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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 March 2003 (06.03.2003)

PCT

(10) International Publication Number
WO 03/019140 A2

(51) International Patent Classification^7: G01N

(21) International Application Number: PCT/US02/26844

(22) International Filing Date: 22 August 2002 (22.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/314,380 23 August 2001 (23.08.2001) US

(71) Applicant: DELTAGEN RESEARCH LABORATORIES, L.L.C. [US/US]; 4570 Executive Drive, Suite 400, San Diego, CA 92121 (US).

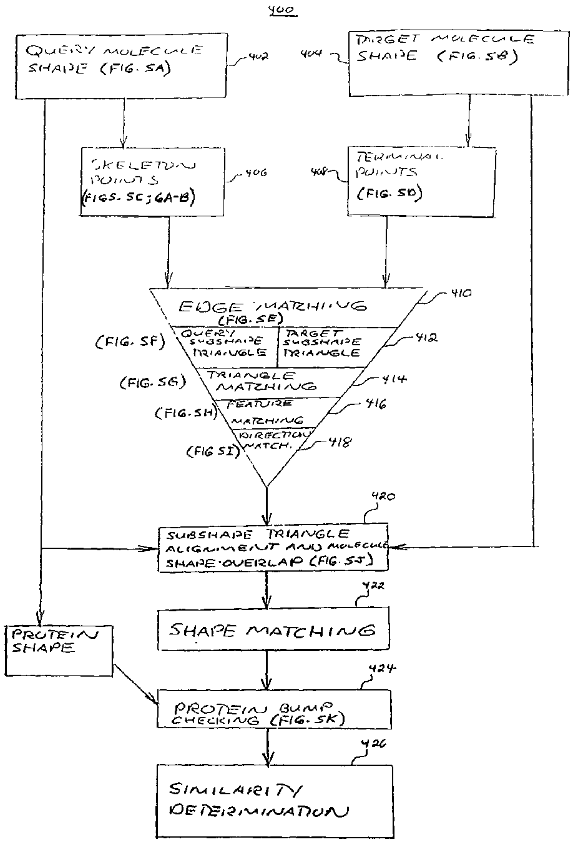
(72) Inventors: PUTTA, Santosh; 34336 Portia Terrace, Fremont, CA 94555 (US). LEMMEN, Christian; Ellhauser Weg 31, 53797 Lohmar (DE). STANTON, Robert, V.; 6774 Saroni Drive, Oakland, CA 94611 (US). BLANEY, Jeffery; 130 Alta Avenue, Piedmont, CA 94611 (US). ROBINSON, Leslie, Ann; 14099 Recuerdo Drive, Del Mar, CA 92014 (US).

(74) Agents: TOBIN, Kent, J. et al.; Townsend and Townsend and Crew LLP, Two Embarcadero Center, Eighth Floor, San Francisco, CA 94111-3834 (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, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

[Continued on next page]

(54) Title: METHOD FOR MOLECULAR SUBSHAPE SIMILARITY MATCHING



(57) Abstract: A database of three-dimensional molecule shapes may be effectively searched for similarity to a query molecule subshape. Triangle matching is first performed between a query subshape triangle representative of a local volume distribution within the query molecule, and a target subshape triangle representative of a local volume distribution within a target molecule of the database. Based upon overlap of the target molecule and query molecule as determined by alignment of matched subshape triangles, shape matching between the aligned query and target molecules is then performed. Comparison of a direction assigned to each vertex of the matched subshape triangles based upon a principal axis of sampled local volume, eliminates unsuitable subshape triangle pairs from consideration prior to the computationally-intensive shape matching step.

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

Published:

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

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

METHOD FOR MOLECULAR SUBSHAPE SIMILARITY MATCHING

CROSS-REFERENCES TO RELATED APPLICATIONS

[01] The instant non-provisional patent application claims priority from U.S. Provisional Patent Application No. 60/314,380, filed August 23, 2001 and incorporated by reference in its entirety for all purposes herein.

BACKGROUND OF THE INVENTION

[02] The field of the instant invention relates to molecular modeling, and in particular to methods for matching three-dimensional molecular structures based upon similarity in molecular subshapes.

[03] The biological activity of many drugs is dependent upon their three-dimensional shape. For example, Fig. 1A shows a simplified schematic view of the interaction of a drug, referred to here as query molecule 100, with a biological macromolecule 102 such as a protein. Drug 100 may function by projecting an active region 100a into a receptor site 102a having a distinctive three-dimensional shape.

[04] C. Burt, in Molecular Similarity Calculations for the Rational Design of Bioactive Molecules in Molecular Modeling and Drug Design, JG Vinter & M Gardner, Eds., MacMillan, (1994) pp. 305-332, has proposed that non-covalent forces dominate drug-receptor interactions, and that these forces can be described in terms of van der Waals and electrostatic effects. .This article is incorporated by reference for all purposes herein. R.B. Silverman in The Organic Chemistry of Drug Design and Drug Action, Academic Press: San Diego, (1992) pp. 44-47 has asserted that the more complementary the fit between drug and receptor, the more potent the drug will be against the receptor. For purposes of describing van der Waals interactions, the term "complementarity" means that the shape of the drug is complementary to the shape of the receptor-binding or enzyme-active site of the receptor structure. This is sometimes referred to as the "lock-and-key" hypothesis. It has thus long been recognized that molecular shape is one of the major determinants of activity.

[05] Frequently during the drug design process, a researcher becomes aware of one or a small number of molecules of known two-dimensional connectivity and/or three-dimensional shape exhibiting affinity to a particular receptor. Based upon these initial experimental results, the researcher seeks to identify additional candidate molecules for screening against the receptor. Several computational procedures that assist in identifying and optimizing this candidate screening process use or take into account molecular shape concepts. One example of such a candidate screening process is the searching of an existing database comprising two-dimensional molecular connectivity information and/or three-dimensional molecule shapes for molecules exhibiting a shape similar to one or more query molecules exhibiting activity against a particular receptor or enzyme. See C. Good et al., "Three-Dimensional Structure Database Searches", Rev. Comput. Chem. 7, 67-117 (1996).

[06] Compounds that are selected from the database in this way can subsequently be screened for activity utilizing *in vitro*, *in vivo*, or *in silico* techniques. If the candidates are indeed found to be active, they may provide a novel starting point for further optimization.

[07] Fig. 1B shows a schematic view of the role of shape similarity matching in drug design. In Fig. 1B, the researcher conducting such *in silico* screening may therefore conduct a search of a database 110 containing two- or three-dimensional information of a large number of molecules, posing a query in the form of the three-dimensional shape 104 of query molecule 100 in order to identify a target molecule shape 112 having features shaped similarly to those of query molecule 100.

[08] While the importance of the entire molecule shape has been recognized in drug design, it can be noted that a subshape (part of the entire molecular shape) may interact with the receptor to determine activity. As a result, two compounds that are dissimilar in their overall shape may be active against the same receptor because parts of their shapes are similar to each other.

[09] Fig. 1C shows a schematic view of the interaction between similar subshapes of two differently-shaped drug molecules with the same receptor. As shown in Fig. 1C, the affinity of a molecule 100 to receptor site 102a is generally dictated by molecule subshape such as active regions 100a or 112a, rather than by the overall shape of the molecule 100 or 112. In attempting to identify similarity between a query molecule of known affinity and a target molecule shape in a database, one may employ a large number of molecule rotation-translation steps to match target molecule shapes with query molecule shapes. However, such shape matching processes may be computationally-intensive, requiring expensive equipment and extended periods of time to perform. These shape matching processes also tend to emphasize overall molecule shape similarity, and may fail to recognize similarity between subshapes of the molecules.

[10] In order to accelerate the shape matching process, other conventional approaches have utilized simplifying techniques. For example, Fig. 2 shows a conventional molecular similarity matching method which identifies the centroid 114 of query molecule shape 104 and the centroid 116 of target molecule 112. Centroids 114 and 116 of query and target molecule shapes 104 and 112 respectively are placed at the origin of a three-dimensional grid 118. The degree of overlap between shapes 104 and 112 is then evaluated.

[11] While providing relatively rapid results, the conventional overall shape matching technique shown in Fig. 2 is of limited utility because it fails to recognize critical subshapes that are important in determining the affinity of a ligand to a receptor.

[12] Therefore, there is a need in the art for methods of molecular similarity matching that accounts for similarity in subshapes.

BRIEF SUMMARY OF THE INVENTION

[13] A database comprising three-dimensional target molecule shapes may be effectively searched for similarity to a subshape of a query molecule. Methods in accordance with embodiments of the present invention thus enable a researcher to identify promising candidates for biological screening experiments during the drug discovery process, based upon only a query molecule of known three-dimensional structure exhibiting affinity to a particular receptor.

[14] In one embodiment of the method, triangle matching is first performed between a query subshape triangle representative of a local volume distribution within the query molecule, and a target subshape triangle representative of a local volume distribution within a target molecule of the database. Shape matching is then performed between the query molecule and the target molecule, based upon overlap of the target molecule and query molecule as determined by alignment of matched query and target subshape triangles. Comparison of a characteristic direction assigned to each vertex of the subshape triangles based upon a principal axis of sampled local volume may eliminate unsuitable subshape triangle pairs from consideration prior to the shape matching step, thereby optimizing computational efficiency.

[15] An embodiment of a method for comparing a query molecule shape with a target molecule shape comprises, sampling a distribution of local volume of the query molecule shape to generate a plurality of skeleton points, each skeleton point including a location and a direction. The direction of each skeleton point is determined by a principal axis of the sampled local volume distribution of the query molecule shape. A distribution of local volume of the target molecule shape is sampled to generate a plurality of terminal points fewer in number than the skeleton points, each terminal point including a location and a

direction. The direction of each terminal point is determined by a principal axis of the sampled local volume distribution of the target molecule shape. A query subshape triangle is created from three skeleton points. A target subshape triangle is created from three terminal points. The query subshape triangle and the target subshape triangle are matched to determine an optimal translation and rotation of the target subshape triangle relative to the query subshape triangle to align corresponding skeleton and terminal points. Directions of the aligned skeleton points and terminal points at corresponding vertices of the matched subshape triangles. The query molecule shape and the target molecule shape are overlapped by alignment of matched target and query subshape triangles, and the overlapped query and target molecule shapes are compared.

[16] In a first step of an embodiment of a method of searching for three-dimensional target molecule shapes matching a query molecule shape, a query molecule shape is provided. A distribution of local volume of the query molecule shape is sampled to generate a plurality of skeleton points. Each skeleton point includes a location and a direction, the direction determined by a principal axis of the sampled local volume distribution of the query molecule shape. A distribution of local volume of a target molecule shape is sampled to generate a plurality of terminal points fewer in number than the skeleton points. Each terminal point includes a location and a direction, the direction determined by a principal axis of the sampled local volume distribution of the target molecule shape. Query subshape triangles are created from three skeleton points. Target subshape triangles are created from three terminal points. Triangle matching values are determined for query/target subshape triangle pairs having an optimal translation and rotation relative to one another. Net direction differences of corresponding skeleton points and terminal points are determined only for vertices of query/target subshape triangle pairs satisfying a triangle matching threshold. The query molecule shape and the target molecule shape are overlapped by alignment of the query subshape triangle and the target subshape triangle only for query/target subshape triangle pairs satisfying a net direction difference threshold. The overlapped query and target molecule shapes are then compared.

[17] These and other embodiments of the present invention, as well as its advantages and features, are described in more detail in conjunction with the text below and attached figures.

BRIEF DESCRIPTION OF THE DRAWINGS

[18] Fig. 1A shows a simplified schematic view of the interaction of a drug with a biological macromolecule.

[19] Fig. 1B shows a schematic view of the role of shape similarity matching in drug design.

[20] Fig. 1C shows a schematic view of the interaction between similar subshapes of two differently-shaped drug molecules with the same receptor.

[21] Fig. 2 shows a schematic view of a conventional molecular shape matching technique;

[22] Fig. 3 shows a schematic view of one embodiment of a molecular subshape matching technique in accordance with the present invention;

[23] Fig. 4 shows a flowchart of steps performed for molecular subshape matching in accordance with one embodiment of the present invention;

[24] Figs. 5A-5K show schematic views of the steps of Fig. 4;

[25] Fig. 6A is a flowchart showing the steps of one embodiment of a method in accordance with the present invention for generating skeleton points from a target molecule shape;

[26] Fig. 6B is a flowchart showing the steps of one embodiment of a method in accordance with the present invention for generating a minimum required number of initial skeleton points;

[27] Fig. 6C shows a schematic view of the iterative process for generating additional initial skeleton points of a query molecule shape;

[28] Fig. 7 shows a schematic view of the succession of various filtering steps performed in accordance with an embodiment of the present invention;

[29] Fig. 8A is a simplified diagram of a computing device for processing information according to an embodiment of the present invention;

[30] Fig. 8B is an illustration of basic subsystems in the computer system of Fig. 8A;

[31] Fig. 9 is a simplified block diagram an embodiment of a software program used to perform subshape matching in accordance with the present invention;

[32] Figs. 10A-10C show the two-dimensional connectivities of a query molecule and two target molecules utilized in an experiment to demonstrate an embodiment of a method in accordance with the present invention; and

[33] Figs. 11A-11E show the three-dimensional alignment of the query molecule and the target molecules obtained from the first experiment.

[34] Fig. 12 is a flow chart showing the steps of a method for applying a subshape-matched target molecule in accordance with an embodiment of the present invention to identify possible drug leads.

[35] Fig. 13A shows the two-dimensional connectivity of the 1tlp and ppp molecules.

[36] Fig. 13B shows the a three-dimensional representation of the 1tlp and ppp ligands aligned as bound to thermolysin.

[37] Fig. 14A shows a simplified cross-sectional view of a two-dimensional representation of molecule shape volume encoding.

[38] Fig. 14B shows a simplified cross-sectional view of a two-dimensional representation of molecule shape surface encoding.

DETAILED DESCRIPTION OF THE INVENTION

[39] Fig. 3 shows a schematic view of operation of a method in accordance with one embodiment of the present invention. A database comprising three-dimensional target molecule shapes 112 may be effectively searched for similarity to a subshape 104a of query molecule 100 having overall shape 104. In one embodiment of the method, triangle matching is first performed between a query subshape triangle 140 representative of a local volume distribution within query molecule 100, and a target subshape triangle 142 representative of a local volume distribution within target molecule 112. Shape matching is then performed between query molecule 100 and target molecule 112, based upon overlap of target molecule 112 and query molecule 100 as determined by alignment of matched query and target subshape triangles 140 and 142, respectively. Comparison of a characteristic direction assigned to each vertex of the subshape triangles based upon a principal axis of the local sampled volume, may filter unsuitable subshape triangle pairs prior to shape matching.

[40] Fig. 4 shows a schematic flow chart of one embodiment of a method 400 in accordance with the present invention for searching for three-dimensional target molecule shapes matching a query molecule shape. Figs. 5A-5K show schematic views of the corresponding steps in the method of Fig. 4.

[41] In a first step 402 of method 400, query molecule shape 104 of Fig. 5A is provided. The shape of the query molecule may be determined from conformational analysis of the molecule. Examples of approaches for analyzing the three-dimensional conformation of a molecule are presented by Smellie et al., "Conformational Analysis by intersection: Ring conformation", Proc. of the 217th Meeting of the ACS, Anaheim (1999), and by Smellie et al., "Conformational Analysis by Intersection", J. Comput. Chem. Vol. _, No. _, pp. _____ (accepted for publication 2002), both of which are incorporated by reference herein.

[42] Alternatively the shape of the query molecule may be determined from experimental results, for example through multi-dimensional NMR spectroscopy studies, circular dichoism

(CD) spectroscopy studies, or x-ray crystallography of the query molecule 100 bound to receptor 102.

[43] In a second step 404 of method 400, target molecule shape 112 of Fig. 5B is provided. Shape 112 of the target molecule may be obtained from information contained in a relevant database, which may contain a direct representation of the configuration of the molecule in three-dimensional space. Examples of formats for presenting representations of molecules in space include, but are not limited to the SMILES format from Daylight Chemical Information Systems, Mission Viejo, California, and described by Weininger, in "SMILES 1. Introduction and Encoding Rules", J. Chem. Inf. Comput. Sci. 28, 31 (1988), incorporated herein by reference, the MOL2 format by Tripos Inc. of St. Louis, Missouri, the MOL and SDF formats of MDL of San Leandro, California, and the PDB format of the Protein Data Bank, http://www.rcsb.org/pdb/, incorporated herein by reference. Where the database does not include such a three-dimensional representation of the molecule, the three-dimensional shape 112 of the target molecule can be derived from conformational analysis of two-dimensional molecular connectivity information.

[44] Once the query and target molecule shapes have been obtained, they may be superimposed onto a three-dimensional cubic grid to facilitate encoding of the molecule shape. Grid spacing, the edge length of each cube of the lattice, may be specified by the user and determines the accuracy of the shape encoding. The volume of a molecule shape may be encoded using a bit vector whose length is proportional to the number of grid points.

