METHOD AND SYSTEM FOR SEARCH OF IMPLICITLY DESCRIBED VIRTUAL LIBRARIES

CROSS-REFERENCES TO RELATED APPLICATIONS

This application claims priority from the following U.S. Provisional Patent Application, the disclosure of which, including all appendices and all attached documents, is incorporated by reference in its entirety for all purposes:

U.S. Provisional Patent Application No. 60/079,750 to Jonathan Greene and John Mount entitled, METHOD AND SYSTEM FOR SEARCH OF IMPLICITLY DESCRIBED VIRTUAL LIBRARIES, filed March 27, 1998.

Further, this application makes reference to the following commonly owned copending U.S. Patent Application, which is incorporated herein in its entirety for all purposes:

U.S. Patent Application Serial No. 09/102,600, in the name of Andrew S. Smellie and Steven L. Teig, entitled, "Method and Apparatus for Conformationally Analyzing Molecular Fragments," filed June 22, 1998.

Further, this application makes reference to U.S. Patent No. 5,307,287.

COPYRIGHT NOTICE

A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

BACKGROUND OF THE INVENTION

The present invention relates generally to the searching for chemical entities with desired physical, chemical or bioactive properties, and specifically to the automated searching of libraries of synthesizable chemical compounds by computer based search and analysis of techniques.

Researchers in the pharmaceutical field, for example, have sought for some time for a way of systematically searching nature for chemical compounds possessing properties which make them ideally suited as medicines. A molecule's structure determines its chemical, physical and bio-active properties. Molecules can have one or more three-dimensional structures. Scientists use a set of convenient parameters, such as bond length, bond angle and torsion angles, to describe the organization of atoms within a molecule that give rise to its molecular structure.

Traditional approaches centered about the chemist in her laboratory using tedious wet chemistry techniques to synthesize chemical compounds, and then perform tests to explore the properties of the compound. The results of these tests were then factored into a new round of synthesis and analysis.

More recently, researchers needing to search for chemical compounds having desirable attributes have turned to computer based methods, rather than subjecting samples of the compounds to chemical analyses in a laboratory. While some of these approaches provide perceived advantages, opportunities to gain further efficiencies and accuracy in the automated search process exist. .

In a commonly owned, copending U.S. Patent Application Serial No. 09/102,600, entitled, "Method and Apparatus for Conformationally Analyzing Molecular Fragments," Smellie and Teig describe a techniques for determining conformations of molecules. A conformation is the spatial arrangement of the atoms in a molecule at any point in time that results from rotation of covalent bonds, bending of bond angles, etc. While this is an important contribution to the field of drug research, there is no method taught for searching a diverse set of chemical compounds for candidates meeting a set of properties, some of which may depend on conformation.

What is needed are techniques for finding compounds having desirable properties by searching and analyzing computer based "libraries" of compound fragments.

SUMMARY OF THE INVENTION

The present invention provides efficient and effective techniques for searching for chemical entities having desired properties. In a particular embodiment according to the present invention, techniques for searching a virtual library of compounds in order to identify component reactants which, when combined, can yield

compounds having a set of desirable properties are provided. Methods and systems according to the present invention enable researchers and scientists to identify promising new chemical compounds in the search for new and better substances.

According to the present invention, techniques including a method for searching a virtual library for compounds of interest are provided. A virtual library can be described implicitly, such as by encoding at least one of a plurality of chemical reactions, each having one or more reactants, enumerating at least one of a plurality of instances of each reactant, and indicating relationships among the reactions and any operational elements. Indications of relationships can comprise in various embodiments, graphical representations, cascade representations and the like. Operational elements can include filters or merges and the like. A searcher describes a hypothesis against which the virtual library can be searched for compounds.

The search process in a particular embodiment comprises a variety of steps, such as a step of enumerating one or more partial products that can be formed from the reactants. A step of determining based upon a potential combinations of partial products that can form compounds matching the hypothesis can also be included in the method. The method can also include a step of determining one or more compound fragments for the partial products. In many embodiments, combinations of compound fragments can be determined using a database join, an intersection operation, and the like. However, alternative embodiments can use other methods of determining fragment combinations that meet the hypothesis. The combination of these steps can provide a method of determining compounds that meet a hypothesis from a virtual library of compound fragments.

In a particular embodiment, a conformational analysis can be performed for the partial products to determine shape of the fragments.

Numerous benefits are achieved by way of the present invention over conventional techniques. The present invention can provide techniques for determining compounds of interest based upon information about fragments without the need to synthesize actual compounds. Further, embodiments according to the present invention can provide techniques for determining compounds of interest based upon information about fragments without the need to create complete models of the compound in a computer. Many embodiments according to the present invention can provide the ability to increase the speed of search by eliminating manipulation of atomic representations or

coordinates. Yet further, some embodiments using the techniques according to the present invention can identify partial fits. Thus, in these embodiments, molecules that fit some but not all of the features of the hypothesis may be identified.

These and other benefits are described throughout the present specification. A further understanding of the nature and advantages of the invention herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1A illustrates a representative client server relationship in accordance with a particular embodiment of the invention;

Fig. 1B illustrates a functional perspective of the representative client server relationship in accordance with a particular embodiment of the invention;

Fig. 1C illustrates an explicitly defined combinatorial library;

Fig. 1D illustrates a representative combination of molecules in an explicitly defined combinatorial library;

Fig. 1E illustrates an implicitly defined combinatorial library;

Figs. 2A-2C depict graphical representations of a virtual library in a particular embodiment of the invention;

Fig. 3A illustrates a representative flowchart of simplified processing in a particular embodiment of the invention;

Fig. 3B illustrates a graphical representation of a fitting of multiple fragments to multiple features in a hypothesis in a particular embodiment according to the invention; and

Fig. 3C illustrates a representative flowchart of simplified search processing in a particular embodiment according to the present invention.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

The present invention provides techniques for searching a virtual library of compounds in order to identify component reactants which, when combined, can yield compounds having desirable properties. Methods and systems according to the present invention enable researchers and scientists to identify promising new chemical compounds in the search for new and better substances.

