US 20130184163A1

(19) **United States**

(12) **Patent Application Publication**

Ross et al.

(10) **Pub. No.: US 2013/0184163 A1**

(43) **Pub. Date: Jul. 18, 2013**

(54) **METHOD FOR IDENTIFYING COMPOUNDS**

(75) Inventors: **Guenther Ross**, Freising (DE); **Udo Ottmann**, Mertingen (DE)

(73) Assignee: **PRIAXON AG**, Munich (DE)

(21) Appl. No.: **13/702,660**

(22) PCT Filed: **Jun. 7, 2011**

(86) PCT No.: **PCT/EP11/02783**

§ 371 (c)(1),
(2), (4) Date: **Mar. 28, 2013**

(30) **Foreign Application Priority Data**

Jun. 7, 2010 (EP) 10005857.7

Publication Classification

(51) **Int. Cl.**
C40B 20/00 (2006.01)

(52) **U.S. Cl.**
CPC ***C40B 20/00*** (2013.01)
USPC **506/2**

(57) **ABSTRACT**

The present invention relates to a method for identifying compounds comprising the steps of: (a) providing a set of compounds; (b) optionally selecting a sub-set from the set of compounds based on one or more specific compound properties; (c) generating a 3D structure of each of the compounds provided and/or selected in step (a) or (b); (d) encoding each 3D structure; (e) providing at least one known compound having at least one desired property and/or providing a target molecule; (f) encoding the 3D structure of (each of) the known compound(s) provided in step (e) and/or the active site of the target molecule provided in step (e); (g) comparing said encoded 3D structure(s) of step (d) with the encoded 3D structure(s) of step (f); and (h) selecting all compounds falling within a specified similarity range.