[45] For example, Fig. 14A shows a simplified cross-sectional view of a two-dimensional representation of volume encoding. Molecule 1400 is represented by shape 1402. Grid points falling within interior region 1404 are assigned an occupancy value of 3. The occupancy value of a grid point decreases gradually moving beyond the van der Waal surface of the molecule. In Fig. 14A, r_w is the van der Walls radius of atom 1400a, $r_1 = (r_w + (r_b/s_g))$, and $r_2 = (r_w + (2r_b/s_g))$, where $r_w =$ van der Waals radius of the heavy atom; $r_b =$ user specified parameter; and $s_g =$ grid spacing. Thus grid points falling within interior surface region 1406 are assigned a value of 2, grid points falling within exterior surface region 1408 are assigned a value of 1, and grid points falling outside of the molecule shape entirely are assigned a value of 0.

[46] As shown below in TABLE A, the occupancy at each grid point may be assigned based on its distance to the closest heavy atom. Two bits are assigned to store the occupancy of an individual grid point, allowing for four distinct values based on whether the grid point

is located at the interior of the molecule, the exterior of the molecule, or at the interior or exterior surface of the molecule.

TABLE A

r_w = van der Waals radius of the heavy atom;
r_b = user specified parameter; and
s_g = grid spacing

Distance (d) From Grid Point To Center Of Closest Heavy Atom	Grid Occupancy Value	Grid Point Location
$d \leq r_w$	3	interior
$r_w < d \leq (r_s + r_b/s_g)$	2	interior surface
$(r_s + r_b/s_g) < d \leq (2r_b/s_g)$	1	exterior surface
$d > (2r_b/s_g)$	0	exterior

[47] While the use of a grid for volume encoding is described, the invention is not limited to this particular technique, and the volumes of molecule shapes can be encoded in other ways, for example based upon the atom coordinates of the atoms comprising the molecule.

[48] Once the volumes of the query molecule has been encoded as just described, in a third step 406, a distribution of local volume of the query molecule shape is sampled to generate a plurality of skeleton points 120 of Fig. 5C. Figs. 6A-6B, discussed below, provide a detailed description of the generation of skeleton points.

[49] Each skeleton point 120 includes a characteristic location 120a and direction 120b, and may further include one or more chemical feature types 120c. The location 120a of a particular skeleton point corresponds to its position within the three-dimensional query molecule shape. The direction 120b of the skeleton point is determined by a principal axis of the sampled local volume corresponding to that particular skeleton point. Chemical feature types 120c of the skeleton point are determined by chemical groups proximate to that skeleton point.

[50] In a fourth step 408, a distribution of local volume of the target molecule shape is sampled to generate a plurality of terminal points 122 of Fig. 5D. The number of terminal points 122 is fewer than the number of skeleton points of the query molecule shape. Like the skeleton points of the query molecule shape, each terminal point 122 includes a characteristic location 122a and direction 122b, and may also include one or more chemical feature types 122c. The location of a particular terminal point corresponds to its position within the three-

dimensional target molecule shape. The direction of the terminal point is determined by a principal axis of the sampled local volume corresponding to the particular skeleton point. Chemical feature types of the terminal points may be determined by the identities of chemical groups proximate to the terminal point. A description of the generation of terminal points is also provided below.

[51] In a fifth step 410, as shown in Fig. 5E, the distance 130a between selected pairs of skeleton points 120 and the distance 130b between selected pairs of terminal points 122 are measured and compared. Based upon the difference between distances 130a and 130b, an edge matching value is calculated.

[52] In a sixth step 412, as shown in Fig. 5F, query subshape triangles 140 and target subshape triangles 142 are assembled from three pairs of skeleton points 120 and terminal points 122, respectively, with corresponding subshape triangle edges formed from pairs of skeleton points and pairs of terminal points. For each subshape triangle pair, the edge matching values just calculated for each edge pair are checked against an edge matching threshold value. Only subshape triangle pairs having all three corresponding edges satisfying this edge matching threshold are passed on to the next step.

[53] Specifically, in seventh step 414, query subshape triangle 140 is matched with target subshape triangle 142 and a triangle matching value is generated. As shown in Fig. 5G, for each query/target subshape triangle pair 144 the triangle matching process produces a triangle matching value and a translation 146 and rotation 148 of the target subshape triangle 142 relative to the query subshape triangle 140. In one embodiment, the triangle matching value may be based upon the root mean square difference (RMSD) between the query and target subshape triangles.

[54] A description of examples of the use of edge matching and triangle matching is described in the following papers: R. Nussinov et al., "Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques", Proceedings of the National Academy of Sciences of the US, Vol. 88, 10495-10499 (1991); and C. Lemmen et al., "Time-efficient flexible superposition of medium-sized molecules", Journal of Computer-Aided Molecular Design, Vol. 11, no. 4, 357-368 (1997). These papers are hereby incorporated by reference for all purposes.

[55] In an eighth step 416, subshape triangle pairs having a triangle matching value satisfying a triangle matching threshold are subjected to feature matching. Specifically, as shown in Fig. 5H, chemical feature types 120c and 122c proximate to corresponding vertices of matched query/target subshape triangle pairs 144 are compared, and a feature difference

calculated. Examples of feature types that may be identified proximate to subshape triangle vertices include hydrogen bond donors, hydrogen bond acceptors, positive charges, negative charges, aromatic groups, and hydrophobic groups.

[56] In a ninth step 418, matched subshape triangles satisfying a feature difference threshold are subjected to direction matching. Specifically, as shown in Fig. 5I, directions 120b and 122b assigned to corresponding skeleton points 120 and terminal points 122 of the vertices of matched query/target subshape triangle pairs 144 are compared, and a net direction difference calculated. One example of this calculation includes taking the sum of the sines (or cosines) of angles θ created by the respective directions of each corresponding skeleton/terminal point. The directions of the skeleton/terminal points considered during this direction matching step reflect rotation/translation of the respective subshape triangles performed during the previous triangle matching step.

[57] In a tenth step 420, molecule volumes corresponding to matched subshape triangles satisfying a direction matching threshold are overlapped. Specifically, as shown in Fig. 5J, query molecule shape 104 and target molecule shape 112 are overlapped. This overlap is based upon alignment of query and target subshape triangles satisfying a direction difference threshold value. The result of this overlap produces three volumes: a volume 190 common to the query and target molecule shapes, a volume 192 of the query molecule shape projecting outside of the target molecule shape, and a volume 194 of the target molecule shape projecting outside of the query molecule shape.

[58] In an eleventh step 422, shape matching is performed for overlapped query and target molecule shapes, and a shape matching value is calculated. The shape matching value generated during this step reflects the overlapped and non-overlapped volumes of the target and query molecule shapes, with a high degree of overlap between the query and target molecule would indicate favored matching.

[59] Specifically, target and query molecule shapes are overlaid onto a three-dimensional cubic grid. At each grid point a shape matching value is determined based on the position of the grid point relative to the surface of the molecule shapes. The shape matching value considers grid points occupying volumes common to the molecule shapes, grid points occupying volumes of the query shape projecting out of the target volume, and grid points occupying volumes of the target shape projecting out of the query shape.

[60] As shown in Fig. 5J, because shape matching in accordance with embodiments of the present invention is performed based upon alignment of subshape triangles 140 and 142, more accurate matching of subshapes between molecules 100 and 112 will result.

[61] The relative size of the target and query molecules may influence the algorithm utilized to quantify overlap between the molecule volumes or surfaces. One approach for calculating overlap between the target and query molecule volumes or shells during shape matching is the protrusion distance calculated according to equation (1):

(1) P = S/V, where:

P = protrusion distance;
S = volume of smaller shape protruding out of larger shape; and
V = volume of the smaller shape.

[62] When the volumes occupied by the query and target molecules are approximately the same size (i.e. the target and query molecules are about the same molecular weight (M.W.)) identification of the "smaller" molecule for purposes of calculation of the protrusion distance becomes less relevant. Accordingly, Tanimoto distance, a second measure of molecular overlap, was developed.

[63] Tanimoto distance may be calculated according to equation (2) as follows:

(2) T = S/U, where:

T = Tanimoto distance;
S = volume not shared by the overlapped molecules; and
U = the union (combined) volume occupied by the overlapped molecules.

Because the Tanimoto distance metric assigns equal weight to the volumes of the query and target molecules regardless of their relative size, this metric is particularly suited for shape matching between query and target molecules of approximately the same size.

[64] The relative size of the target and query molecules may also influence the particular shapes of the molecules that are overlapped during the shape matching step. For example, the chemical activity of a molecule is determined predominantly by shapes and functionalities presented on the surface of the molecule and available to interact with the chemical or biological environment, including other molecules such as receptors or enzymes. Accordingly, during shape matching it may be valuable to emphasize the importance of overlap or non-overlap between portions of the molecular shapes proximate to the surface. This is particularly true when the target/query molecules are substantially different in size

such that alignment may result in the smaller molecule being positioned entirely within the interior of the larger molecule; such similarity in subshape within the interior of the molecules would be considered less relevant to predicting chemical behavior.

[65] In order to compensate for this effect, embodiments of subshape matching processes in accordance with the present invention may utilize a shape matching step based upon overlap between the volumes of shells representing the surface regions (i.e. outlines) of the aligned molecular shapes, rather than the entire volume of the molecular shapes. In this surface matching approach, interior volumes of the molecule shapes are not considered during calculation of overlap, whether using the protrusion or Tanimoto distance measures previously described.

[66] Thus, in calculating grid occupancy values according to the volume approach in accordance with the embodiment of the invention as described above in connection with Fig. 14A, grid points falling within the interior of the molecule shape are accorded a high value. By contrast, Fig. 14B shows a simplified cross-sectional view of an alternative embodiment in accordance with the present invention for use in surface matching, wherein grid points falling within the interior 1404 or exterior 1412 of molecule shape 1402 are accorded no value, and grid points occurring in shells 1414, 1416, 1418, 1420, or 1422 at or near the surface are assigned higher grid occupancy values, thereby creating a molecule volume in the shape of a shell.

[67] To summarize, the relative sizes of the target and query molecules may influence both the choice of shapes (i.e. molecule volume or molecule shell volume) that are overlapped during the shape matching process, and the algorithm utilized to quantify that overlap. Some examples of suggested combinations of these approaches are presented in TABLE B below.

TABLE B

"drug-like" molecule: M.W. ~ 400-700
molecule "fragment": M.W. ~ 100 or less
"large" molecule: M.W. ~ 1000 or greater

Query Molecule	Target Molecule	Applications(s) (Discussed below)	Overlap Type	Distance Measure
drug-like	drug-like	database searching	volume or surface	protrusion or Tanimoto
drug-like	large	docking	volume or surface	protrusion
fragment	drug-like	template evaluation	surface	protrusion

[68] In a twelfth step 424, a protein bump checking step is performed for query and target molecule shapes that satisfy a shape matching threshold. This protein bump checking step is depicted in Fig. 5K. Query molecule 100 is oriented within receptor 102a of macromolecule shape 102 which is typically a protein but could be another type of macromolecule such as a nucleic acid . Based upon alignment of subshape triangles, target molecule shape 112 is then substituted for the query molecule shape 104 within receptor 102a, and an overlap 196 between target molecule shape 112 and protein shape 102 is identified. A protein overlap value reflecting overlap between the target molecule shape and the protein shape is then calculated. This protein overlap value reflects the volume overlapped between the query and target molecule shapes. Unlike the prior shape matching value just calculated, a high degree of overlap between the protein shape and the target molecule shape would indicate a disfavored interaction between target and receptor.

[69] Finally, in a step 426, target molecule shapes satisfying a protein overlap threshold are identified as close matches. These ultimately selected target molecules are favorable candidates for actual screening experiments to determine affinity to the receptor.

[70] As is evident from the above discussion, one aspect of embodiments of methods in accordance with the present invention is that shape matching between query and target molecules is based upon alignment of subshape triangles representative of a local molecule volume distribution. These subshape triangles are in turn generated from skeleton/terminal points resulting from sampling of local volume distributions.

[71] Fig. 6A shows a flow chart illustrating the steps of one embodiment of a method in accordance with the present invention for generating skeleton points. In a first step 602 of skeleton point generation process 600, the known three-dimensional shape of a query molecule is overlaid onto a three-dimensional grid.

[72] In the next step 604, grid points falling within the query molecule shape are identified. In a step 608, a local volume is sampled at each of encompassed points by positioning spheres of a consistent radius at each encompassed grid point, and then calculating the fraction of the volume of the sphere occupied by the query molecule shape. Spheres falling below a minimum volume fraction are determined as defining the boundaries of the query molecule shape. In step 608, centers of these minimum volume fraction spheres are clustered together to produce initial skeleton points.

[73] In general, only a small number of initial skeleton points are generated. While only three such initial skeleton points are required to construct the subshape triangle utilized in the

method, optimal results have been realized with embodiments of methods in accordance with the present invention that utilize five initial skeleton points.

[74] Where fewer than the minimum number of required initial skeleton points have been generated by steps 602, 604, 606, and 608 just described, additional initial skeleton points may be generated. Specifically, Fig. 6B is a flowchart showing the steps of one embodiment of a method in accordance with the present invention for generating additional initial skeleton points. In a first step 652 of Fig. 6B, a pair (A and B) of existing initial skeleton points is selected.

[75] In the next step 654 of Fig. 6B, middle point C is determined along a line segment joining initial skeleton points A and B. In a step 656, a sphere having center O and radius r is placed at middle point C. In a step 658, centroid D of the query molecule shape volume contained within the sphere is computed.

[76] In a step 660, the distance between points O and D is computed. In a step 662, distance OD is compared to a threshold valve. If this distance OD is greater than a threshold value, in a step 664, the sphere is moved such that O coincides with D. Steps 658, 660, and 662 are then repeated until the distance between sphere center O and centroid D is less than the threshold valve. As shown in Fig. 6C, this iterative procedure attempts to place point D within query molecule shape 104, even if the original midpoint of the line AB was located outside of query molecule shape 104.

[77] When the distance between sphere center O and centroid D is less than the threshold valve, in a step 666, the distance d from centroid D to the closest initial skeleton point is computed. In a step 668, the location of centroid D and corresponding distance d is stored. As shown in a step 670, a determination as to whether centroids have been determined for all pairs of skeleton points is made. Where additional pairs of untested initial skeleton points remain, steps 652, 654, 656, 658, 660, 662, and, if necessary, step 664, are repeated until all pairs of initial skeleton points have served as a basis for the generation of centroids.

[78] In a step 672, a new initial skeleton point is chosen from the stored points such that the new initial skeleton point selected corresponds to the maximum distance d.

[79] Once the minimum number of initial skeleton points have been generated, supplemental skeleton points are then added to characterize remaining portions of the query molecule shape. Returning to Fig. 6A, in a step 612, a center of a sphere of a given radius is placed at each grid point encompassed within the query molecule shape. The centroid of the query molecule shape volume inside the sphere is then computed and stored, along with the

volume fraction of the molecule shape falling within the sphere. Spheres falling above a minimum volume function are determined as defining the backbone of the query molecule.

[80] In a subsequent clustering step 614, centroids that are close to one another are removed. In one embodiment, this filtering step 614 can involve starting from each skeleton point (initial and supplemental) and drawing a sphere of a particular radius. All stored high volume fraction centroids falling within the sphere are considered. The centroid of the sphere having the largest corresponding volume fraction is added as a supplemental skeleton point. Remaining centroids within the sphere are discarded. This process is repeated over all the initial and supplemental skeleton points generated.

[81] The supplemental skeleton points, together with the initial skeleton points, define the backbone/skeleton of the query molecule shape. The total number of skeleton points varies according to the overall size of the query molecule, with a range of between 25-100 skeleton points being typical.

[82] While the above description has illustrated generation of skeleton points of a query molecule shape, a truncated version of the same process is utilized to sample local volumes of the target molecule shape and thereby generate terminal points. The primary difference between the process for generating skeleton points of the query molecule and the process for generating terminal points of the target molecule is that supplemental terminal points are not generated. This is because the desirable character of the three-dimensional query molecule shape has already been established through its affinity to the receptor of interest. Therefore, a relatively large number of skeleton points representing a high density of the query molecule shape are helpful in order to accurately characterize query molecule subshapes.

[83] By contrast, a target molecule shape is merely one of many present in the database. Accordingly, in certain embodiments in accordance with the present invention only the initial terminal points are generated for a target molecule shape. Supplemental terminal points are not generated. Limiting the number of terminal points in this manner reduces the number of target subshape triangles, and thus the number of possible combinations of matched query and target subshape triangles, to a quantity manageable by the processing power generally available to personal computers or workstations.

[84] Methods of molecular similarity matching in accordance with embodiments of the present invention offer a number of advantages over conventional methods. One advantage is that the method recognizes similarity between molecule subshapes. This is because shape matching is performed upon query and target molecules aligned according to matched subshape triangles that reflect local, rather than overall molecule shape.

[85] Still another advantage of methods in accordance with embodiments of the present invention is efficiency in the allocation of computing power. The use of multiple subshape triangles as a basis for similarity matching increases the number of possibilities that must be considered. Calculations of the present method requiring the most computational power are triangle matching, shape matching, and protein bump checking steps.

[86] Accordingly, embodiments in accordance with the present invention perform a number of filtering steps to eliminate unpromising subshape orientations from further consideration prior to performance of the computationally-intensive steps. The results of this filtering is shown in Fig. 7, which schematically depicts the successively fewer number of possible combinations that must be evaluated by each method step.