Embodiments according to the present invention provide methods and systems for locating compounds having desirable bioactive or other attributes by searching libraries of compound fragments for candidates that meet a set of requirements, called a hypothesis. In many embodiments, both the library and the hypothesis can be specified by the searcher prior to search. Hypotheses may be any of a plurality of forms, such as pharmacophores, pseudo-receptor models and the like. A pharmacophore comprises a set of relative positions in space which should be occupied by atoms of a specific type. For further description about hypotheses, such as pharmacophores and pseudo-receptor models, reference may be had to U.S. Patent, Nos. 5,526,281, 5,025,388, 5,307,287; M. Hahn, J. Med. Chem. 1995 V. 38, pp. 2080-2090 and references cited therein; T. Martin, J. Med. Chem. 1992 V. 35 pp. 2145-2154 and references cited therein.

Combinatorial chemical libraries can be used to assist scientists and researchers in the searching for chemical compounds possessing desirable properties. Libraries of compounds can be described explicitly, for example, by enumerating specifically each compound in a database. Searches of such libraries can become computationally expensive as the size of the library increases when each compound is to be examined individually. For example, a combinatorial chemical library, such as a peptide library, formed by enumerating all possible combinations of a set of chemical building blocks, called reactants, can contain millions, billions or more compounds. Search time, and hence cost, increases with the size of the library. Thus, in some embodiments, a virtual library wherein compounds are described implicitly, i.e., comprised of specified building blocks combined in specified ways can be used. For example, a tri-peptide virtual library may be implicitly described as, "all sequences of three amino acids, chosen from a list of 20." By contrast, an explicit description would be a complete list of 20x20x20 = 8000 compounds.

In some embodiments according to the present invention, optimization methods can be used for searching virtual libraries. Optimization methods enumerate a sample of one or more compounds in the library, evaluate these enumerated compounds against the hypothesis, and based upon the result of this evaluation, generate a new sample of compounds from the library targeted to better fitting the hypothesis.

In alternative embodiments, systematic search methods that evaluate fragments of compounds rather than whole compounds, against the hypothesis and then assemble the results of these evaluations can be used. The fragments may be organized

into a tree data structure, with small fragment nodes having as children nodes representing larger fragments that contain the smaller fragments. At the end, leaf nodes of the tree represent complete compounds. Such a tree may be searched in a systematic way, such as depth-first or breadth-first, with unfruitful branches being pruned. In examining the fragment associated with each node of the tree, one may determine conformers of the fragment and poses that fit them to a three-dimensional hypothesis, or other analytical information about the fragment. A conformer is the spatial arrangement of the atoms in a molecule at any point in time that results from rotation of parts of the molecule about covalent bonds and the "bending" of bond angles.

Some embodiments can include intersection search techniques that incorporate the tree search of two or more trees from a common ancestor fragment comprising connected atoms. The results from these searches are combined by an intersection operation.

In a linking technique, small disconnected functional groups involved in binding can be positioned at locations within the receptor model or pharmacophore. Molecular fragments which can link to these groups can then be identified. Linking methods can be useful in performing de novo design. In de novo design techniques, a set of compounds can be built from a list of simple fragments, typically single atoms or rings, without regard to specific reactions. A principle advantage of this approach is that it can produce a practically infinite size library.

In a build-up technique, a set of fragments is identified which can form compounds in the library through the attaching of non-overlapping fragments. Desirable positions of fragments within a receptor model or pharmacophore can be identified. Then adjacent fragments can be attached in order to determine the positions of larger fragments. These steps can be repeated until a molecule having a desirable structure is found.

The method for searching a virtual library of synthesizable compounds for compounds meeting specific criteria in a particular embodiment according to the present invention is implemented in the C++ programming language and is operational on a computer system such as shown in Fig. 1A. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. Many embodiments according to the present invention may be implemented in a client-server environment,

but a client-server environment is not essential. Fig. 1A shows a conventional client-server computer system which includes a server 20 and numerous clients, one of which is shown as client 25. The use of the term "server" is used in the context of the invention, wherein the server receives queries from (typically remote) clients, does substantially all the processing necessary to formulate responses to the queries, and provides these responses to the clients. However, server 20 may itself act in the capacity of a client when it accesses remote databases located at another node acting as a database server.

The hardware configurations are in general standard and will be described only briefly. In accordance with known practice, server 20 includes one or more processors 30 which communicate with a number of peripheral devices via a bus subsystem 32. These peripheral devices typically include a storage subsystem 35, comprised of memory subsystem 35a and file storage subsystem 35b, which hold computer programs (e.g., code or instructions) and data, set of user interface input and output devices 37, and an interface to outside networks, which may employ Ethernet, Token Ring, ATM, IEEE 802.3, ITU X.25, Serial Link Internet Protocol (SLIP) or the public switched telephone network. This interface is shown schematically as a "Network Interface" block 40. It is coupled to corresponding interface devices in client computers via a network connection 45.

Client 25 has the same general configuration, although typically with less storage and processing capability. Thus, while the client computer could be a terminal or a low-end personal computer, the server computer is generally a high-end workstation or mainframe, such as a SUN SPARC™ server. Corresponding elements and subsystems in the client computer are shown with corresponding, but primed, reference numerals.

The user interface input devices typically includes a keyboard and may further include a pointing device and a scanner. The pointing device may be an indirect pointing device such as a mouse, trackball, touch pad, or graphics tablet, or a direct pointing device such as a touch screen incorporated into the display. Other types of user interface input devices, such as voice recognition systems, are also possible.