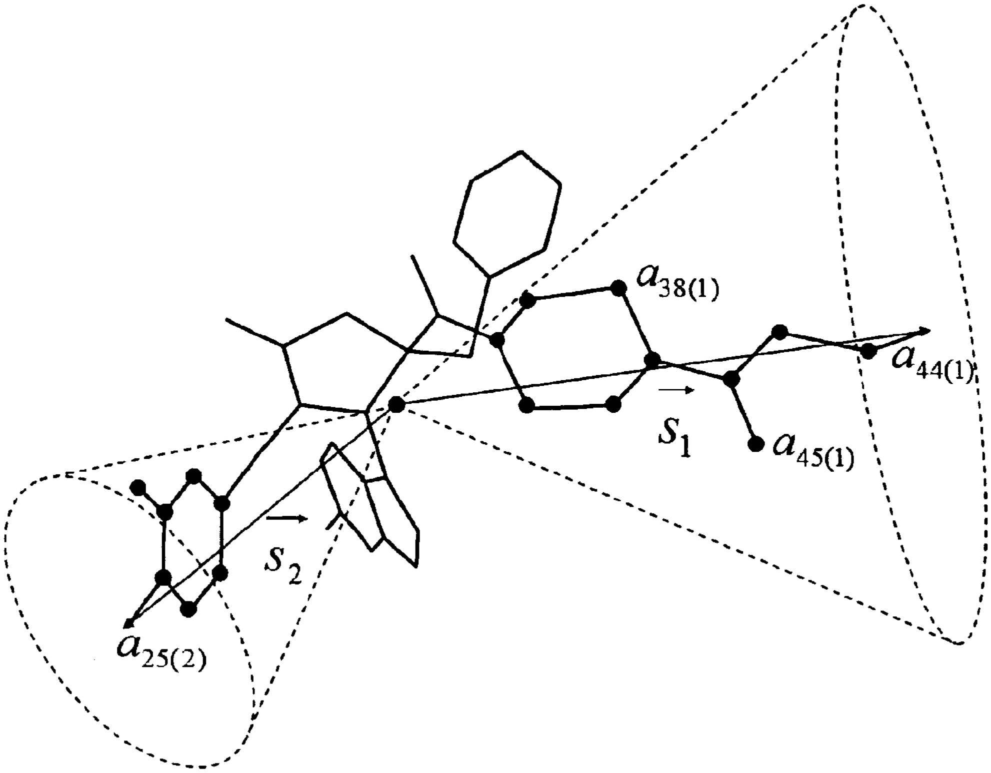


FIGURE 1: Construction of "super-substituents" represented by shape vectors

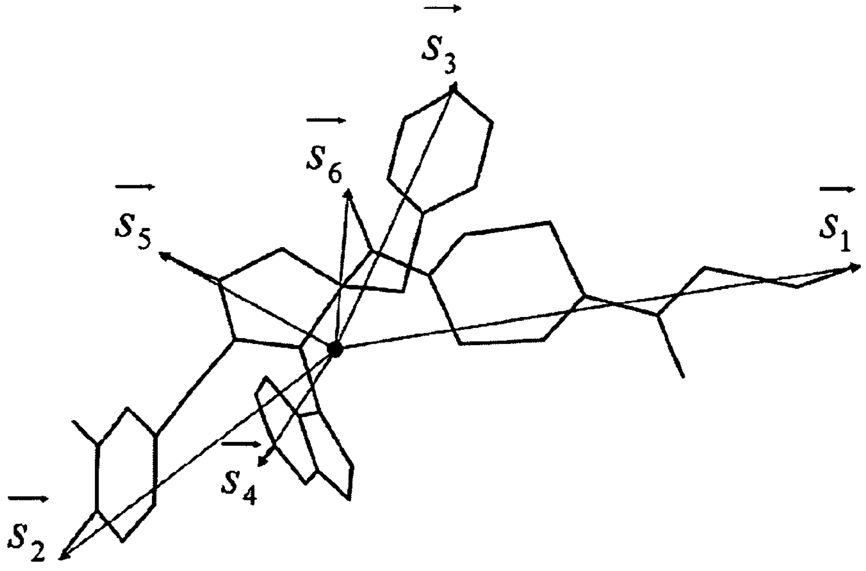


FIGURE 2: Distorted octahedral orientation of “super-substituents”