[87] Fig. 7 shows that edge matching initially filters out a large number of potential subshape triangles of disproportionate size prior to the triangle matching step. In this manner, the number of possible subshape triangles pairs to be created from all skeleton points and all terminal points is substantially reduced.

[88] Similarly, prior to alignment of the molecules and performance of the computationally-intensive shape matching step, feature matching and direction matching serve to filter additional unpromising matched subshape triangle pairs. In this manner, the possible number of material subshape triangles is successively reduced from on the order of one million to on the order of 100 or fewer.

[89] Additional filtering may be accomplished during the shape matching step to eliminate unpromising molecular overlaps. Specifically, surface encoding refines volumes of molecule shapes into shells or surfaces, removing from consideration alignments not implicating the molecule surface considered especially relevant to chemical activity. Such surface matching is particularly valuable in reducing unhelpful matches involving a smaller molecule shape aligned wholly within the interior of a larger molecule shape.

[90] Finally, the protein bump checking performed by embodiments of methods in accordance with the present invention ensures that only the most promising matched pairs of query/subshape triangles are considered. This is particularly important given the complexity of most receptor shapes and the large quantity of computing resources that must be allocated to describe overlap between the target and a receptor shape.

[91] The examples and embodiments described herein are for illustrative purposes only. Various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the

appended claims. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives.

[92] For example, some of the above method steps can be omitted or separated or combined, depending upon the particular embodiment. Thus a similarity matching method lacking a feature matching step would fall within the scope of the present invention. Similarly, a method in accordance with an embodiment of the present invention may dispense entirely with the protein matching step, basing subshape matching solely upon overlap of molecule shapes based upon alignment of matched subshape triangles.

[93] Moreover, in some embodiments the order in that the steps are performed can also be changed in order without limiting the scope of the invention claimed herein. Thus a method could perform feature matching either before or after direction matching, and still remain within the scope of the present invention.

[94] While the previous discussion has focused upon methods, a computer system for performing these methods is also a part of the present invention. Accordingly, Fig. 8A is a simplified diagram of a computing device for processing information according to an embodiment of the present invention. This diagram is merely an example which should not limit the scope of the claims herein. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives. Embodiments according to the present invention can be implemented in a single application program such as a browser, or can be implemented as multiple programs in a distributed computing environment, such as a workstation, personal computer or a remote terminal in a client server relationship.

[95] Fig. 8A shows a computer system 810 including a display device 820, a display screen 830, a cabinet 840, a keyboard 850, and a mouse 870. Mouse 870 and keyboard 850 are representative "user input devices." Mouse 870 includes buttons 880 for selection of buttons on a graphical user interface device. Other examples of user input devices are a touch screen, light pen, track ball, data glove, microphone, and so forth. Fig. 8A is representative of but one type of system for embodying the present invention. It will be readily apparent to one of ordinary skill in the art that many system types and configurations are suitable for use in conjunction with the present invention. In a preferred embodiment, computer system 810 includes a Pentium^{TM} class based computer, running LINUX or Windows^{TM} NT operating system by Microsoft Corporation. However, the apparatus is easily adapted to other operating systems and architectures by those of ordinary skill in the art without departing from the scope of the present invention.

[96] As noted, mouse 870 can have one or more buttons such as buttons 880. Cabinet 840 houses familiar computer components such as disk drives, a processor, storage device, etc. Storage devices include, but are not limited to, disk drives, magnetic tape, solid state memory, bubble memory, etc. Cabinet 840 can include additional hardware such as input/output (I/O) interface cards for connecting computer system 810 to external devices external storage, other computers or additional peripherals, further described below.

[97] Fig. 8B is an illustration of basic subsystems in computer system 810 of Fig. 8A. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art will recognize other variations, modifications, and alternatives. In certain embodiments, the subsystems are interconnected via a system bus 875. Additional subsystems such as a printer 874, a keyboard 878, a fixed disk 879, a monitor 876, which is coupled to a display adapter 882, and others are shown. Peripherals and input/output (I/O) devices, which couple to an I/O controller 871, can be connected to the computer system by any number of means known in the art, such as a serial port 877. For example, serial port 877 can be used to connect the computer system to a modem 881, which in turn connects to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus allows a central processor 873 to communicate with each subsystem and to control the execution of instructions from system memory 872 or the fixed disk 879, as well as the exchange of information between subsystems. Other arrangements of subsystems and interconnections are readily achievable by those of ordinary skill in the art. System memory, and the fixed disk are examples of tangible media for storage of computer programs, other types of tangible media include floppy disks, removable hard disks, optical storage media such as CD-ROM's and bar codes, and semiconductor memories such as flash memory, read-only-memories (ROM), and battery backed memory.

[98] Fig. 9 presents a simplified block diagram of one embodiment of a software program 900 used to perform subshape matching in accordance with the present invention. Program 900 exhibits a two-tier architecture that includes an interface 902 as the first tier. During the various steps of embodiments of the subshape matching process just described, user 901 may interact with external features and other components of program 900 through interface 902. Interface 902 may be written in the PYTHON programming language.

[99] Interface 902 is in communication with a database 903 containing information relevant to a large number of possible target molecules. This information may be organized in the form of a number of searchable categories, such as molecule name, biological activity, two-dimensional connectivity, and three-dimensional shape molecular shape. An example of

database 903 is the MDL Drug Data Report™ (MDDR) available from MDL Information Systems, Inc. of San Leandro, California.

[100] The second tier of software program 900 includes a C++ library 906 and may further include a visualization module 908. Visualization module 908 enables the display and manipulation of various shapes during subshape matching in accordance with embodiments of the present invention. Visualization module 908 is in communication with molecular imaging tool 950 for display of three-dimensional molecular shapes. A number of independent software programs are available to serve as a molecular imaging tool, including the WebLab™ software program manufactured by Accelrys, Inc. of San Diego, California.

[101] C++ library 906 is formed from a number of components. A conformer generation module 930 enables the generation of three-dimensional conformers based upon two-dimensional molecular connectivities. A skeleton/terminal point generation module 932 performs the localized volume sampling and point generation techniques performed on the three-dimensional molecular volumes as described above in conjunction with Figs. 6A-B. An edge matching module 934 performs edge matching of distances between skeleton point pairs and terminal point pairs as previously described.

[102] A triangle matching module 936 performs the matching of query/target subshape triangle pairs assembled from the skeleton/terminal points. A feature matching module 938 compares the chemical environment proximate to the vertices of matched query/target subshape triangles. A direction matching module 940 of C++ library 906 compares the principal axes of local sampled volume at the vertices of the matched subshape triangles.

[103] A shape matching module 942 performs alignment of the matched subshape triangles, overlaps the query molecule shape and the target molecule shape based upon this alignment, and then calculates relative overlap between query and target molecule shapes as previously described in connection with Fig. 5J. A protein matching module 944 performs alignment of the target molecule shape within the active site of the protein, and calculates relative overlap between protein and target molecule shapes as previously described in connection with Fig. 5K. Protein matching module 944 may obtain three-dimensional protein shapes from database 903.

[104] Embodiments of subshape matching methods and codes in accordance with the present invention may be utilized in performing a number of applications in connection with the discovery and testing of pharmaceutical compounds.

[105] One possible application for methods of subshape matching in accordance with embodiments of the present invention is in conformational evaluation of potential therapeutic compounds. For example, experimental results may reveal that several ligands exhibit affinity to a particular receptor, but the actual three-dimensional orientation of only one of the ligands is known or suspected. Under such circumstances, an embodiment of subshape similarity matching in accordance with the present invention could utilize the known shape of the bound ligand as the query shape, and various conformers of the other active ligands as the target shapes. Such an conformational evaluation would reveal for the target molecules the conformation(s) in which they may bind to the receptor. Further the three-dimensional orientations of these target molecule binding conformations can also be studied to reveal key ligand-receptor interactions during the binding process.

[106] In order to test the accuracy of one embodiment of a method of subshape matching in accordance with the present invention, the method was utilized to evaluate conformers for active compounds. Specifically, an experiment was performed comparing the molecular alignment predicted by molecular subshape matching versus the actual molecular alignment as revealed by crystallographic data. Crystallographic data was available showing the alignment of one large ligand (fibrinogen), and each of two smaller ligands (NAPAP, PPACK) within the same receptor of a protein (thrombin). Using this empirically-derived alignment information, the larger ligand (fibrinogen) was utilized as a query molecule shape for subshape matching from a database containing the smaller ligands as target molecule shapes. As subshape matching performed in this example occurred between target molecules (NAPAP and PPACK) and a query molecule (fibrinogen) of approximately the same size, the protrusion overlap distance measure was evaluated during the shape matching step.

[107] TABLE C shows the results of this subshape matching experiment. The last column of TABLE C quantifies the difference in RMSD between the molecular alignment predicted by subshape matching in accordance with an embodiment of the present invention, and the actual molecular alignment empirically determined from prior crystallographic studies.

TABLE C

Query Molecule = Fibrinogen

Target Molecule	# of possible alignments of subshape triangle pairs after triangle matching	# of possible alignments of subshape triangle pairs after direction matching	# of possible alignments of subshape triangle pairs after shape matching	RMSD of optimal predicted alignment vs. actual crystal alignment (Å)
NAPAP	6438	649	48	0.53
PPACK	8556	957	108	0.35

[108] Figs. 11A-11E show the three-dimensional orientation of the fibrinogen query molecule as aligned in the thrombin receptor, and the NAPAP and PPACK target molecules, respectively, as obtained from this experiment.

[109] Another possible application for methods of subshape matching in accordance with embodiments of the present invention is searching of a database of three-dimensional molecule shapes to identify molecules possessing subshapes similar to one or more generic molecules active against a particular receptor or enzyme. Thus in a second experiment, subshape matching against a set of 83,178 small molecules from the MDDR database was performed utilizing the NAPAP ligand (known to bind to thrombin) as the query molecule. Assuming similarity in subshape to NAPAP serves as an indicia of activity against thrombin, the number of thrombin-active compounds in the resulting pool of matched target molecules would be expected to be higher than in the original set of 83,000+ molecules. As subshape matching performed in this example occurred between a query molecule (NAPAP) of about roughly the same size as the target molecules of the MDDR database, the Tanimoto distance overlap approach was employed to evaluate the shape matching step.

[110] The database search performed in this example enabled determination of the versatility of the code for handling a wide variety of molecules, the approximate time required to perform the subshape matching, and the behavior of different shape comparison measures. A conformer-generating program was utilized to generate an average of about 65 low-energy conformers for each of the 83,178 compounds. The total conformer library was then searched for subshape similarity to the NAPAP ligand shape posed as bound to the thrombin receptor.

[111] Calculations were run on a LINUX cluster of 20 1.7 GHz Athlon™ and 20 1.3 GHz Pentium 4™ processors. Comparative time required for the jobs on the two processor types resulted in only negligible differences in overall speed. The jobs were run in a trivially parallel fashion, with each processor receiving an equal number of compounds for analysis at the initiation of the run. This allocation resulted in some underutilization of computing resources due to variability in the number of conformers, stereo-isomers, compound size and number of matches, with some processors finishing significantly before others (~15 min in runs averaging 6 hours). This could easily be corrected through a dynamic load balancing application, which would send individual compounds to the processors as they became free. However, as both jobs were completed overnight, additional load balancing was deemed unnecessary.

[112] For each of the volume or surface shape matching algorithms, matches were calculated using three appropriate shape threshold cutoff values utilizing the Tanimoto distance measure. These runs were performed using a 19.5x15x12.5 Å box, with triangle and direction threshold values of 0.5 and 2.6 respectively. Additionally a loose feature matching restriction was employed requiring three matched features at a distance of 1.25Å.

[113] Results from this second experiment are illustrated below in TABLE D For each search criteria, TABLE D lists the total number of selected compounds passing all of the subshape matching filters (N_{aTotal}), along with the corresponding number of these selected compounds which are annotated as "thrombin inhibitors" in the MDDR (N_{ap}). The final number presented in TABLE D for each search is the enrichment value (E_t) for activity in the compounds selected for each experiment. This enrichment is calculated according to Equation (3) below as the density of thrombin actives in the selected set over the density in the original pool of compounds.

(3) $E_t = (N_{ap}/N_{aTotal})/(N_{ip}/N_{iTotal})$; where:

E_t = enrichment;
N_{ap} = number of "thrombin active" target molecules selected;
N_{aTotal} = total number of target molecules selected;
$N_{ip} = 624$ = number of "thrombin active" molecules from the MDDR set; and
$N_{iTotal} = 83{,}178$ = total number of molecules from the MDDR set.

TABLE D

Query Molecule = NAPAP

Overlap Type	Shape Threshold = Tanimoto Distance (T) x 100	N_{aTotal}	N_{ap}	E_t
Volume	40	48	16	44.1
Volume	45	1493	124	11.0
Volume	50	17213	410	3.1
Surface	55	78	300	50.9
Surface	60	3159	188	7.9
Surface	65	34162	518	2.0

[114] The data presented in TABLE D indicates a substantial correlation between the number of matched molecules and the particular shape threshold imposed. Thus changes of 5 units in the shape threshold value may correspond with changes of one or two orders in magnitude in the number of matches. The Tanimoto measure is apparently able to differentiate thrombin actives compounds with significant enrichment, using either the volume or surface overlap approaches. Thus where a high amount of overlap (corresponding to a tight Tanimoto distance cutoff and a low shape threshold value) is used, substantial enrichments of 44.1 for the overlap algorithm and of 50.9 for the surface matching algorithm were evident. The magnitude of E_t falls off as required overlap is reduced (corresponding to a looser Tanimoto distance cutoff distance and a higher shape threshold value) and more of the pool is selected (i.e. N_{aTotal} increases). However, a reasonable enrichment of 11.0 is observed when ~2% (1493 molecules) of the total data set (83,178 molecules) are matched using the surface overlap approach. While the magnitudes of reasonable enrichment figures would of course vary with particular applications, for a database search as described above, an enrichment of ten or more could be significant.

[115] The effectiveness of the database search was evaluated based upon designation of a particular molecule as “thrombin active” by the MDDR. However, several thousand compounds not labeled as “thrombin active” by the MDDR are annotated as "antiaggretory" or "GB2A/3B" inhibitors, labels which may also indicate thrombin activity. Accordingly exclusion of hits from compounds annotated as "antiaggretory" or "GB2A/3B" inhibitors from this study should result in an underestimation of enrichment obtained utilizing subshape matching in accordance with the present invention, and thus an underestimation of the true therapeutic performance of the subshape matching technique as compared with results from a more homologously assayed data set.

[116] Another possible application for embodiments of subshape matching in accordance with the present invention is in the field of template evaluation. Specifically, combinatorial chemistry approaches typically involve creating large numbers of organic compounds in parallel by linking a large number of chemical building blocks in all possible combinations.

[117] Typically, combinatorial synthesis utilizes a molecular scaffold as a starting point, mixing these scaffolds with molecules that react with the scaffold to form side chains. Discussion of the use of scaffolds in drug design is presented by Lee et al., "Scaffold Architecture and Pharmacophoric Properties of Natural Products and Trade Drugs: Application in the Design of Natural Product-Based Combinatorial Libraries", J. Comb. Chem, 3, 284-89 (2001), and by Lewell et al., "RECAP-Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry", J. Chem. Inf. Comput. Sci., 38, 511-522 (1998), both of which are incorporated by reference herein for all purposes.

[118] In deciding upon the particular combinatorial scaffold and molecules for additive reaction with same, a research may seek to ascertain the manner in which the synthesized molecule, including its side chains, will interact with a receptor or enzyme. Thus an embodiment of shape matching in accordance with the present invention, comprises the steps of identifying subshape similarity between a query shape defined by a molecule known to be active toward a particular receptor, and a target shape defined by a template or side chain to the template. Thus in the case of template evaluation, the query molecule could typically correspond in size to a drug-like molecule having a molecular weight of between about 400-700, with the target molecule corresponding to a molecule fragment having a molecular weight of 100 or less.

[119] A further possible application for embodiments of subshape matching in accordance with embodiments of the present invention is the superposition or alignment of molecules on the basis of their molecular shape, and sometimes also in combination with other molecular features such as the presence of certain chemical functionalities. Such spatial superpositions may not be immediately apparent to a user, particularly when viewing subshape similarities and taking into account conformational flexibility. For example, Fig. 13A shows the 2-D representations of two ligands (1tlp and ppp) which do not apparently share common structural features or shapes. However, experimental evidence has revealed the 1tlp and ppp ligands to be bound to the protein thermolysin in the manner indicated by the superimposed three-dimensional shapes of Fig. 13B

[120] While not intuitively obvious, such molecular superpositions may be useful for determining binding modes, or key interactions between ligand and receptor, see Bohm et al., "What Can We Learn from Molecular Recognition in protein-Ligand Complexes for the Design of New Drugs?", Angew Chem. Int. Ed. Engl. 35-2588-2614 (1996), hereby incorporated by reference for all purposes..