The user interface output devices typically include a printer and a display subsystem, which includes a display controller and a display device coupled to the controller. The display device may be a cathode ray tube (CRT), a flat-panel device such as a liquid crystal display (LCD), or a projection device. The display controller provides control signals to the display device and normally includes a display memory for storing

the pixels that appear on the display device. The display subsystem may also provide non-visual display such as audio output.

The memory subsystem typically includes a number of memories including a main random access memory (RAM) for storage of instructions and data during program execution and a read only memory (ROM) in which fixed instructions are stored. In the case of Macintosh-compatible personal computers the ROM would include portions of the operating system; in the case of IBM-compatible personal computers, this would include the BIOS (basic input/output system).

The file storage subsystem provides persistent (non-volatile) storage for program and data files, and typically includes at least one hard disk drive and at least one floppy disk drive (with associated removable media). There may also be other devices such as a CD-ROM drive and optical drives (all with their associate removable media). Additionally, the computer system may include drives of the type with removable media cartridges. The removable media cartridges may, for example be hard disk cartridges, such as those marketed by Syquest and others, and flexible disk cartridges, such as those marketed by Iomega. One or more of the drives may be located at a remote location, such as in a server on a local area network or at a site of the Internet's World Wide Web.

In this context, the term "bus subsystem" is used generically so as to include any mechanism for letting the various components and subsystems communicate with each other as intended. With the exception of the input devices and the display, the other components need not be at the same physical location. Thus, for example, portions of the file storage system could be connected via various local-area or wide-area network media, including telephone lines. Similarly, the input devices and display need not be at the same location as the processor, although it is anticipated that the present invention will most often be implemented in the context of PCs and workstations.

Bus subsystem 32 is shown schematically as a single bus, but a typical system has a number of buses such as a local bus and one or more expansion buses (e.g., ADB, SCSI, ISA, EISA, MCA, NuBus, or PCI), as well as serial and parallel ports. Network connections are usually established through a device such as a network adapter on one of these expansion buses or a modem on a serial port. The client computer may be a desktop system or a portable system.

Fig. 1B is a functional diagram of the computer system of Fig. 1A. This diagram is merely an illustration and should not limit the scope of the claims herein. One

of ordinary skill in the art would recognize other variations, modifications, and alternatives. Fig. 1B illustrates a server 20, and a representative client 25 of a multiplicity of clients which may interact with the server 20 via the internet 45 or any other communications method. Blocks to the right of the server are indicative of the processing components and functions which occur in the server's program and data storage indicated by block 35a in Fig. 1A. A TCP/IP "stack" 44 works in conjunction with Operating System 42 to communicate with processes over a network or serial connection attaching Server 20 to internet 45. Web server software 46 executes concurrently and cooperatively with other processes in server 20 to make data objects 50 and 51 available to requesting clients. A Common Gateway Interface (CGI) script 55 enables information from user clients to be acted upon by web server 46, or other processes within server 20. Responses to client queries may be returned to the clients in the form of a Hypertext Markup Language (HTML) document outputs which are then communicated via internet 45 back to the user.

Client 25 in Fig. 1B possesses software implementing functional processes operatively disposed in its program and data storage as indicated by block 35a' in Fig. 1A. TCP/IP stack 44', works in conjunction with Operating System 42' to communicate with processes over a network or serial connection attaching Client 25 to internet 45. Software implementing the function of a web browser 46' executes concurrently and cooperatively with other processes in client 25 to make requests of server 20 for data objects 50 and 51. The user of the client may interact via the web browser 46' to make such queries of the server 20 via internet 45 and to view responses from the server 20 via internet 45 on the web browser 46'.

Fig. 1C illustrates a representative diagram of a simplified explicitly defined 3 x 3 combinatorial library 100, which can reside in system memory subsystem 35a and/or file storage subsystem 35b of Fig. 1A. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. A virtual library of compounds can include compounds that in theory are synthesizable, but typically have not yet been synthesized. Other virtual libraries can be built that can include, for example, known synthesizable compounds, or known non-synthesizable compounds without departing from the scope of the present invention. Some virtual libraries may be described in an explicit fashion by listing each compound in the library, while other

virtual libraries may be described implicitly. Representative combinatorial library 100 has been defined by possible combinations of three molecules arranged in rows with three molecules arranged in columns, giving rise to a tabular format. Combinatorial library 100 includes row molecules, such as a first molecule 102, a second molecule 104 and a third molecule 106. Other and different molecules can be included as row molecules in some embodiments. These row molecules can be combined with molecules arranged in columns of combinatorial library 100, including a first molecule 108, a second molecule 110 and a third molecule 112. Other and different molecules can be included as column molecules in some embodiments. For example, molecule 114 in combinatorial library 100 can be formed by a reaction of row molecule 102 and column molecule 108. Similarly, molecule 116 can be formed by reacting molecule 102 and molecule 110. In this manner, members of the combinatorial library can be explicitly enumerated. Each member can be derived from a combination of a row and a column molecule. The foregoing description is intended to be merely illustrative and not restrictive. Molecules 102, 104, 106, 108, 110, 112, 114 and 116 are merely examples of some of the many types of molecules and reactants that can be used to specify a combinatorial library, such as library 100 in a particular embodiment. Other reactions can be used without departing from the scope of the present invention.

Fig. 1D illustrates one such combination of a row molecule and a column molecule such as described in Fig. 1C, to produce a resultant molecule in the library. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. Fig. 1D illustrates a molecule 120 being combined with a molecule 122 to form a composite molecule 124, which then can be fit to the features of a hypothesis. Molecule 120, molecule 122 and molecule 124 are merely examples of some of many reactants and molecules that can be used to specify one or more libraries in this particular embodiment. Other molecules can be used without departing from the scope of the present invention.