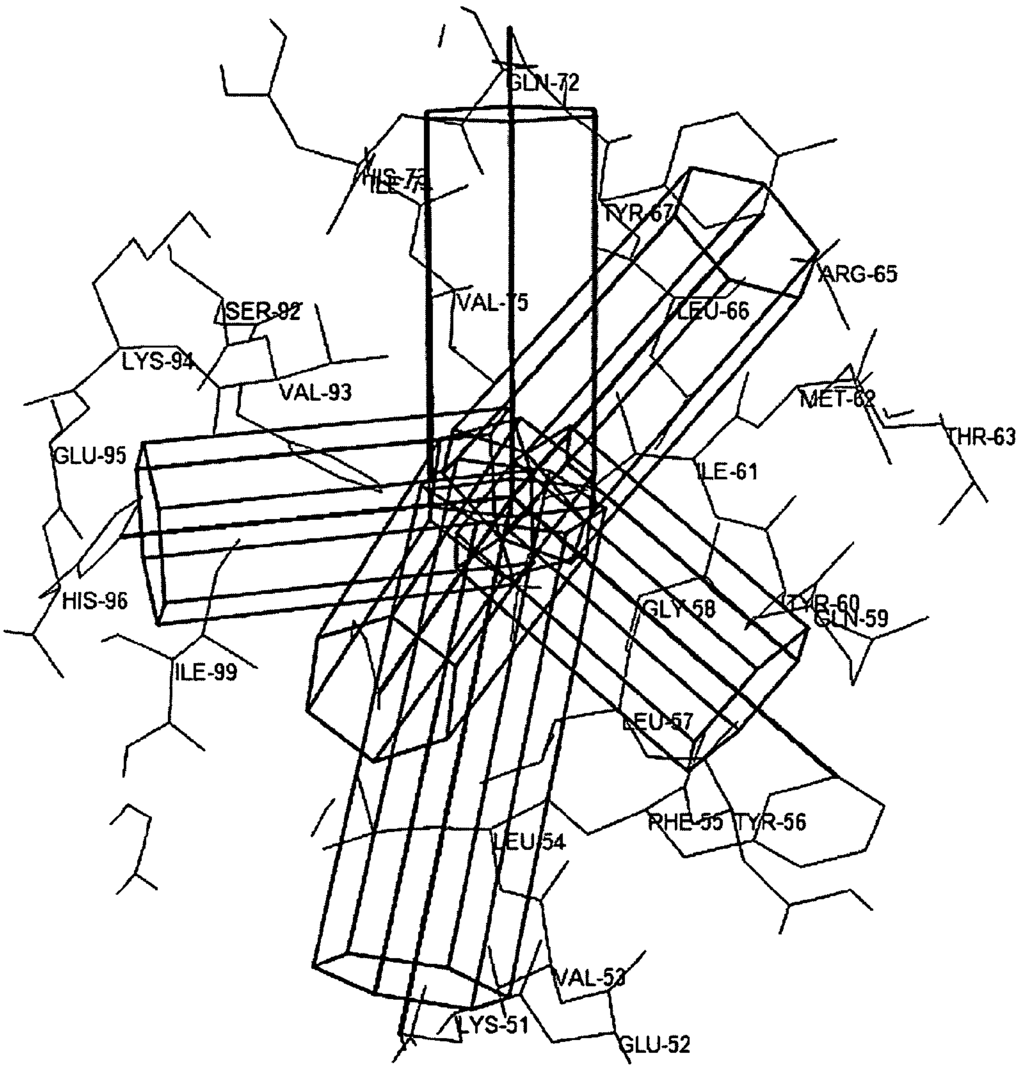


FIGURE 3: Regions of ligand accessible space calculated for the active site of mdm2