[121] Yet another example of a possible application for subshape similarity matching in accordance with embodiments of the present invention is in performing docking studies. Docking involves the creation of a shape of the space representing the receptor site itself, rather than the shape of a ligand known to bind to that receptor. In performing docking studies, complementarity in shape between the target molecule and the receptor is important because short contacts will typically result in high repulsive energies. Where the three-dimensional structure of the receptor is known, the shape of the active site can be deduced from the receptor's atomic coordinates using techniques and programs known in the art. For example, the use of protein structural information for drug design is often referred to as Structure-Based Drug Design. Recent examples of compounds discovered by such techniques are reviewed by Murcko et al., "Chapter 29 – Structure-Based Drug Design", Ann. Rep. Med. Chem. 34, 297-306 (1999), incorporated herein by reference for all purposes. Docking studies are described in detail in AR Leach, "Molecular Modeling, Principles and Applications", 2nd Ed, Prentice Hall, pp. 661-668 (2001), incorporated herein by reference for all purposes.

[122] Accordingly an embodiment of a method of performing docking studies in accordance with the present invention comprises the steps of identifying a target molecule exhibiting subshape similarity with a query shape defined by the binding volume present on a particular receptor structure. Such a binding volume would typically correspond in size to the shape occupied by a larger molecule (i.e. M.W. > 1000).

[123] Still another example of an application for subshape similarity matching in accordance with embodiments of the present invention is in the field of Qualitative Structure Activity Relations (QSAR). The QSAR technique is explained in detail by Hoekman et al in Exploring QSAR, ACS, Washington, D.C. (1995). In particular 3D QSAR studies, rely on alignment of multiple molecules with alignment superimposition often based on molecular shape. An example of using subshape technology for this application may involve utilizing a low energy conformation of the largest of the molecules as a query shape. The target shape can then be obtained from the remaining molecules to compare to the query shape. Once a

consistent alignment has been obtained for all the molecules, follow-up methods such as Comparative Molecular Field Analysis can be employed.

[124] Yet another example of an application for subshape similarity matching in accordance with embodiments of the present invention is in performing shape cataloging in conjunction with machine learning. One approach to shape cataloging is described in detail by Putta et al., "A Novel Shape-Feature Based Approach to Virtual Library Screening", J. Chem. Inform. and Comput. Sci. __, pp. ____(2002), incorporated by reference for all purposes herein.

[125] In an initial step of a shape cataloging process, a shape catalog is generated from three-dimensional conformations of molecules known to be active. This step may involve comparison of subshapes in accordance with an embodiment of the present invention to pick the most diverse set of shapes for the catalog. In such an application, shapes of conformers known to be active could serve as the query molecule shape and other conformers of known active compounds could serve as the target molecule shape.

[126] In a second step of a shape cataloging application, a shape in the catalog may be used as the query shape, with the target shape obtained from molecules in a input data set comprising both active and inactive molecules. In such a second step, if a match is found between the query shape and the target shape, locations of the chemical features on the target shape are marked. Based on these chemical feature locations and the subshape matched target molecules, a fingerprint in the form of a bit string can be generated for molecules of the input data set.

[127] In the third step of the shape cataloging process, the fingerprints are analyzed using various machine learning techniques to identify a small number of query shape and chemical feature location on them, that are important for activity. In a fourth step of the shape cataloging process, a query shape identified through machine learning analysis of the fingerprints can then be used as the query molecule shape to search through a database of target molecules to identify compounds of the database suitable for biological assaying.

[128] Once a target compound exhibiting subshape similarity with a query shape has been identified through a subshape matching method in accordance with an embodiment of the present invention, the target molecule may be further evaluated as a potentially useful lead candidate in the process of drug discovery. FIG. 12 is a flow chart showing the steps of a method 1200 for applying a subshape-matched target molecule in accordance with an embodiment of the present invention to identify possible drug leads.

[129] In the first step 1202, a target molecule exhibiting desirable subshape characteristics is identified as described in detail above. The first step 1202 of the method 1200 shown in FIG. 12 thus corresponds to step 422 of FIG. 4.

[130] In a second step 1204, the target molecule identified by subshape matching in accordance with an embodiment of the present invention is procured. The target molecule can be procured in a number of ways. One approach is to synthesize the molecule in the library. Such synthesis can comprise conventional techniques, or more efficiently can employ combinatorial synthesis strategies wherein large numbers of organic compounds are created in parallel by linking chemical building blocks in all possible combinations. Such combinatorial synthesis approaches may involve solid phase synthesis wherein the molecules are anchored to beads, or may involve solution phase synthesis wherein the molecules are present in solution. Either or both solid or solution phase combinatorial synthesis techniques could be utilized to procure a subshape-matched target molecule identified in accordance with embodiments of the present invention.

[131] Another alternative approach for procuring a target molecule identified by subshape matching in accordance with the present invention is to purchase existing molecules from commercial sources. Examples of commercial sources of molecules suitable for procuring members of a gene family screening library created in accordance with an embodiment of the present invention include, but are not limited to, Pharmacopeia Inc. of Princeton, New Jersey, Sigma-Aldrich Corp. of St. Louis, Missouri, Maybridge Plc.of Tintagel, Cornwall U.K., Chembridge Corp. of San Diego, CA, and Albany Molecular Research, of Albany, NY.

[132] In a third step 1206 of flowchart 1200, the procured target molecule identified by subshape matching in accordance with the present invention can be screened for activity. Screening of the target molecule can take the form of biological assays conducted outside of living tissue (*in vitro*). As is well known to one of skill in the art, examples of assay formats for measurement of enzyme activity or receptor binding include, but are not limited to, electrophoresis, scintillation proximity, ELISAs, immunoprecipitation, western blotting, and bead-based methods. Examples of detection techniques for application with biological assays include, but are not limited to, the use of time-resolved fluorescence, resonance energy transfer (FRET), fluorescence polarization, radioisotopic tracers, and chemiluminescent or colorimetric substrates. Other in vitro screening techniques for use in conjunction with gene family screening libraries created in accordance with the present invention include, but are not limited to, binding assays, enzyme activity assays, and cell-based assays such as functional assays and metabolism assays.

[133] One or more of the screening techniques described above can be performed with different levels of throughput. High-throughput screening of compounds is a standard approach in pharmaceutical research to discover new lead compounds for drug design. High-throughput screening typically involves the use of ninety-six or a greater number of wells per plate. Such high-throughput screening methods have discovered novel molecules, dissimilar to known ligands, that nevertheless bind to the target receptor at micromolar or submicromolar concentrations. Examples of the use of high throughput *in vitro* screening to identify active molecules of a screening library are described by McGovern et al., "A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening", J. Med. Chem. 45, 1712-1722 (2002), and Golebiowski et al, "Lead compounds discovered from libraries", Curr. Opin Chem. Biol.., 5, 273-284 (2001), both of which are incorporated by reference herein for all purposes. Medium or low-throughput formats can also be utilized to screen the target molecules identified by subshape matching in accordance with embodiments of the present invention.

[134] Alternatively, or in conjunction with *in vitro* testing, procured subshape-matched molecules in accordance with the embodiments of the present invention can be subjected to screening in living tissue (*in vivo*). Such *in vivo* assays include but are not limited to evaluation of a subshape-matched molecule activity in rodents, dogs, primates, or any other species. This evaluation may include testing of the molecules in a suitable pharmacological model of a particular disease state, wherein physiological or behavioral changes in an animal are monitored. Such animals may be normal (wild-type) or genetically-modified, or may be subject to a particular experimental protocol. Data produced from *in vivo* assays may include but is not limited to physical examination, histological (organ/tissue) or behavioral observations, post-mortem examinations, and gene-expression analyses from tissue samples of animals exposed to library molecules. For example, subshape-matched molecules may effectively reduce the size, weight and/or adipose tissue density of animals fed a high-fat diet, as a model for human obesity and diabetes, or may produce a response associated with reduced anxiety in a behavioral test, or may alter normal gene-expression in a given tissue as a result of interacting with an appropriate biological target.

[135] In addition to *in vitro* and *in vivo* testing of members of a target molecule revealed by subshape matching in accordance with embodiments of the present invention, screening "*in silico*" – within the silicon of the integrated circuits comprising a computer processor or memory, is emerging as an increasingly useful technique. *In silico* screening, also known as virtual screening, relies upon electronic representations of the molecules in two- or three-

dimensions, rather than upon the physical molecules themselves. *In silico* screening may permit a researcher to rapidly compare and evaluate similarity between subshape-matched target molecule and other structures, such as receptors or other molecules with previously-demonstrated activity against a particular receptor. While not replacing entirely bioassays that attempt to reproduce the *in vitro* and *in vivo* behavior of a molecule in chemical and biological environments, respectively, *in silico* screening has emerged as a useful tool for drug development. *In silico* screening is described in general by Terstappen et al. in "*In silico* research in Drug Discovery", Trends in Pharmacological Sciences, Vol. 22 No. 1 (2001), incorporated by reference herein for all purposes. An example of *in silico* screening of combinatorial libraries across a gene-family has been described by Aronov et al., "Virtual Screening of Combinatorial Libraries Across a Gene Family: in Search of Inhibitors of Giardia lambia Guanine PhosphoribosyltransferaseAntimicrob Agents Chemother., 45,2571-6 (2001), incorporated by reference herein for all purposes

[136] *In silico* screening of subshape-matched target molecules can utilize known ligand connectivities and/or fragments derived from known ligands. Examples of criteria for *in silico* screening include the use of structurally-definite molecular substructures (e.g., privileged [sub]structures), structurally-definite molecular fragments, structurally-definite chemical scaffolds or structurally-definite sidechains A description of the concept of privileged substructures is presented by Patchett et al., in "Chapter 26: Privileged Structures – An Update", Ann. Rep. Med. Chem 35, 289 (2000), hereby incorporated by reference in its entirety for all purposes.

[137] Alternatively, *in silico* screening can comprise searching the subshape-matched target molecules with at least one class of structurally-abstract molecule descriptors. The selection of the class of structurally-abstract molecule descriptors can be based on any suitable structurally-abstract characteristic, feature or property of a molecule. Therefore, structurally-abstract molecule descriptor classes include pharmacophore descriptors, atom path-length descriptors, BCUT descriptors, and other biophysical descriptors (e.g., solubility) known to one skilled in the art. A discussion of BCUT descriptors is given by Pearlman et al. in "Metric Validation and the Receptor-Relevant Subspace Concept", J. Chem. Inf. Comput. Sci, 39, 28-35 (1999), hereby incorporated by reference for all purposes.

[138] Target molecules identified by subshape matching which evidence desirable activity *in vitro, in vivo, in silico*, or in some combination thereof, against certain desired objectives are designated as 'hits', and may be validated and further optimized to identify leads and ultimately, drug candidates and drugs. A typical sequence of screening utilizing maximum efficiency of resources is initial screening of subshape matched target molecules *in silico*, followed by *in vitro* screening of subshape matched target molecules revealed as promising *in silico*, followed by *in vivo* screening of subshape matched target molecules revealed as promising *in vitro*. However, this order of testing is not required, and the various techniques could be employed in any order to screen a target molecule identified by subshape matching in accordance with an embodiment of the present invention, for suitability for use as a drug.

[139] While the above is a full description of the specific embodiments, various modifications, alternative constructions and equivalents may be used. Accordingly, the above description and illustrations should not be taken as limiting the scope of the present invention which is defined by the appended claims.

WHAT IS CLAIMED IS:

1. A method of comparing a query molecule shape with a target molecule shape, the method comprising:

sampling a distribution of local volume of the query molecule shape to generate a plurality of skeleton points, each skeleton point including a location and a direction, the direction determined by a principal axis of the sampled local volume distribution around the skeleton point;

sampling a distribution of local volume of the target molecule shape to generate a plurality of terminal points fewer in number than the skeleton points, each terminal point including a location and a direction, the direction determined by a principal axis of the sampled local volume distribution around the terminal point;

creating a query subshape triangle from three skeleton points;

creating a target subshape triangle from three terminal points;

matching the query subshape triangle and the target subshape triangle to determine an optimal translation and rotation of the target subshape triangle relative to the query subshape triangle to align skeleton and terminal points corresponding to vertices of the subshape triangles;

comparing directions of the aligned corresponding skeleton points and terminal points;

overlapping the query molecule shape and the target molecule shape by alignment of matched target and query subshape triangles; and

comparing the overlapped query and target molecule shapes.

2. The method of claim 1 further comprising:

calculating triangle edge distances from skeleton points of query subshape triangles and from terminal points of target subshape triangles; and

identifying a difference between query subshape triangle edges and target subshape triangle edges prior to matching the subshape triangles.

3. The method of claim 1 wherein:

each terminal point may further comprise a characteristic feature type representative of a chemical environment of the sampled local volume distribution around the terminal point;

each skeleton point may further comprise a characteristic feature type representative of a chemical environment of the sampled local volume distribution around the skeleton point; and

the method further comprises comparing feature types of corresponding terminal and skeleton points prior to overlapping the query molecule shape and the target molecule shape.

4. The method of claim 1 further comprising:

orienting the query molecule shape within a receptor defined by a macromolecule shape;

overlapping the query molecule shape and the target molecule shape by alignment of matched target and query subshape triangles; and

comparing an overlap between the target molecule shape and the macromolecule shape.

5. A method of searching three-dimensional target molecule shapes matching a query molecule shape, the method comprising:

providing a query molecule shape;

sampling a distribution of local volume of the query molecule shape to generate a plurality of skeleton points, each skeleton point including a location and a direction, the direction determined by a principal axis of the sampled local volume distribution of the query molecule shape;

sampling a distribution of local volume of a target molecule shape to generate a plurality of terminal points fewer in number than the skeleton points, each terminal point including a location and a direction, the direction determined by a principal axis of the sampled local volume distribution of the target molecule shape;

creating query subshape triangles from three skeleton points;

creating target subshape triangles from three terminal points;

determining triangle matching values for query/target subshape triangle pairs having an optimal translation and rotation of the target subshape triangle relative to the query subshape triangle;

determining net direction differences of corresponding skeleton points and terminal points only for query/target subshape triangle pairs whose triangle matching value satisfies a triangle matching threshold;

overlapping the query molecule shape and the target molecule shape by alignment of the query subshape triangle and the target subshape triangle only for query/target triangle pairs whose net direction difference satisfies a net direction difference threshold; and

comparing the overlapped query and target molecule shapes.

6. The method of claim 5 further comprising:

calculating triangle edge distances from skeleton points of query subshape triangles and from terminal points of target subshape triangles; and

generating an edge matching value from a difference between query subshape triangle edges and target subshape triangle edges, such that the triangle matching value is determined only for query/target subshape triangle pairs whose edge matching valve exceeds an edge matching threshold.

7. The method of claim 5 wherein each terminal point may further comprise a characteristic feature type representative of a chemical environment of the sampled local volume distribution of the target molecule shape, and each skeleton point may further comprise a characteristic feature type representative of a chemical environment of the sampled local volume distribution of the query molecule shape, the method further comprising:

determining a feature difference by comparing feature types of corresponding skeleton points and terminal points only of query/target subshape triangle pairs satisfying the triangle matching threshold, such that the direction difference is determined only for query/target subshape triangle pairs having a feature difference satisfying a feature difference threshold.

8. The method of claim 5 wherein the query molecule shape is oriented within a receptor defined by a macromolecule shape, and the query molecule shape and the target molecule shape are overlapped by alignment of target and query subshape triangles, the method further comprising comparing an overlap between the target molecule shape and the macromolecule shape.

9. A computer programming product for comparing a query molecule shape with a target molecule shape, the product comprising:

code for sampling a distribution of local volume of the query molecule shape to generate a plurality of skeleton points, each skeleton point including a location and a direction, the direction determined by a principal axis of the sampled local volume distribution around the skeleton point;

code for sampling a distribution of local volume of the target molecule shape to generate a plurality of terminal points fewer in number than the skeleton points, each terminal point including a location and a direction, the direction determined by a principal axis of the sampled local volume distribution around the terminal point;

code for creating a query subshape triangle from three skeleton points;

code for creating a target subshape triangle from three terminal points;

code for matching the query subshape triangle and the target subshape triangle to determine an optimal translation and rotation of the target subshape triangle relative to the query subshape triangle to align skeleton and terminal points corresponding to vertices of the subshape triangles;

code for comparing directions of the aligned corresponding skeleton points and terminal points;

code for overlapping the query molecule shape and the target molecule shape by alignment of matched target and query subshape triangles;

code for comparing the overlapped query and target molecule shapes;

an interface in communication with said codes and with a user; and

a computer readable storage medium for holding said codes and said interface.

10. The product of claim 9 further comprising:

code for calculating triangle edge distances from skeleton points of query subshape triangles and from terminal points of target subshape triangles; and

code for identifying a difference between query subshape triangle edges and target subshape triangle edges prior to matching the subshape triangles.

11. The product of claim 9 further comprising:

code for identifying a characteristic feature type representative of a chemical environment of the sampled local volume distribution around the terminal point;

code for identifying a characteristic feature type representative of a chemical environment of the sampled local volume distribution around the skeleton point; and code for comparing feature types of corresponding terminal and skeleton points prior to overlapping the query molecule shape and the target molecule shape.

12. The product of claim 9 further comprising:

code for orienting the query molecule shape within a receptor defined by a macromolecule shape;

code for overlapping the query molecule shape and the target molecule shape by alignment of matched target and query subshape triangles; and

code for comparing an overlap between the target molecule shape and the macromolecule shape.