Fig. 1E illustrates a simplified diagram of a representative implicitly defined combinatorial library 101 in a particular embodiment according to the present invention. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. As in library 100 depicted in Fig. 1C, library 101 is

defined across molecules arranged along rows and columns. In this particular embodiment, molecule 130, molecule 132 and molecule 134, are arranged across rows and molecule 136, molecule 138 and molecule 140 are arranged across columns. Molecules defined by the combination of these row and column molecules need not be enumerated explicitly. The combination of molecules, such as a row molecule 132 with a columnar molecule, such as molecule 140, can be defined by means of one or more chemical reactions, such as a first reaction 150, which is a reductive animation reaction. Reaction 150 is merely an example of one of many reactions that can be used to specify one or more molecules in library 101 according to this particular embodiment. Other reactions can be used without departing from the scope of the present invention. In this particular example, reaction 150 is followed by another reaction, a deprotect reaction 152. Reaction 152 is merely an example of one of many reactions that can be used to specify one or more molecules in library 101 in this particular embodiment according to the present invention. Other reactions can be used without departing from the scope of the present invention. Thus, in this particular example, the contents of this 3 X 3 library can be defined implicitly by its columnar and row inputs and the reactions upon these inputs which produce various outputs.

Practical virtual libraries comprise complex combinations of information about reactions, merge operations and filter operations and other similar operations. Fig. 2A illustrates a representative example virtual library 201 in a particular embodiment according to the present invention. Virtual library 201 can reside in storage system 35' of server 20, for example. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. Virtual library 201 comprises a first set of intermediates that can be produced by a first reaction 202 from an instance list A 216 and an instance list B 218 and a second set of intermediates that can be produced by a reaction 204 from an instance list C 220 and an instance list D 222 which can be input into a merge operation 206. The result of the merge can be the union of the compounds from its inputs. These can be passed through a first filter 208. Filters can be used to select a subset of the compounds provided at their input. Filters may select molecules based on size, substructures such as those that are toxic or reactive, a diverse or informative subset, as well as those fitting one or more hypotheses, such as the hypotheses having forms as described herein. The output of filter 208 becomes a reactant along with an instance list

E 224 in a third reaction 210 and a reactant along with an instance list F 226 in a fourth reaction 212. A result of reaction 210 and a result of reaction 212 can be merged by a second merge 214. It is noteworthy that according to the invention no particular reactions need actually take place, but rather these reactions can be incorporated into a large database comprising the virtual library. However, preparing one or more compounds by conducting actual reactions does not depart from the scope of the present invention. It is a novel aspect of the method described in this particular embodiment that it provides the capability to search virtual libraries comprising a filter, such as filter 208, followed by further reactions, such as reaction 210 and reaction 212.

Fig. 2B illustrates a representative diagram of another example virtual library 203. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. Virtual library 203 comprises a first set of intermediates produced by a first reaction 232 from an instance list A 242 and an instance list B 244 and a second set of intermediates produced by a reaction 234 from an instance list C 246 and an instance list D 248 which may be input to a merge operation 236. The output of merge 236 becomes a reactant along with an instance list E 250 in a third reaction 238. The results of third reaction 238 are input to a filter 240. It is a novel aspect of the method described by this embodiment that it provides the capability to search virtual libraries comprising a merge, such as merge 236, followed by further reactions, such as reaction 238.

Fig. 2C illustrates a representative diagram of a yet further example virtual library 205. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. Virtual library 205 comprises a first reaction 252 from an instance list A 260 and an instance list B 262. The intermediate formed by reaction 252 becomes a reactant along with an instance list E 264 in a second reaction 254 and a reactant along with an instance list F 266 in a third reaction 256. The results of reaction 254 and reaction 256 are merged by a merge 258. It is a novel aspect of the method described by this embodiment that it provides the capability to search virtual libraries comprising common intermediates, such as the intermediate formed by reaction 252, subject to two or more alternative reactions, such as reaction 254 and reaction 256.

Fig. 3A illustrates a representative flow chart 303 of simplified processing steps in describing a virtual library, such as virtual library 201 of Fig. 2A. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. In a first step 310, chemical reactions are encoded. Then, in a step 312, reactants in the library are enumerated. Then, in a step 314, one or more relationships among the reactions, reactants and operational elements are specified to the system, such as by a cascade 203 of Fig. 2B, for example. Some embodiments can use other techniques to specify such relationships, such as graphs, spreadsheets, tables and the like without departing from the scope of the present invention. Then, in a step 316, a hypothesis is described to the system. The details of the specific processing of each of these steps will be described below. It is noteworthy that a presently preferable embodiment according to the present invention is not limited to creating possible compound fragments in the library prior to search.

Fig. 2A illustrates a representative cascade description 201 of an example virtual library in one particular format suitable for input into a search program. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. The cascade description of Fig. 2A illustrates reaction (synthesis), filtering and merge operations in a graphical representation. Other ways of representing a virtual library can be used rather than the cascade representation in various embodiments without departing from the scope of the present invention. For example, operations can be represented to a computer in plurality of ways, such as a listing of nodes or operations, graphical representations, charts, spreadsheets and the like. Such representations can comprise connections that indicate relationships between nodes and one or more parameters for each operation. For example, names of reactants, hypotheses, filter constraints and the like, can be specified. The cascade description of a presently preferable embodiment can incorporate a hierarchical arrangement.

Reactions comprising the virtual library can be described in a computer-readable form. Methods for encoding reactions are well known in the art and an example is given here to be illustrative rather than limiting. A reaction description comprises a substructure search query for each reactant and a transformation diagram. The substructure search query can identify a relevant chemical functional group in a valid

instance of the reactant. A transformation diagram comprises a list of operations indicating which atoms are deleted or added and which bonds are made, broken or changed. Such encoding is further described in "Daylight Toolkit Theory Manual," Daylight Chemical Information Systems, Santa Fe, NM; "Myriad Users Manual," Afferent Systems Inc., San Francisco, CA, the entire contents of which are incorporated herein by reference for all purposes.

