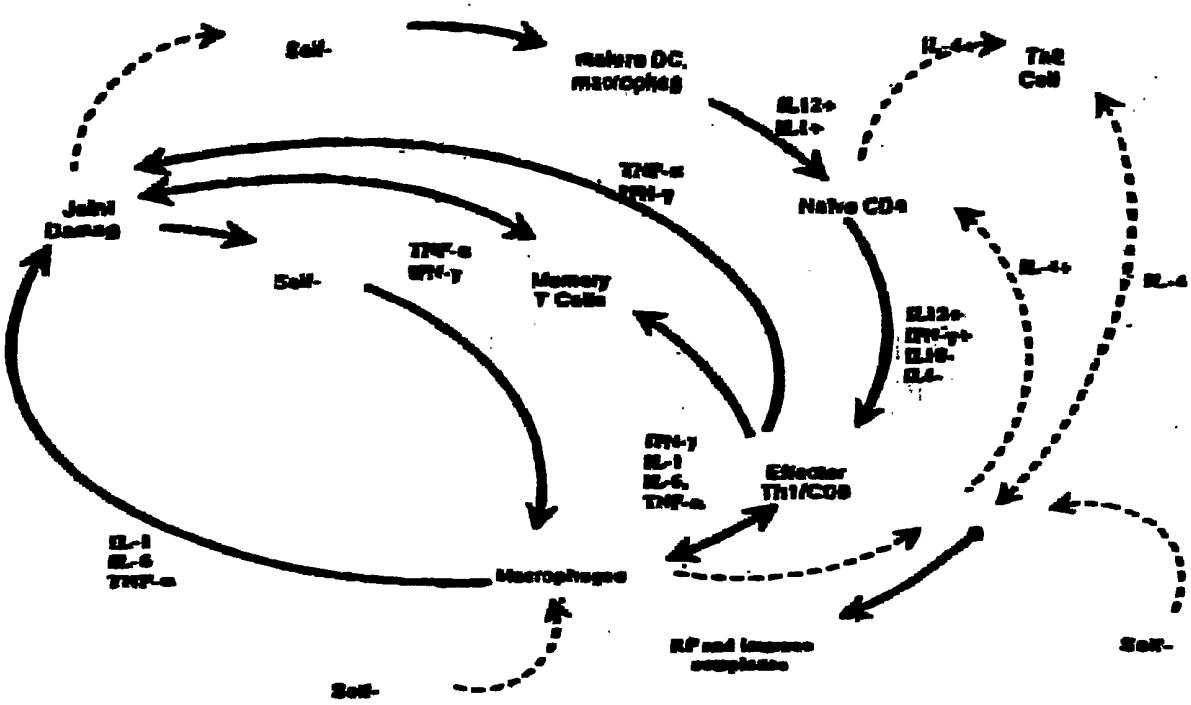

48
External Resources
30
Data Extraction/Clean Module
44
Knowledge Database

FIG. 9

FIG. 10



# FIG. 11


Self-
mature DC,
macrophag
IL12+
IL1+
Th2
Cell
TNF-α
IFN-γ
Joint
Damag
Naïve CD4
TNF-α
IFN-γ
Self-
Memory
T Cells
IL12+
IFN-γ+
IL10-
IL4-
IL-4+
IL-4
IFN-γ
IL-1
IL-6,
TNF-α
Effector
Th1/CD8
IL-1
IL-6
TNF-α
Macrophages
RF and immune
complexes
Self-
Self-