13. A method of identifying a molecule as a drug lead candidate, the method comprising:

identifying a query molecule shape;

identifying a target molecule shape;

sampling a distribution of local volume of the query molecule shape to generate a plurality of skeleton points, each skeleton point including a location and a direction, the direction determined by a principal axis of the sampled local volume distribution around the skeleton point;

sampling a distribution of local volume of the target molecule shape to generate a plurality of terminal points fewer in number than the skeleton points, each terminal point including a location and a direction, the direction determined by a principal axis of the sampled local volume distribution around the terminal point;

creating a query subshape triangle from three skeleton points;

creating a target subshape triangle from three terminal points;

matching the query subshape triangle and the target subshape triangle to determine an optimal translation and rotation of the target subshape triangle relative to the query subshape triangle to align skeleton and terminal points corresponding to vertices of the subshape triangles;

comparing directions of the aligned corresponding skeleton points and terminal points;

overlapping the query molecule shape and the target molecule shape by alignment of matched target and query subshape triangles;

comparing the overlapped query and target molecule shapes; and procuring the target molecule where comparison of the query and target molecule shapes indicates a threshold similarity in the subshapes of the target and query molecules.

14. The method of claim 13 further comprising:

calculating triangle edge distances from skeleton points of query subshape triangles and from terminal points of target subshape triangles; and

identifying a difference between query subshape triangle edges and target subshape triangle edges prior to matching the subshape triangles.

15. The method of claim 13 wherein:

each terminal point may further comprise a characteristic feature type representative of a chemical environment of the sampled local volume distribution around the terminal point;

each skeleton point may further comprise a characteristic feature type representative of a chemical environment of the sampled local volume distribution around the skeleton point; and

the method further comprises comparing feature types of corresponding terminal and skeleton points prior to overlapping the query molecule shape and the target molecule shape.

16. The method of claim 13 wherein:

prior to overlapping the target molecule shape and the query molecule shape, encoding the query molecule shape into a first volume shell and the target molecule shape into a second volume shell, the first and second volume shells corresponding to a surface of the query molecule and of the target molecule, respectively; and

comparing the overlapped query and target molecule shapes comprises comparing overlap between the first and second volume shells.

17. The method of claim 13 wherein:

identifying the query molecule shape comprises identifying a known conformer of a first ligand as bound to a receptor;

identifying the target molecule shape comprises identifying a conformer of a second ligand known to be active against the receptor; and comparing the overlapped query and target molecule shapes comprises performing a conformer analysis to determine a likely conformer of the second ligand as bound to the receptor.

18. The method of claim 13 wherein:

identifying the query molecule shape comprises identifying a known conformer of a first ligand as bound to a receptor;

identifying the target molecule shape comprises identifying a conformer of a molecule selected from a database; and

comparing the overlapped query and target molecule shapes comprises performing a database search to determine a likelihood that the molecule will bind to the receptor.

19. The method of claim 13 wherein:

identifying the query molecule shape comprises identifying a known ligand to a receptor;

identifying the target molecule shape comprises identifying a scaffold molecule suitable for combinatorial synthesis; and

comparing the overlapped query and target molecule shapes comprises determining the suitability of the scaffold molecule for synthesizing molecules to test for activity against the receptor.

20. The method of claim 13 wherein:

identifying the query molecule shape comprises identifying a shape from a shape catalog complied from conformers of molecules known to be active against a receptor;

identifying the target molecule comprises identifying a conformer of a molecule selected from a pool of a mixture of molecules that are active and inactive against the receptor; and

comparing the overlapped query and target molecule shapes comprises compiling a catalog of bit-string fingerprints representing a location of chemical features in molecules of the pool.

21. The method of claim 13 wherein the query molecule shape is identified on the basis of at least one of activity demonstrated by a query molecule toward a particular receptor or family of receptors, similarity in structure between the query molecule and a

molecule known to demonstrate activity toward the particular receptor, and similarity in structure between the query molecule and a receptor volume defined by the particular receptor or a family of receptors.

22. The method of claim 13 wherein the target molecule shape is identified on the basis of at least one of presence in a available database of molecules, activity demonstrated by the target molecule toward a particular receptor or family of receptors, a known three-dimensional structure, and similarity in structure between the target molecule and a molecule demonstrating activity toward the particular receptor.

23. The method of claim 13, wherein:

the query and target molecule shapes identified are of approximately the same size; and

comparing the overlapped query and target molecule shapes comprises comparing a volume of the overlapped target and query molecule shapes to determine a protrusion distance.

24. The method of claim 13, wherein:

the query and target molecule shapes identified are of substantially different sizes; and

comparing the overlapped query and target molecule shapes comprises comparing a volume of surface shells of the overlapped target and query molecule shapes to determine a Tanimoto distance.

25. The method of claim 13, wherein the procuring step comprises obtaining a physical sample of the target molecule.

26. The method of claim 25, wherein the physical sample of the target molecule is obtained through the technique selected from the group consisting of purchasing the physical sample from a commercial vendor, isolating the physical sample from a natural source, and synthesizing the molecule in the laboratory.

27. The method of claim 25, wherein the target molecule is synthesized utilizing combinatorial chemistry techniques.

28. The method of claim 13, wherein the procuring step comprises:

generating a three-dimensional representation of a conformer of the target molecule in space; and

storing the three-dimensional representation of the conformer in a computer-readable storage medium.

29. The method of claim 13 further comprising screening the procured target molecule for activity.

30. The method of claim 29, wherein the screening step is selected from the group comprising determining an activity of the target molecule against another molecule through in vitro or in vivo experimentation.

31. The method of claim 29, wherein the screening step comprises comparison of a three-dimensional representation of a conformer of the target molecule with a three-dimensional representation of another molecule.

32. The method of claim 31, wherein the screening step comprises comparison of a three-dimensional representation of a conformer of the target molecule with a three-dimensional representation of a portion of a macromolecule defining a receptor binding volume.

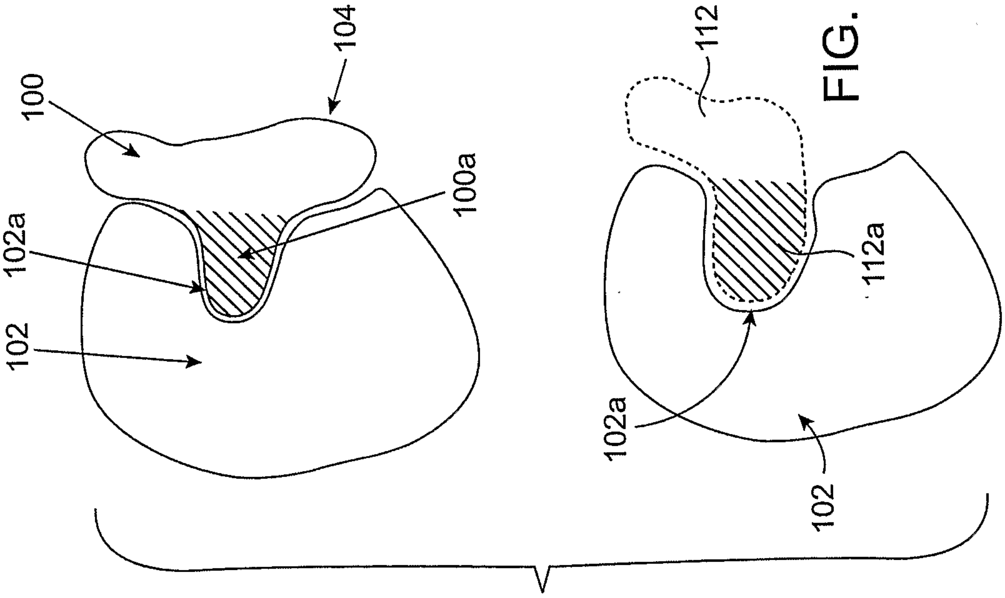
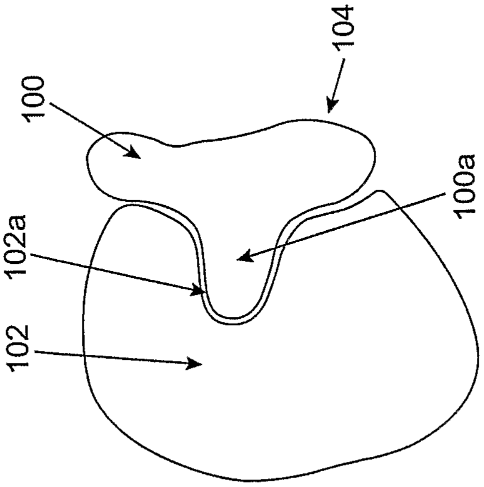
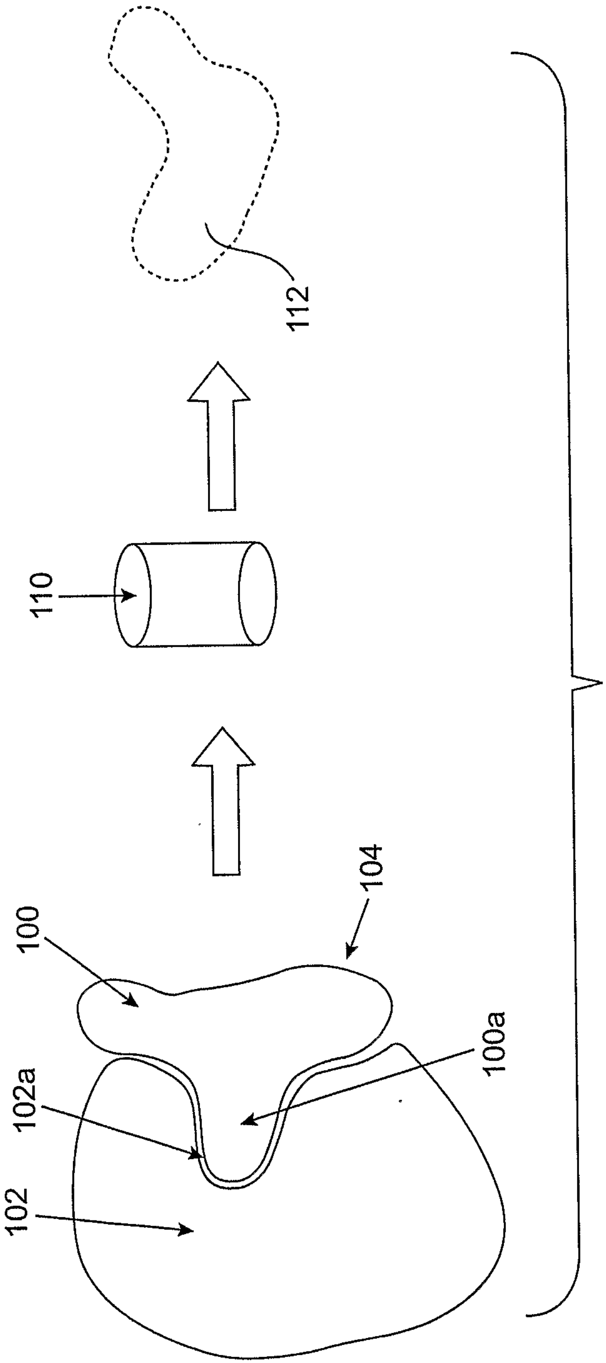


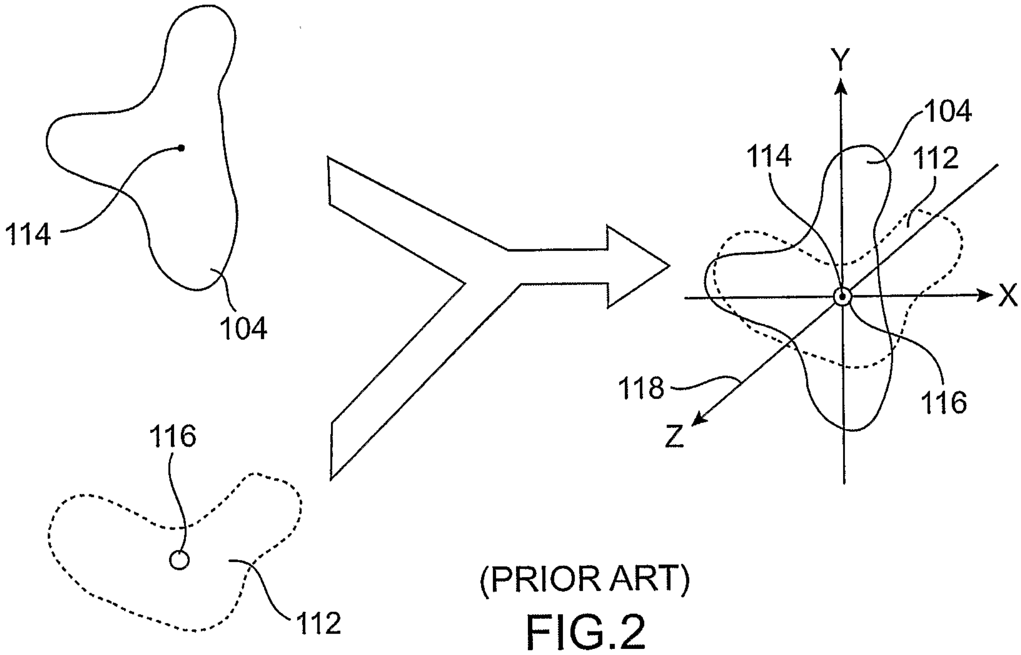
FIG. 1C



(PRIOR ART)
FIG. 1A



(PRIOR ART)
FIG. 1B



(PRIOR ART)
FIG.2

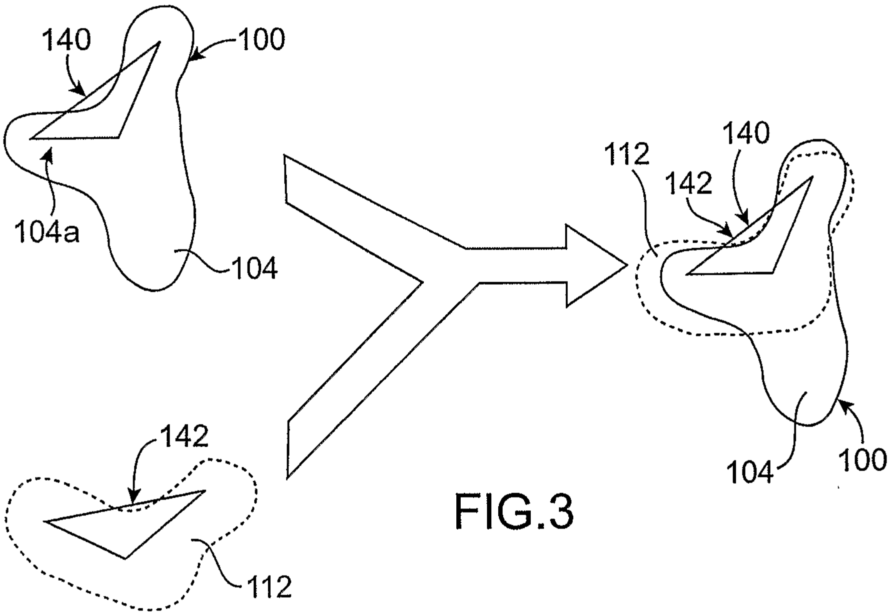


FIG.3

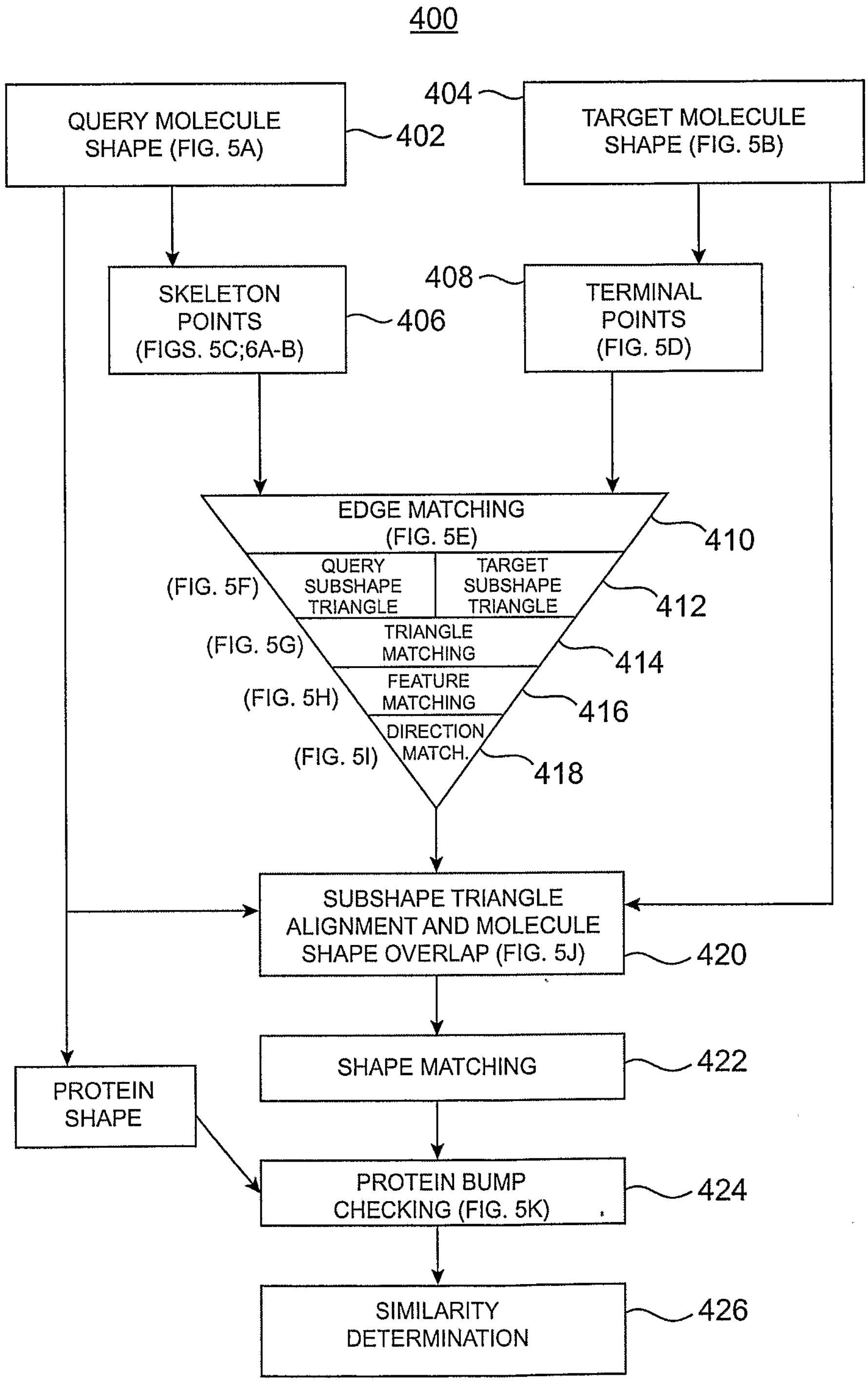

400
QUERY MOLECULE SHAPE (FIG. 5A)
402
404
TARGET MOLECULE SHAPE (FIG. 5B)
SKELETON POINTS (FIGS. 5C;6A-B)
406
408
TERMINAL POINTS (FIG. 5D)
EDGE MATCHING (FIG. 5E)
410
(FIG. 5F)
QUERY SUBSHAPE TRIANGLE
TARGET SUBSHAPE TRIANGLE
412
(FIG. 5G)
TRIANGLE MATCHING
414
(FIG. 5H)
FEATURE MATCHING
416
(FIG. 5I)
DIRECTION MATCH.
418
SUBSHAPE TRIANGLE ALIGNMENT AND MOLECULE SHAPE OVERLAP (FIG. 5J)
420
SHAPE MATCHING
422
PROTEIN SHAPE
PROTEIN BUMP CHECKING (FIG. 5K)
424
SIMILARITY DETERMINATION
426