A reactant comprises a component of a generic reaction. For example, in a peptide bond formation coupling an acid and an amine, there are two reactants, the acid and the amine. Each particular acid or amine which may be used as a reactant is referred to as an instance of the reactant. For each primary reactant in the cascade, a list of valid instances can be specified. The lists are shown as Instance Lists A-F in the example of Fig. 2A. In select embodiments, these lists may take the form of disk-resident files in one or more standard formats such as SMILES or MOL files, or chemical databases. Other formats can also be used without departing from the scope of the present invention. Reference may be had for further description of standard formats to "Daylight Toolkit Theory Manual," Daylight Chemical Information Systems, Santa Fe, NM; "CTFile Formats," June 1997, MDL Information Systems, Inc., San Leandro, CA., the entire contents of which are incorporated herein by reference for all purposes.

A hypothesis can comprise a structure-activity model that can provide information about a molecule's biological activity or other property based upon the molecule's two-dimensional (connectivity) or three-dimensional (conformational) structure, or other properties. The hypothesis may be one of many forms, including the following; a receptor model from a crystal structure, a pseudo receptor model inferred from structure activity data, a three-dimensional pharmacophore possibly with excluded volumes, two-dimensional or three-dimensional similarity to a reference compound, a comparative molecular field analysis ("COMFA") or similar model or any combination of the above. Comparative molecular field analysis techniques well known in the art include the technique described in U.S. Patent No. 5,307,287. Other types of hypotheses can also be used many embodiments without departing from the scope of the present invention. Embodiments having hypotheses based on three-dimensional structures of compounds can provide searches of potential conformations of each molecule. In some embodiments, possible poses of the conformer (its alignment to the hypothesis) can be considered as well.

Fig. 3B illustrates a simplified flow diagram of a generalized representative search process of a virtual library, such as virtual library 201 of Fig. 2A in a particular embodiment according to the present invention. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. In a first step A, hypothesis 301, which in this particular embodiment is a pseudo receptor, is defined as having a first feature 302 and a second feature 304. Next, in a step B, a molecular fragment 306 is found which fits first feature 302 of hypothesis 301. In a step C, a second molecular fragment 308 is found which fits second feature 304 of hypothesis 301. Next, in a step D, a determination that molecular fragments 306 and 308 are consistent and are likely to overlap to form a complete molecule that fits the hypothesis is performed. Then in a step E, the complete molecule 309 is found to simultaneously fit both first feature 302 and second feature 304 of hypothesis 301. Fragments that overlap by one or more bonds can be effective for reducing the number of conformers to be considered because such overlapping fragments can fit a plurality of portions of the hypothesis, such as first feature 302 and second feature 304 of hypothesis 301 as illustrated by Fig. 3B.

Fig. 3C illustrates a representative flowchart 305 of the simplified processing steps in searching a virtual library, such as virtual library 201 of Fig. 2A in a particular embodiment according to the present invention. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art would recognize other variations, modifications, and alternatives. In a first step 320, a list of prototypes for each reactant is created. Then, in a step 321, a list of prototype products is formed. In step 322 of Fig 3C, the fragments in the virtual library are enumerated. In a presently preferable embodiment, step 322, enumerating fragments in the virtual library can comprise forming a partial product for each reactant based upon the prototype products determined in step 321. Then, in some embodiments, an optional step 324 of performing a conformational analysis on a partial product formed in step 322 can be included. Next, in a step 325, fragments fitting the hypothesis are enumerated. In a presently preferable embodiment, database tables containing fragments fitting the hypothesis are formed. Next, in a decisional step 326, a determination is made whether there are further conformers to process. If this is so, then processing returns to step 324. Otherwise, in a decisional step 327, a determination is made whether there are further partial products to process. If this is so, then processing returns to step 322 to process the

next partial product. Otherwise, in a step 328, combinations of fragments that meet the hypothesis are determined. In a presently preferable embodiment, a join operation on the database tables formed in step 325 is performed to form a list of candidate compounds. The order of these steps is illustrative of a particular embodiment, but is not requisite to carry on the invention. Thus, these steps may be re-ordered or combined without deviating from the invention.

A prototype is a smallest possible instance meeting the requirements for the reactant. For example, if the reaction requires an acid, HCOOH would be a suitable prototype. A plurality of prototypes may be used for a reactant to describe a plurality of instances to a sufficient detail. Prototypes may be specified for each reactant manually, or they may be generated automatically from a list of instances by a breadth first search of limited depth from the key functional group (e.g. the COOH in the acid list) or other similar method. An illustrative example is given in the following pseudocode:

```
// Enumerate prototypes for set of instances of a given reactant
// and a given depth B.
Set PrototypeList to empty set.
For each instance I of reactant {
    Find atoms FG in I forming the functional group(s) playing a role in the relevant reactions.
    Find all other neighboring atoms N within B bonds of atoms FG by breadth-first search.
    Construct fragment of instance I composed of atoms FG and N.
    Add fragment to PrototypeList if it is not already in it
            }
```

A list of prototype products useful in labeling partial products is formed according to the pseudocode in Table 1.

```
// Make a list of prototype products (to be used below for labeling atoms) PrototypeProductList = empty
For each combination of prototypes, one for each reactant {
        React prototypes according to cascade to form prototype product.
        Add prototype product to PrototypeProductList.
        }
```

Table 1

A partial product is a compound formed by instances of one or more reactants and prototypes of the remaining reactants. In a presently preferred embodiment, a fragment is a partial product with an instance of one reactant and prototypes for the rest. In an alternative embodiment, a fragment could be a partial product formed by instances of two or more but not all reactants, and prototypes of the remaining reactants.

In another alternative embodiment, the fragments may be defined without requiring a one-to-one correspondence with partial products when fragments can be enumerated without enumerating whole compounds. For example, consider chains of length 6 atoms or less that are contained within one or more partial products.

In preferable embodiments of the present invention, fragments are such that a conformational analysis of a fragment in isolation includes conformers of the fragment which occur when the fragment is put in a larger molecular context. Characteristics of such fragments and ways of selecting them are described in the commonly owned copending U.S. Patent Application No. 09/102,600, incorporated by reference above.

A conformation is a spatial arrangement of the atoms in a molecule at any point in time that results from rotation of covalent bonds. Thus, a molecule is capable of adopting an many conformations since bonds in the molecule can rotate substantially in a plurality of small increments. Other motions, such as "bending" of bonds, can also occur. In practice, however, there tends to be a finite number of important conformational states of a molecule as a result of stearic interactions (collisions) between atoms at certain locations during rotation about a given covalent bond. Those states with minimal steric interactions have a lower potential energy and are called the preferred conformations. For example, an ethane molecule can rotate about its central bond throughout 360 degrees, but spends most of its time at positions near 60 degrees, near 180 degrees or near 300 degrees of rotation, its preferred conformations.

In a preferred embodiment, step 324 comprises identifying at least one of a plurality of representative conformations for the fragment using one of the techniques known in the art, for example that described in the commonly owned copending U.S. Patent Application No. 09/102,600, incorporated herein by reference in its entirety for all purposes. Each conformer can then be fit to the hypothesis (a pose determined) using any of the means known in the art. An identification of possible binding features in the fragment can be made. Prototype reactants may be chosen with some minimum depth, as is referred to as parameter B in the pseudo code of Table 1 above, to accommodate instances where features at boundaries between reactants exist. In select embodiments, conformational analysis and fitting operations can be combined into one operation. These are further described in U.S. Patent No. 5,526,281 (cited above); Y. Martin, J. Med. Chem. 1992 V. 35 pp. 2145-2154 and references cited therein (cited above).

In a presently preferred embodiment, fragments corresponding to partial products can be identified by instances and prototypes giving rise to the fragment. In a particular embodiment conformer and pose can be identified for the fragment, and can be represented symbolically in database tables, for example.

In a presently preferred embodiment, identifying conformer and pose can be facilitated by labeling the atoms of the fragment. Preferably these labels are applicable in many particular contexts in which the fragment may appear. Atoms derived from a prototype, or atoms derived from an instance of a reactant but which correspond to an atom in a prototype of that instance, can be labeled with the name of the atom in the prototype, for example. In a particular embodiment, a label may be the number of a prototype and the number of a relevant atom in the prototype. Other labeling and identification paradigms can be used without departing from the scope of the present invention. Correspondences between prototypes and instances may be determined automatically by any of a plurality of techniques known to those of ordinary skill in the art, such as subgraph isomorphism. Reference may be had to publications, such as "Introduction to Algorithms" by T. Cormen, et. al. (1989) for further details on such techniques. Atoms that do not correspond to a prototype atom can still correspond to some atom in an instance of a reactant. In this latter case, in a particular embodiment, the instance number and the number of the atom within the instance may be used as a label. Alternatively, an instance number alone may be used as a label. Other labeling and identification paradigms can be used without departing from the scope of the present invention.

Conformer and pose can be identified using any of a plurality of techniques in various embodiments according to the present invention. In a presently preferred embodiment, for hypotheses comprising three dimensional pharmacophores, the correspondence between fragment atoms and pharmacophore features with which they align to can provide an indication of the conformation and pose. In an alternative embodiment, additional locations relative to features in a pharmacophore can be defined and used to supplement the correspondence of the first technique. In a yet further embodiment, for hypotheses comprising receptor models, includes matching atoms to a plurality of defined locations in the receptor model, as shown in Fig. 3B. Such locations may be a plurality of spaced locations within a binding cavity. Alternatively, locations of high interaction energy, or a set of bottlenecks in the cavity, such as narrow spots

between more capacious regions or the ends of pockets, could also be used, for example. In a still further embodiment, the conformer can be identified by internal coordinates, such as torsions or bond angles, among the atoms, and the pose by specifying locations with respect to the hypothesis of a plurality of atoms. Finally, embodiments can also include any of these methods in any combination.

In a presently preferable embodiment, fragments fitting one or more of the hypotheses are recorded in one or more tables in a database. However, embodiments can enumerate fragments in any of a wide variety of ways, such as linked lists, files, tree data structures, specialized data structures and the like without departing from the scope of the present invention. Tables can comprise information about the structure of the fragment such as combinations of reactant instances which give rise to the fragment, the features or locations in the hypothesis that the fragment fits, conformer and pose information for the fragment, and the like. Some embodiments will not contain all of these types of information, while many embodiments can also include other information as well without departing from the scope of the present invention. In a presently preferable embodiment a database join operation can be performed upon such tables to form a list of mutually consistent sets of fragments. Other operations for determining combinations of fragments can also be performed, such as for example an intersection of the fragment data, and the like, without departing from the scope of the present invention.

Table 2 shows pseudocode comprising the forming of partial products and fragments, the conformational analysis of the fragments and their fit to the hypothesis, and the labeling of the fragments and their entry into the tables in a representative example embodiment according to the present invention. Steps can be added, deleted or reordered without departing from the scope of the present invention.

```
//Form Tables Characterizing Partial Products' Fits to Hypothesis

For each reactant R {
   Create an empty table for characterizations of those partial products using an instance of reactant R.
   Table has a column for each reactant and a column for each feature in the hypothesis.

For each combination of an instance of R and a prototype of each other reactant {
   Form partial product P using the combination of instance and prototypes according to the reactions in
   the cascade.
   Use the partial product P as a fragment.

   For each conformation and pose of P that fits hypothesis {
      // Loop over labelings of atoms in P
```

```
For each prototype product C from PrototypeProductList {
  For each subgraph isomorphism of C into P {
      // Use isomorphism to label atoms of P
      Clear label of each atom in P.
      For each atom A in C {
          Find corresponding atom in P and label it with A.
          }
          Label all remaining unlabeled atoms in P with reactant R.

          // Record characterization of P under this labeling in table.
          Create a new row in table.
          For each reactant R' {
              Enter the instance or prototype of R' used to make P in the column for R' of the current
              row.
          }
          For each feature or location F in hypothesis {
              Find the atom in P that aligns with F.
              Enter the label of the atom in the column for F in current row.
              If there is no such atom, enter null instead.
          }
          } // end; for isomorphisms
        } // end; for prototype products
      } // end; for conformations and poses
    } // end; for partial products
  } // end; for each reactant
```