FIG. 4

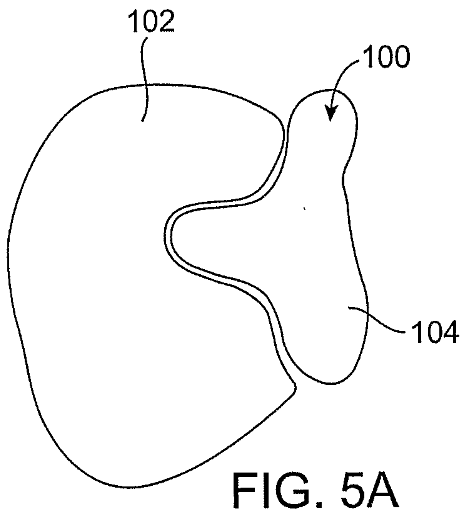


FIG. 5A

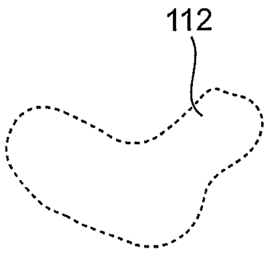


FIG. 5B

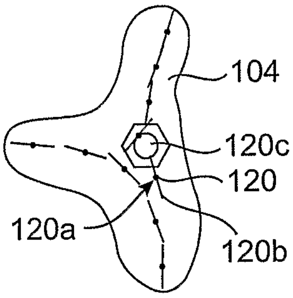


FIG. 5C

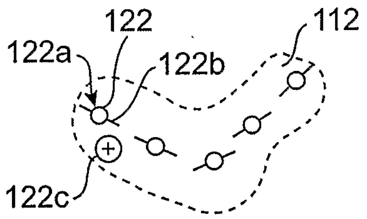


FIG. 5D

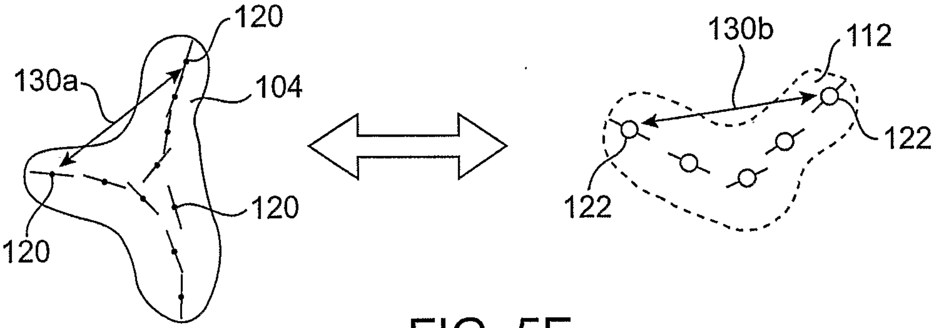


FIG. 5E

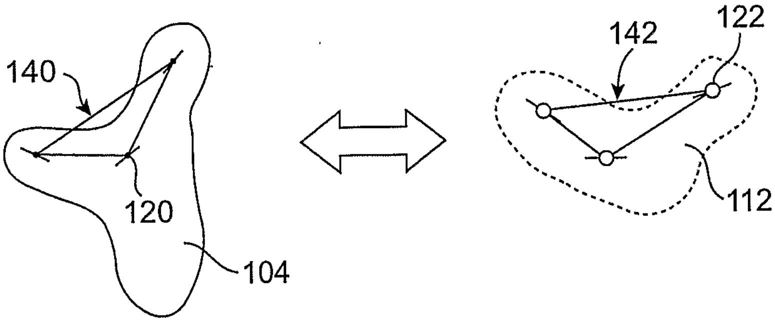


FIG. 5F

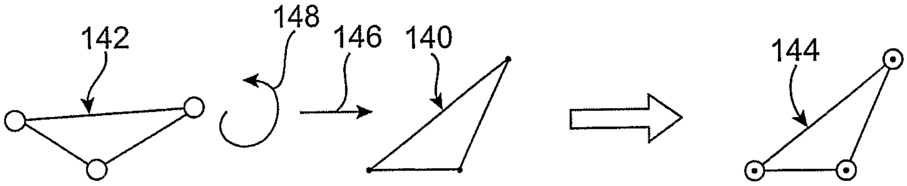


FIG. 5G

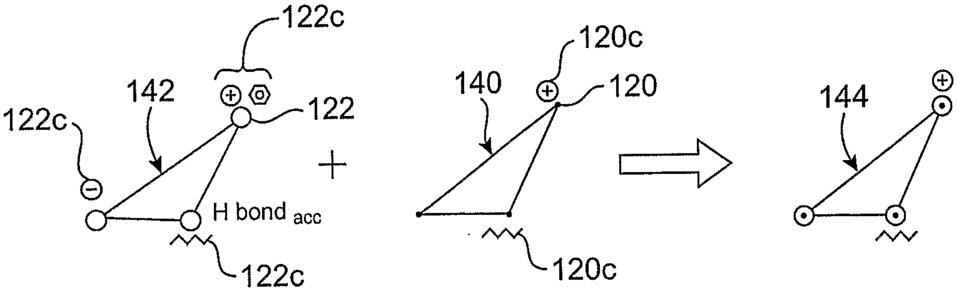


FIG. 5H

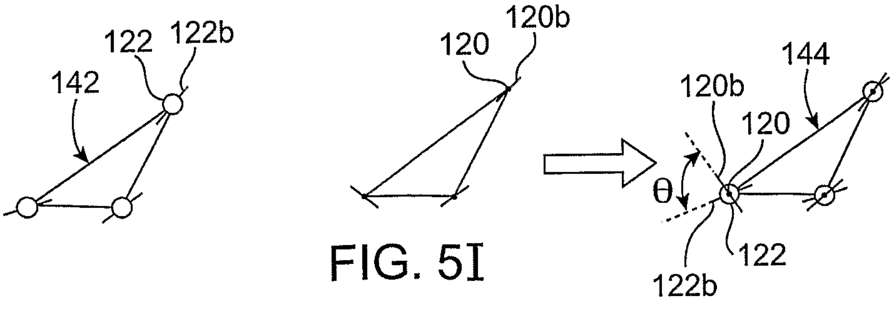


FIG. 5I

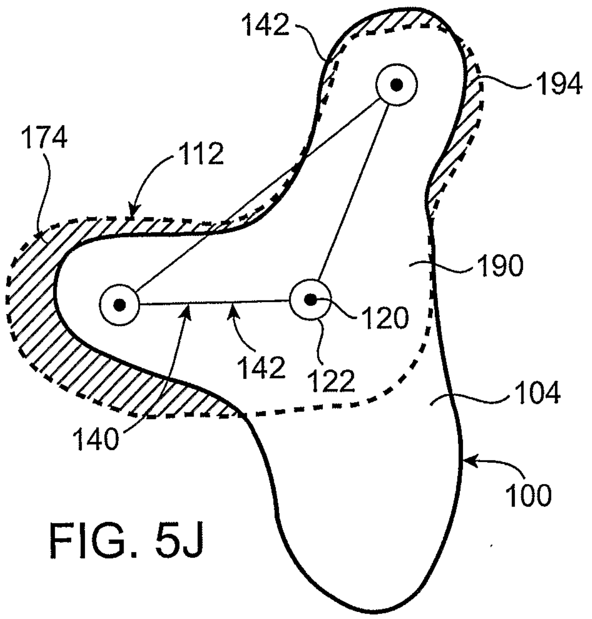


FIG. 5J

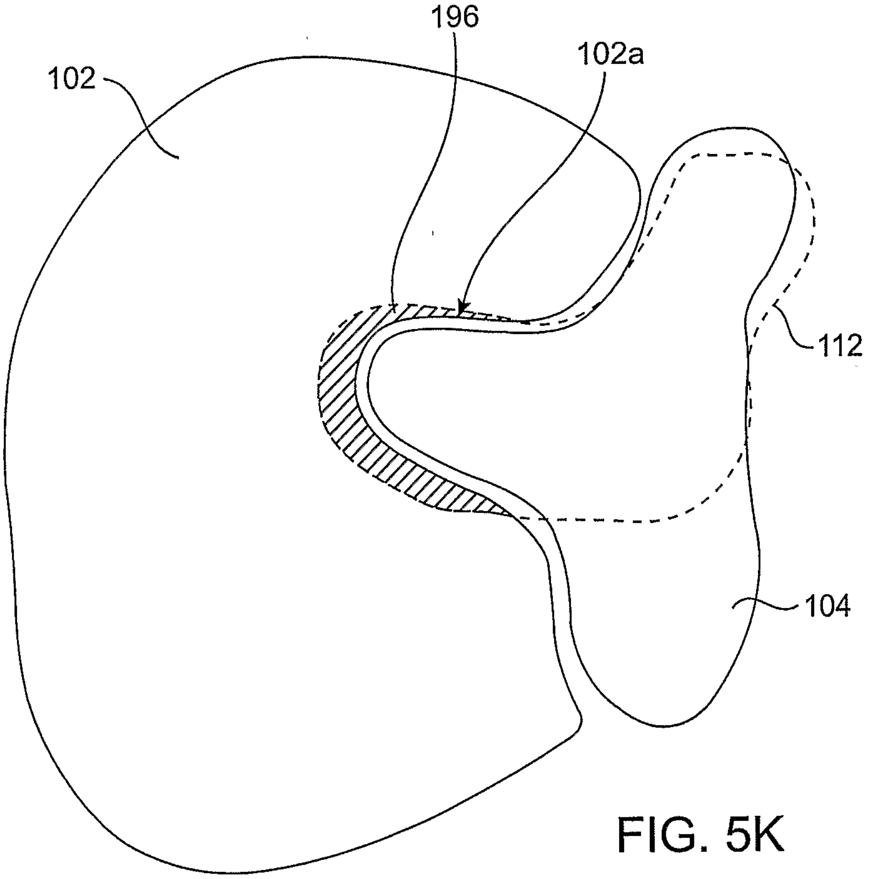


FIG. 5K

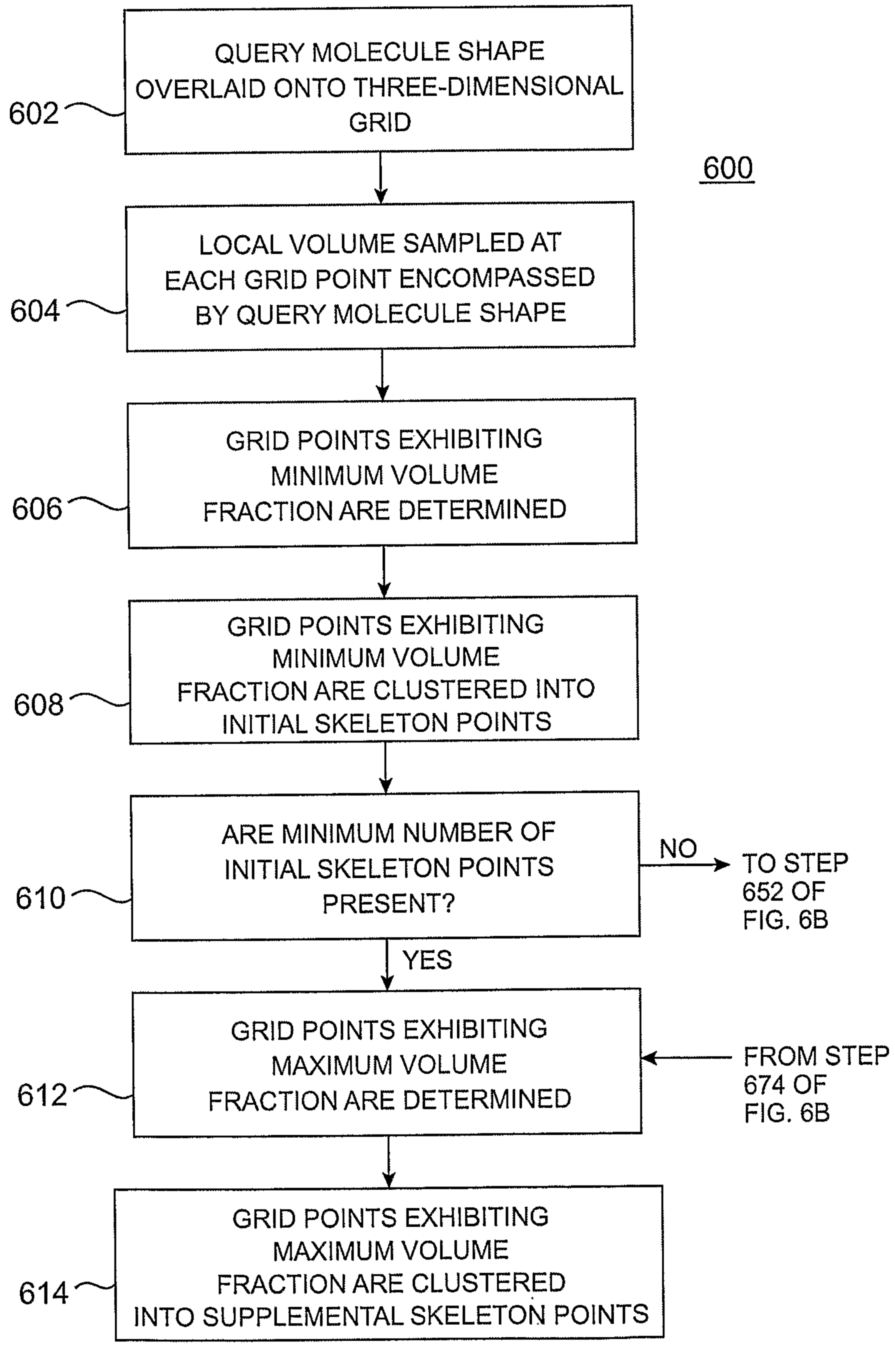

QUERY MOLECULE SHAPE OVERLAID ONTO THREE-DIMENSIONAL GRID
602
600
LOCAL VOLUME SAMPLED AT EACH GRID POINT ENCOMPASSED BY QUERY MOLECULE SHAPE
604
GRID POINTS EXHIBITING MINIMUM VOLUME FRACTION ARE DETERMINED
606
GRID POINTS EXHIBITING MINIMUM VOLUME FRACTION ARE CLUSTERED INTO INITIAL SKELETON POINTS
608
ARE MINIMUM NUMBER OF INITIAL SKELETON POINTS PRESENT?
610
NO
TO STEP 652 OF FIG. 6B
YES
GRID POINTS EXHIBITING MAXIMUM VOLUME FRACTION ARE DETERMINED
612
FROM STEP 674 OF FIG. 6B
GRID POINTS EXHIBITING MAXIMUM VOLUME FRACTION ARE CLUSTERED INTO SUPPLEMENTAL SKELETON POINTS
614