Table 2

A join operation in the relational database arts comprises an operation performed upon tables of one or more databases having at least one column label common to both tables. The join of the tables is defined as a third table whose column labels are the union of the column labels of the two input tables. This resultant table includes combinations of rows from the two input tables that have consistent entries in the common columns. By joining the tables of fragments, a table of combinations of fragments that comprise compounds meeting the hypothesis can be formed. Each combination of fragments can form a complete molecule that is likely to fit the hypothesis. Note that it is not necessary to guarantee that each such molecule actually fit the hypothesis, only that the probability of this is high and that molecules that fit the hypothesis can be included in the result of the join. In a particular embodiment, false positives that may occur can be screened out by a subsequent check to see if the complete compound indeed fits the hypothesis. Other methods of determining from fragments combinations that can form molecules with a high probability of meeting the hypothesis can also be used, such as intersection operations and the like, without departing from the scope of the present invention.

Table 3 shows pseudocode of operations in joining tables of reactants in a database in a particular representative example embodiment according to the present invention. Steps can be added, deleted or reordered without departing from the scope of the present invention.

```
// Join Operation on Database
// Note: A partition of the features of the hypo is a function f[] mapping
// each feature F to a particular reactant R=f[F] (indicating the atom
// satisfying the feature came from reactant R) or to a special value
// 0=f[F] (indicating the atom satisfying the feature appeared in a
// prototype)
For each partition of the features f[] {
  For each reactant R {
      Select rows of table for R such that for all F the entry in the column for F is:
                  a)  some prototype atom A if f[F]=0,
                  b)  the reactant R if f[F]=R, and
                  c)  null if f[F] is one other value.
      }
Join the selected rows of each table, considering as the common columns those columns corresponding to F
such that f[F]=0.
Each row in the resulting table indicates a candidate compound.
}
```

Table 3

In a particular representative example embodiment according to the present invention, a search of a virtual library of 20x20x20=8000 tripeptides is discussed. In this particular example, three reactants numbered R1, R2 and R3, each having 20 instances, corresponding to the 20 amino acids are discussed. These are denoted "gly", "ala", "phe", etc. Each reactant has one prototype: NH2-CH2-COOH. The prototype is denoted "xxx". Non-hydrogen atoms in the prototype are denoted An, Ac1, Ac2, Ao1, Ao2. A hypothesis having three features: a feature F1, comprising a Carbonyl oxygen; a feature F2, comprising a Phenyl ring and a feature F3, comprising a Phenyl ring is specified.

A partial product is described in a database wherein reactant R1 is phe, reactant R2 and reactant R3 are prototypes. This partial product can align with the hypothesis such that atom Ao2 of the prototype used for reactant R2 satisfies feature F1 and the phenyl group of the phe can satisfy feature F2. Feature F3 is left unsatisfied. This gives rise to a row in table 4 for R1 as follows:

TABLE R1

R1	R2	R3	F1	F2	F3
phe	xxx	xxx	Ao2	R1	null

Table 4

A second partial product is described wherein reactant R2 is gly and R1 and R2 are prototypes. The second partial product can align with the hypothesis such that an atom of the gly that corresponds to atoms Ao2 of the prototype satisfies feature F1. Neither of the other two features is satisfied. This gives rise to a row in table 5 for R2 as follows:

TABLE R2

R1	R2	R3	F1	F2	F3
xxx	gly	xxx	Ao2	null	null

Table 5

A third partial product is described wherein reactant R3 is phe and R1 and R2 are prototypes. The third partial product can align with the hypothesis such that atom Ao2 of the prototype used for reactant R2 satisfies feature F1 and the phenyl group of the phe can satisfy feature F3. Feature F2 is left unsatisfied. This gives rise to a row in the table 6 for R3 as follows:

TABLE R3

R1	R2	R3	F1	F2	F3
xxx	xxx	phe	Ao2	null	R3

Table 6

Among the possible partitions of features F1, F2 and F3 is f[F1]=0, f[F2]=R1 and f[F3]=R3. In this partition, column F1 is treated as common when table 4, table 5 and table 6 are joined. The result table 7 includes a result row arising from the three rows shown above:

JOIN OF R1, R2, R3 TABLES FOR PARTITION 0,R1,R3

R1	R2	R3	F1	F2	F3
phe	gly	phe	Ao2	R1	R3

Table 7

This indicates that the compound phe-gly-phe is a candidate for satisfying the hypothesis.

In conclusion the present invention provides for a method and system for searching a virtual library of synthesizable chemical compounds in order to identify select component reactants which, when combined, will yield compounds having a set of desirable properties. One advantage of some embodiments according to the present invention is that the speed limiting aspect of a search is done by a purely symbolic computation that does not require manipulation of atomic representations or coordinates. Many embodiments will also enable the identification of partial fits. In these embodiments, molecules that fit some but not all of the features of the hypothesis may be identified.

Other embodiments of the present invention and its individual components will become readily apparent to those skilled in the art from the foregoing detailed description. As will be realized, the invention is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the spirit and the scope of the present invention. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive. It is therefore not intended that the invention be limited except as indicated by the appended claims.

WHAT IS CLAIMED IS:

1. A computer-based method for searching a plurality of compounds comprising the steps:

describing a virtual library of compound fragments, said compound fragments combinable to form said compounds; and

searching said virtual library for compound fragments meeting a hypothesis, wherein said compound fragments being combinable to form said compounds, wherein said compound fragments are not instantiated prior to said searching.