FIG. 6A

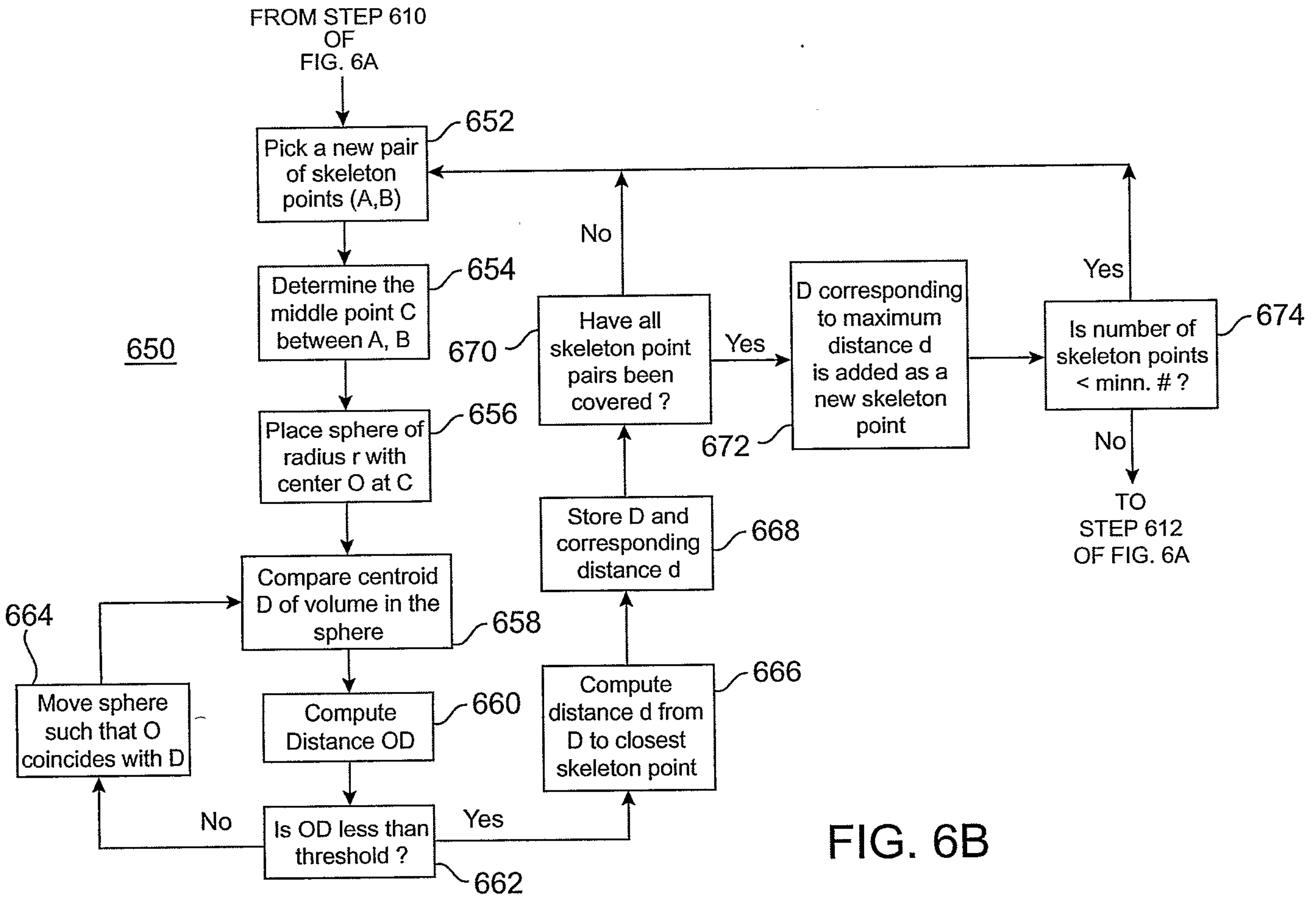

FROM STEP 610 OF FIG. 6A
650
Pick a new pair of skeleton points (A,B)
652
Determine the middle point C between A, B
654
Place sphere of radius r with center O at C
656
Compare centroid D of volume in the sphere
658
Compute Distance OD
660
Is OD less than threshold ?
662
No
Yes
Move sphere such that O coincides with D
664
Compute distance d from D to closest skeleton point
666
Store D and corresponding distance d
668
Have all skeleton point pairs been covered ?
670
No
Yes
D corresponding to maximum distance d is added as a new skeleton point
672
Is number of skeleton points < minn. # ?
674
Yes
No
TO STEP 612 OF FIG. 6A

FIG. 6B

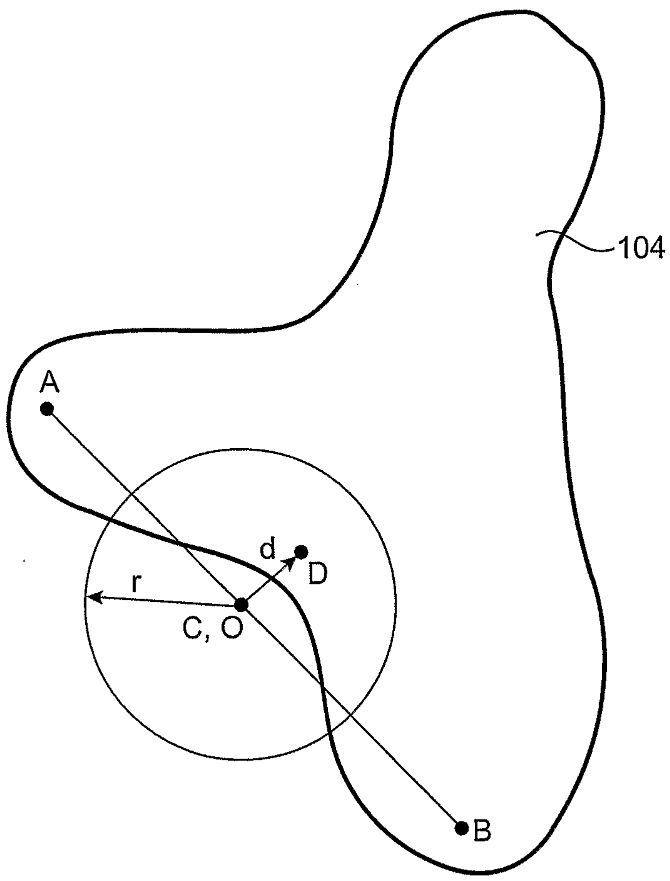


FIG. 6C

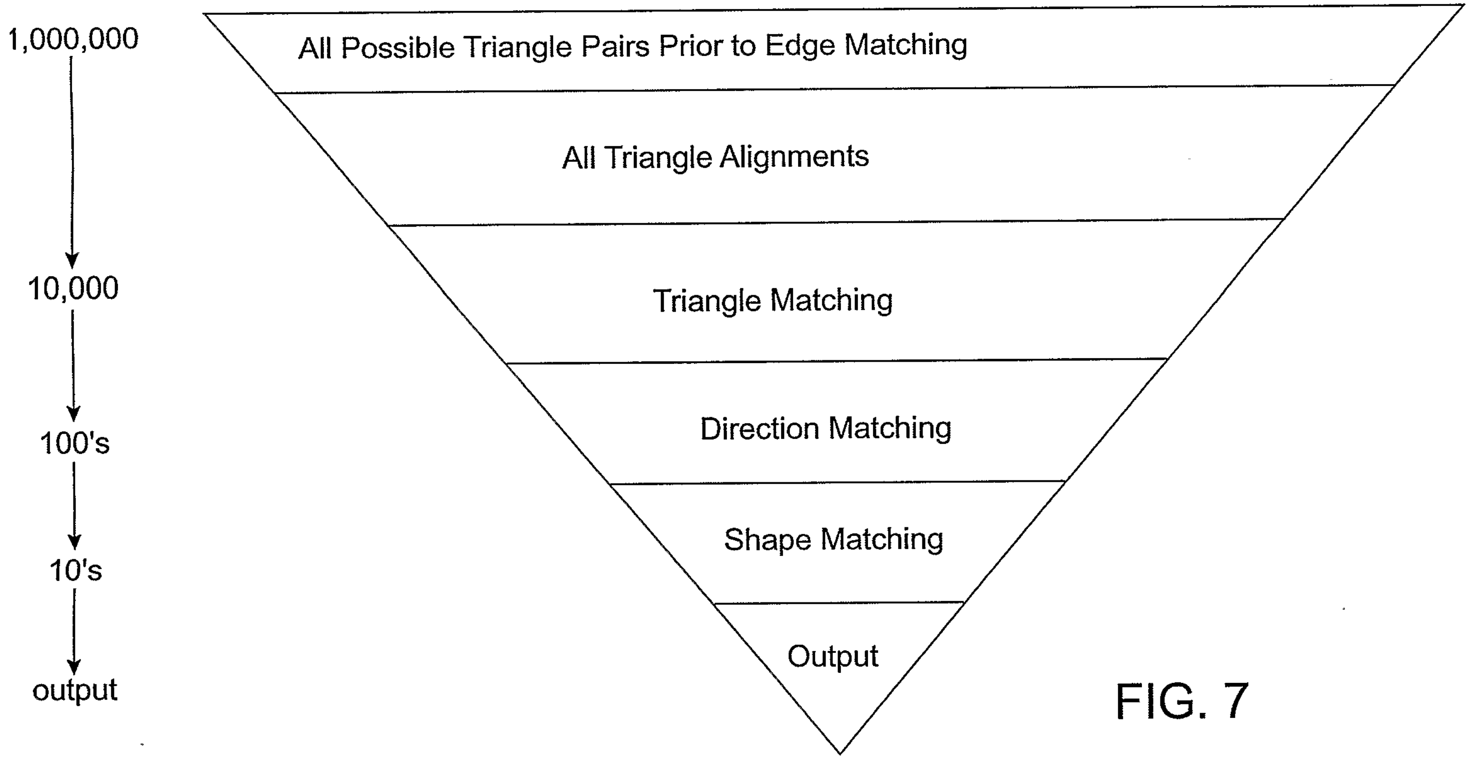

1,000,000
10,000
100's
10's
output
All Possible Triangle Pairs Prior to Edge Matching
All Triangle Alignments
Triangle Matching
Direction Matching
Shape Matching
Output

FIG. 7

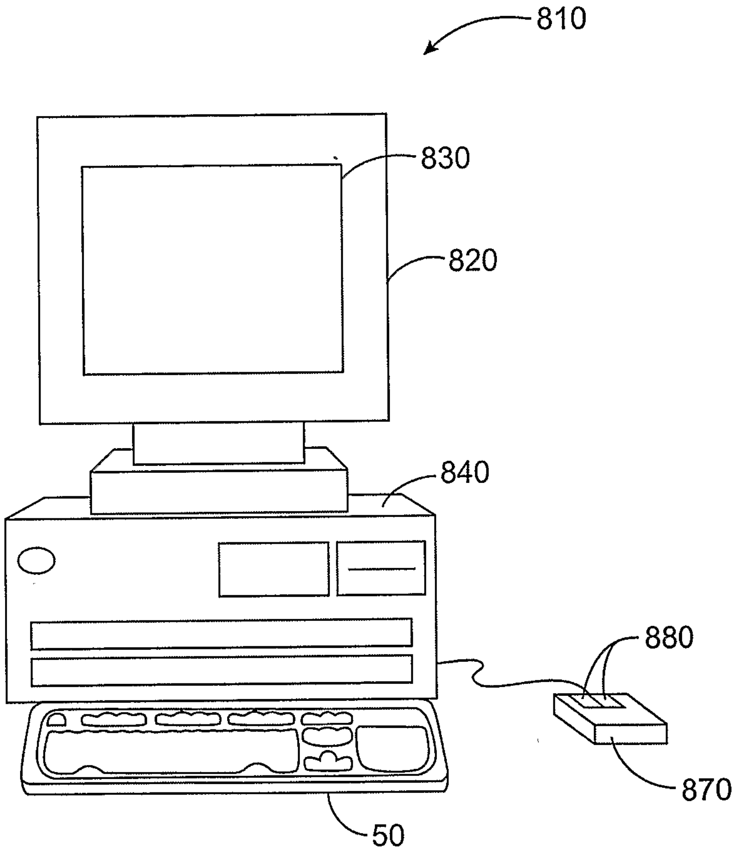


FIG. 8A

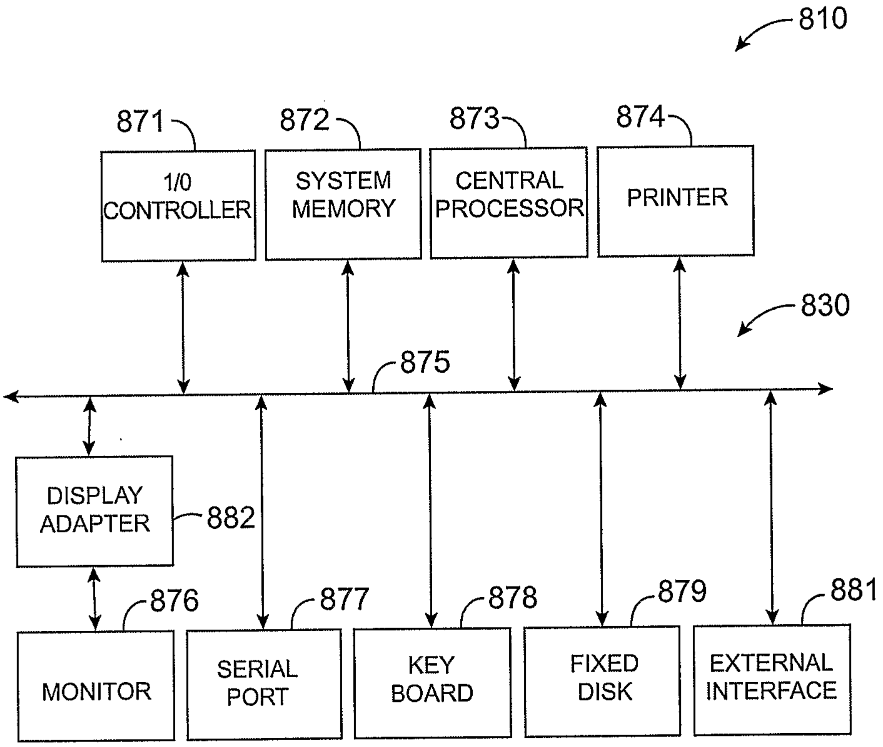

810
871
872
873
874
1/0 CONTROLLER
SYSTEM MEMORY
CENTRAL PROCESSOR
PRINTER
830
875
DISPLAY ADAPTER
882
876
877
878
879
881
MONITOR
SERIAL PORT
KEY BOARD
FIXED DISK
EXTERNAL INTERFACE

FIG. 8B

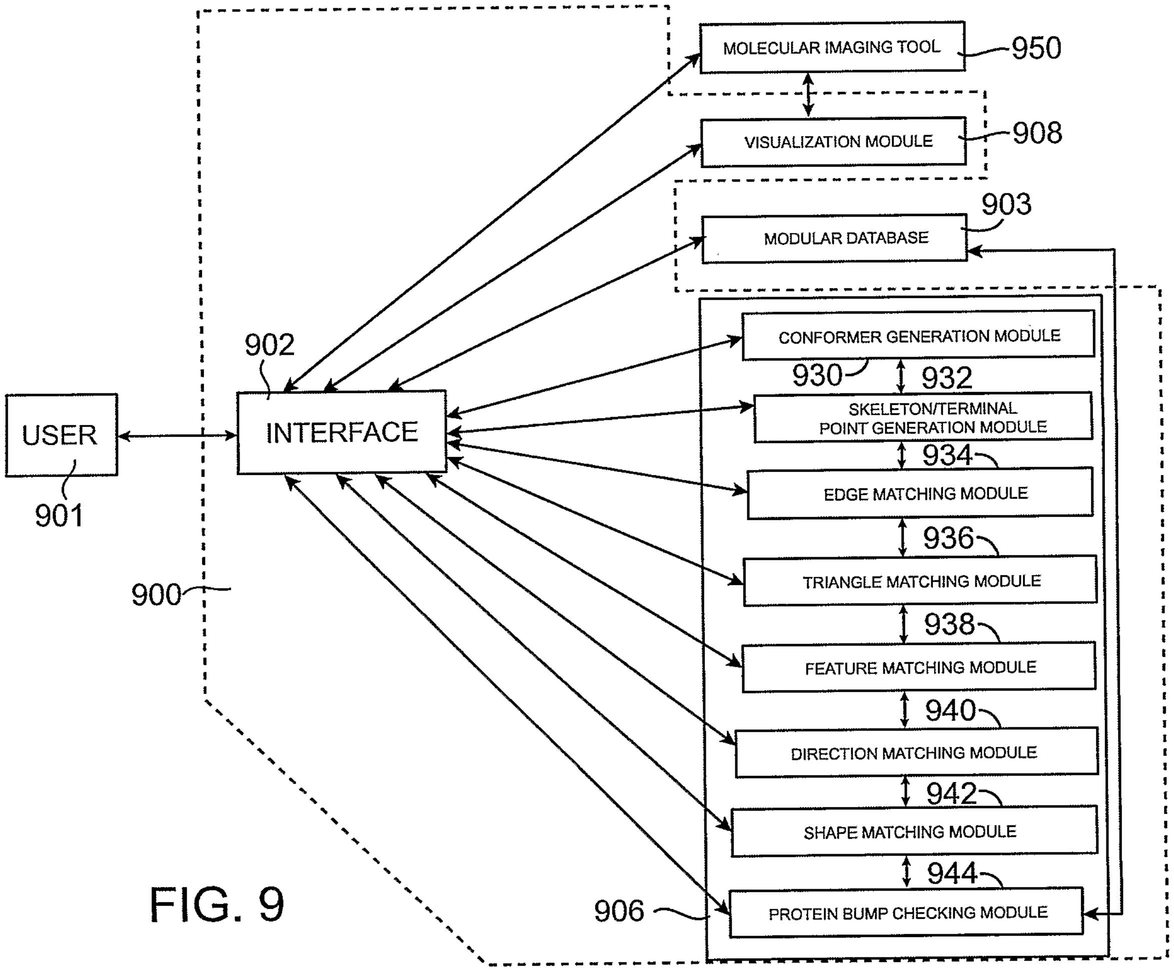
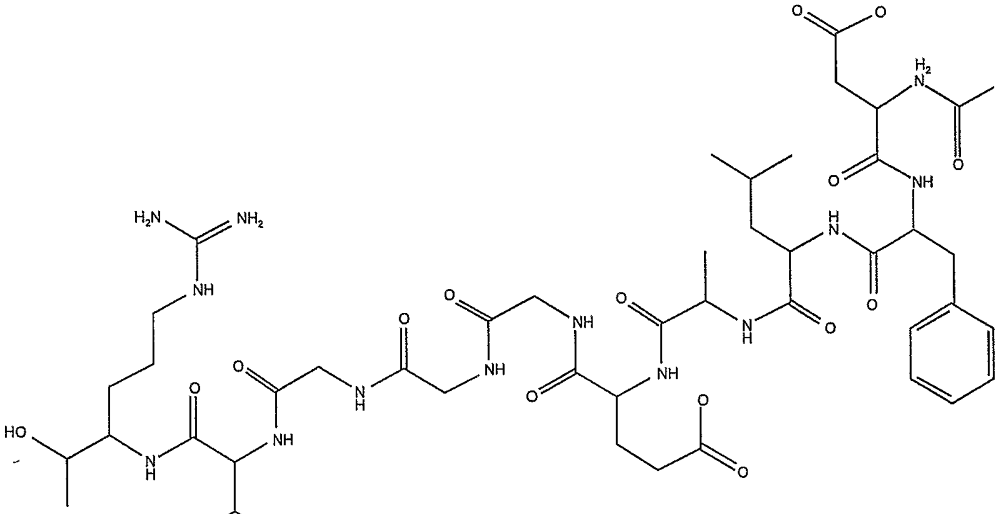

USER
901
INTERFACE
902
900
MOLECULAR IMAGING TOOL
950
VISUALIZATION MODULE
908
MODULAR DATABASE
903
CONFORMER GENERATION MODULE
930
932
SKELETON/TERMINAL POINT GENERATION MODULE
934
EDGE MATCHING MODULE
936
TRIANGLE MATCHING MODULE
938
FEATURE MATCHING MODULE
940
DIRECTION MATCHING MODULE
942
SHAPE MATCHING MODULE
944
PROTEIN BUMP CHECKING MODULE
906

FIG. 9



Fibrinogen

FIG. 10A

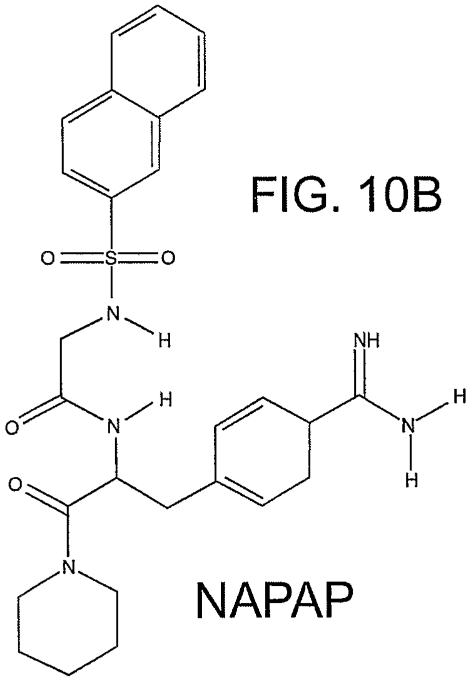
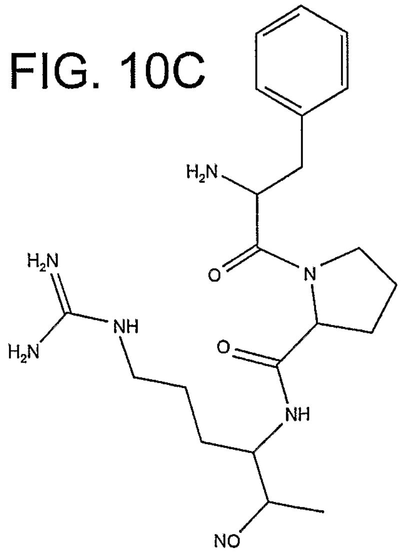


FIG. 10B

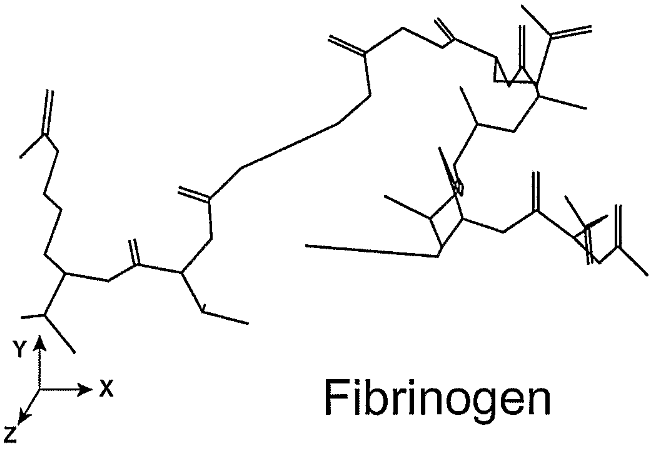
NAPAP

FIG. 10C



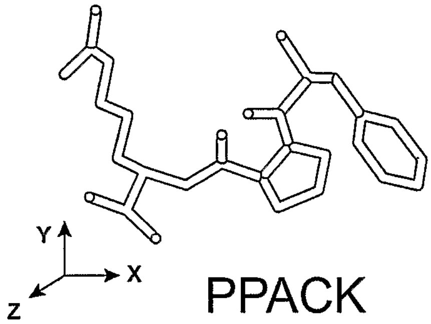
PPACK

FIG. 11A



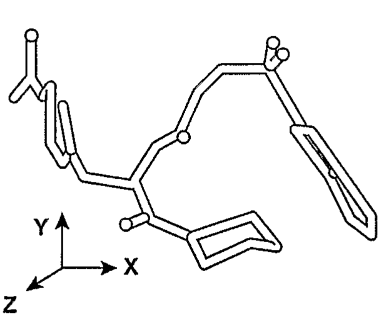
Fibrinogen

FIG. 11B



PPACK

FIG. 11C



NAPAP

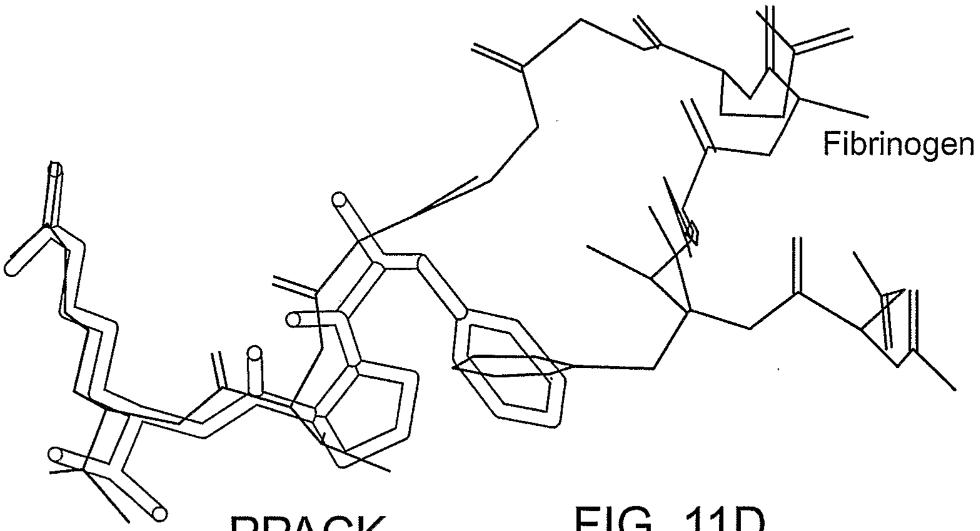


FIG. 11D

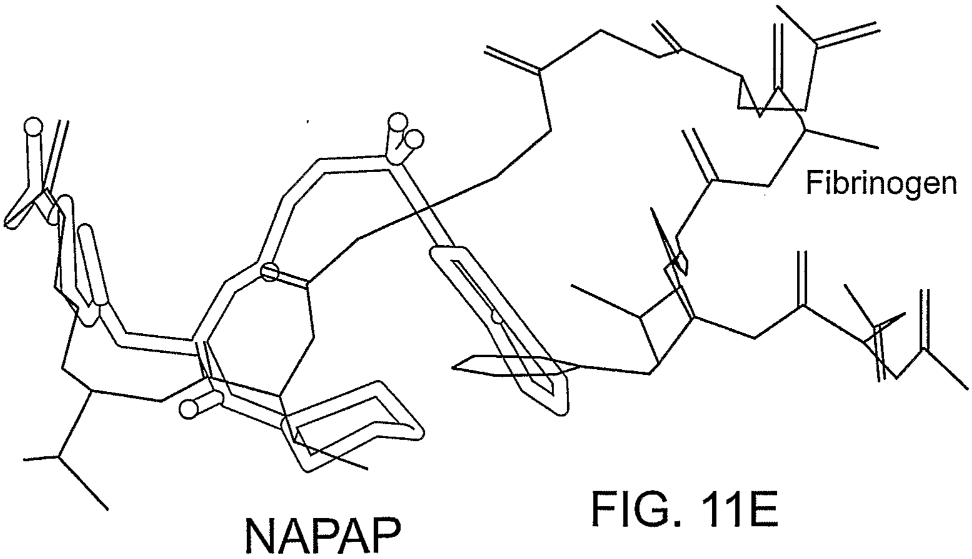


FIG. 11E

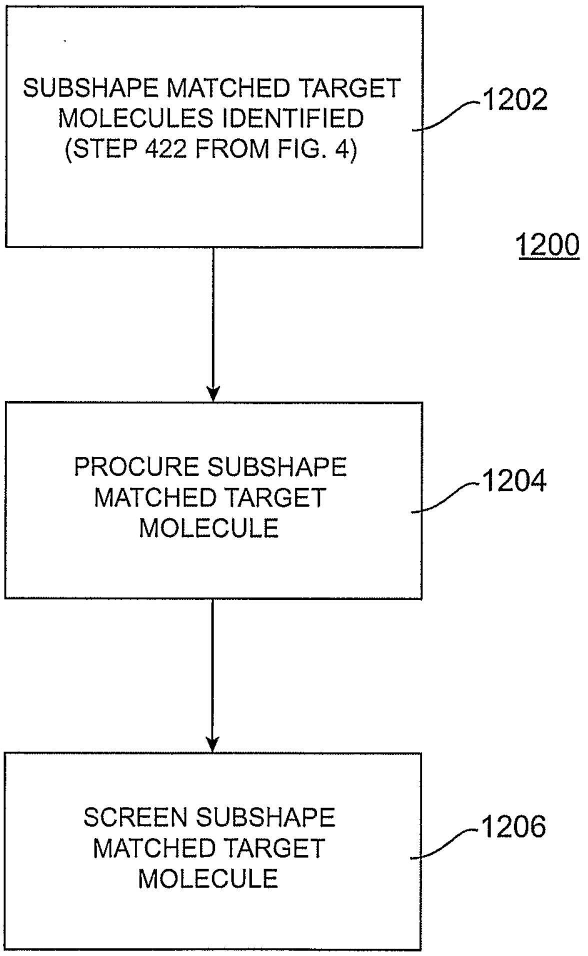
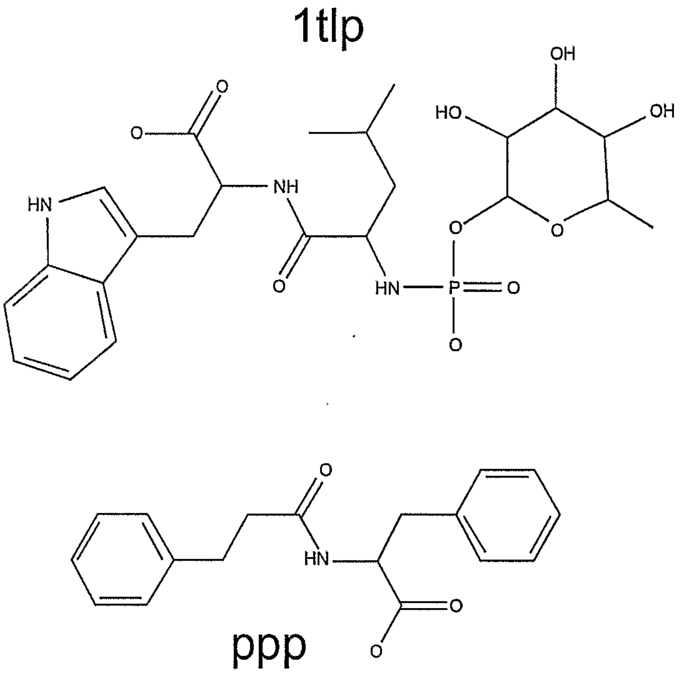
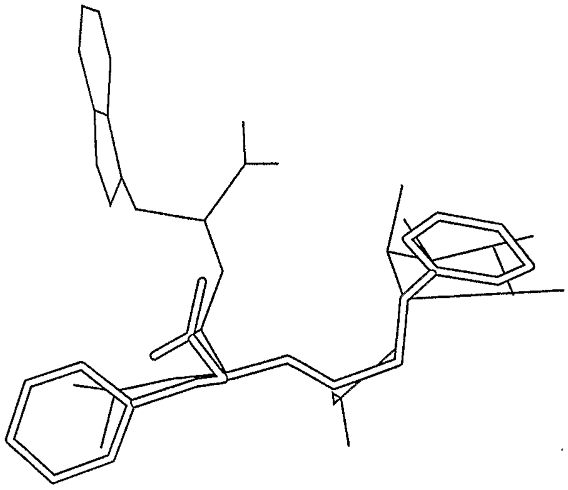

SUBSHAPE MATCHED TARGET MOLECULES IDENTIFIED (STEP 422 FROM FIG. 4)
1202
1200
PROCURE SUBSHAPE MATCHED TARGET MOLECULE
1204
SCREEN SUBSHAPE MATCHED TARGET MOLECULE
1206

FIG. 12



(PRIOR ART)
FIG. 13A



(PRIOR ART)
FIG. 13B

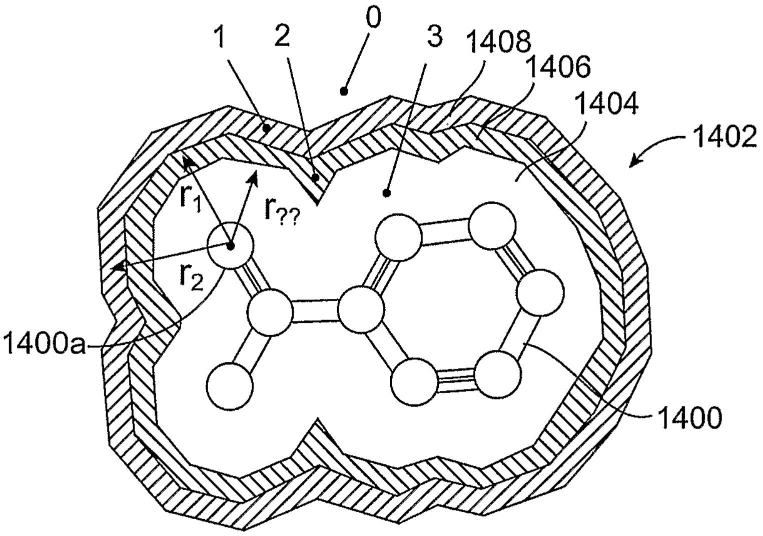


FIG. 14A

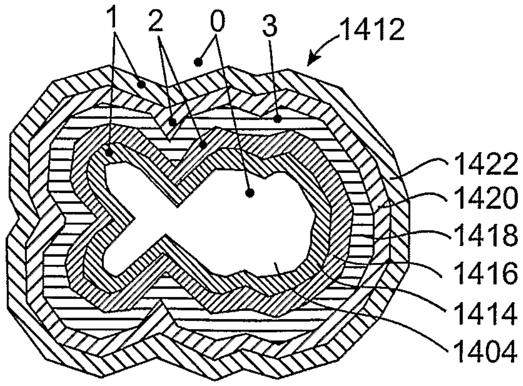


FIG. 14B