2. The method of claim 1 wherein said describing step further comprises:

encoding at least one of a plurality of chemical reactions, each chemical reaction having at least one of a plurality of reactants;

enumerating instances of each reactant in said plurality of chemical reactions;

specifying at least one of a plurality of relationships between said reactions and between said reactions and said reactants; and

providing a hypothesis.

3. The method of claim 1 wherein said searching further comprises:

determining for each reactant at least one of a plurality of prototypes;

determining partial products from each of said reactants;

determining compound fragments from said partial products, each compound fragment fitting said hypothesis; and

enumerating at least one of a plurality of compounds from said compound fragments, said at least one of a plurality of compounds meeting said hypothesis.

4. The method of claim 3 further comprising performing conformational analysis for said partial products and thereupon determining said compound fragments based on said conformational analysis.



5. The method of claim 4 wherein said performing conformational analysis step further comprises identifying compound fragments by at least one of a plurality of internal conformational coordinates.

6. The method of claim 1 wherein said hypothesis further comprises a receptor model from a crystal structure.

7. The method of claim 4 wherein said hypothesis further comprises a pseudo receptor model formed from a plurality of structure activity data.

8. The method of claim 7 wherein said performing conformational analysis step further comprises matching compound fragments to at least one of a plurality of locations in said pseudo receptor model.

9. The method of claim 4 wherein said hypothesis further comprises a three dimensional pharmacophore.

10. The method of claim 9 wherein said performing conformational analysis step further comprises matching compound fragment features to pharmacophore features.

11. The method of claim 1 wherein said hypothesis further comprises a three dimensional similarity to a reference compound.

12. The method of claim 1 wherein said searching step can be performed interactively with a user.

13. A computer based system for searching a plurality of compounds comprising:

means for describing a virtual library of compound fragments, said compound fragments combinable to form said compounds, said virtual library further comprising:

a hypothesis; and

a graph, said graph further comprising:

a first reaction node, having a first intermediate product;

a second reaction node, having a second intermediate product;



a merge node, disposed to combine said first intermediate product with said second intermediate product to form a merged intermediate product; and

means for searching said virtual library for compounds meeting said hypothesis.

14. The computer based system of claim 13 further comprising at least one filter node, said filter node disposed to operate on said merged intermediate product.

15. A computer based system for searching a plurality of compounds comprising:

means for describing a virtual library of compound fragments, said compound fragments combinable to form said compounds, said virtual library further comprising:

a hypothesis; and

a graph, said graph further comprising:

a plurality of reaction nodes, including a first reaction node, a second reaction node and a third reaction node, said first reaction node having a first intermediate product, said second reaction node having a second intermediate product;

a merge node, disposed to combine said first intermediate product and said second intermediate product, forming a merged intermediate product as input to said third reaction node;

at least one filter node, said filter node disposed to operate on a result from said third reaction node; and

means for searching said virtual library for compounds meeting said hypothesis.

16. A computer based system for searching a plurality of compounds comprising:

means for describing a virtual library of compound fragments, said compound fragments combinable to form said compounds, said virtual library further comprising:

a hypothesis; and

a graph, said graph comprising:

a plurality of reaction nodes, including a first reaction node, having a first intermediate product, a second reaction node, having a second intermediate

* C L M S *

product and a third reaction node, having a third intermediate product; wherein said first intermediate product is disposed to be an input into said second reaction node and said third reaction node;

a merge node, disposed to combine said second intermediate product and said third intermediate product, forming a merged intermediate product; and

means for searching said virtual library for compounds meeting said hypothesis.

17. A computer programming product for searching a plurality of compounds comprising:

code for describing a virtual library of compound fragments, said compound fragments combinable to form said compounds;

code for searching said virtual library for compound fragments meeting a hypothesis, wherein said compound fragments being combinable to form said compounds, wherein said compound fragments are not instantiated prior to said searching; and

a computer readable storage medium for holding said codes.

18. The computer programming product of claim 17 wherein said code for describing further comprises:

code for encoding at least one of a plurality of chemical reactions, each chemical reaction having at least one of a plurality of reactants;

code for enumerating instances of each reactant in said plurality of chemical reactions;

code for specifying at least one of a plurality of relationships between said reactions and between said reactions and said reactants; and

code for providing a hypothesis.

19. The computer programming product of claim 17 wherein said code for searching further comprises:

code for determining for each reactant at least one of a plurality of prototypes;

code for determining partial products from each of said reactants;

code for determining compound fragments from said partial products, each

compound fragment fitting said hypothesis; and

code for enumerating at least one of a plurality of compounds from said compound fragments, said at least one of a plurality of compounds meeting said hypothesis.

20. The computer programming product of claim 19 further comprising code for performing conformational analysis for said partial products and thereupon determining said compound fragments based on said conformational analysis.

21. The computer programming product of claim 20 wherein said code for performing conformational analysis further comprises code for identifying compound fragments by at least one of a plurality of internal conformational coordinates.

22. The computer programming product of claim 17 wherein said hypothesis further comprises a receptor model from a crystal structure.

23. The computer programming product of claim 20 wherein said hypothesis further comprises a pseudo receptor model formed from a plurality of structure activity data.

24. The computer programming product of claim 23 wherein said code for performing conformational analysis step further comprises code for matching compound fragments to at least one of a plurality of locations in said pseudo receptor model.

25. The computer programming product of claim 20 wherein said hypothesis further comprises a three dimensional pharmacophore.

26. The computer programming product of claim 26 wherein said code for performing conformational analysis step further comprises code for matching compound fragment features to pharmacophore features.

27. The computer programming product of claim 17 wherein said hypothesis further comprises a three dimensional similarity to a reference compound.

28. The computer programming product of claim 17 wherein said code for searching can operate interactively with a user.

* C L M S *