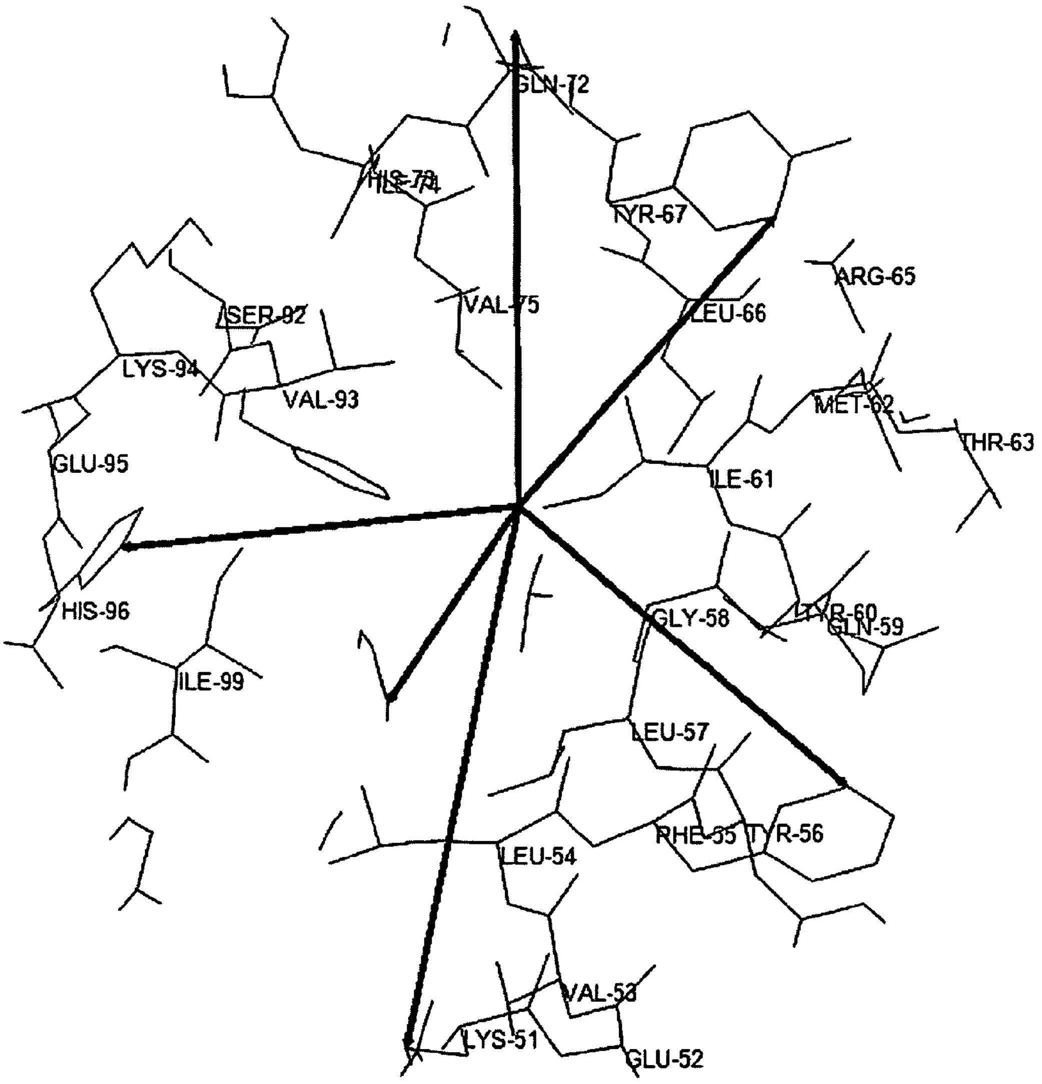


FIGURE 4: Shape vectors of the DPSM descriptor of the active site of mdm2

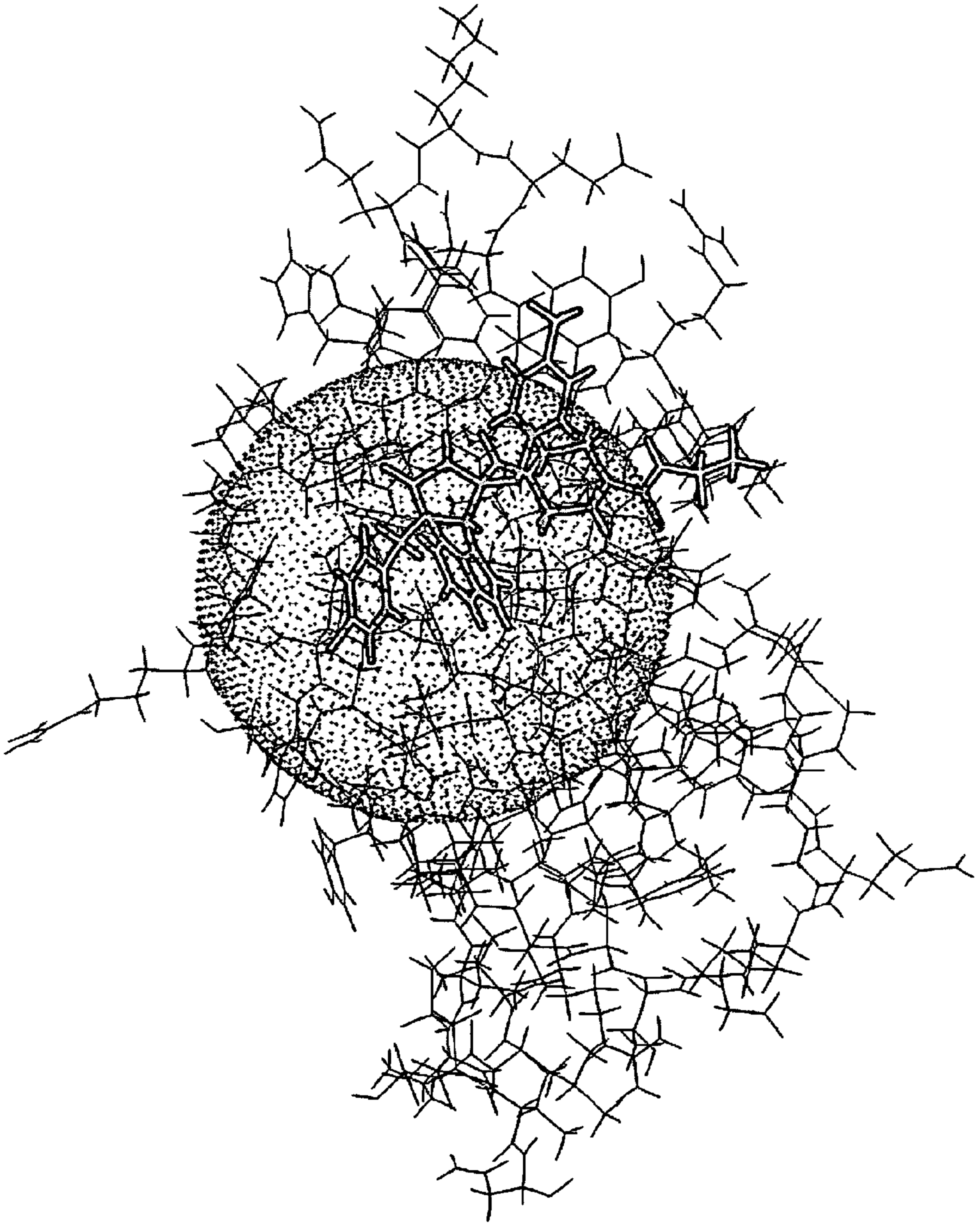


FIGURE 5: Active site of mdm2 predicted by the algorithm

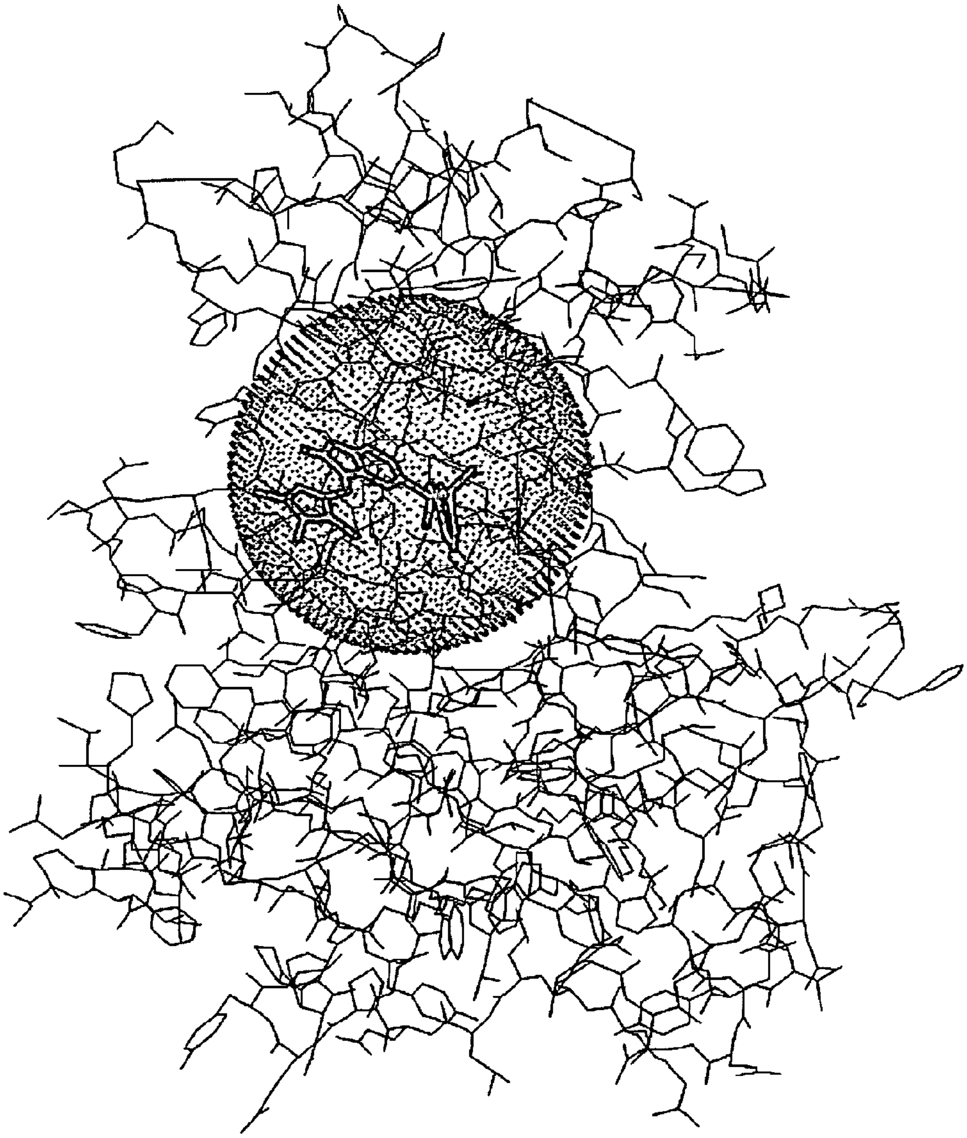


FIGURE 6: Active site of c-met predicted by the algorithm

METHOD FOR IDENTIFYING COMPOUNDS

[0001] In the drug discovery process the identification of new, active chemical entities is the key step to success. During the last two decades the application of "brute force" approaches such as high throughput screening has not yielded the desired results. In consequence smarter, more focused and less resource consuming technologies are required.

[0002] Once a target has been identified and passed the first stages of validation and investigation this knowledge base has to be used as efficiently as possible to discover and develop structural classes of compounds that show activity on the target and can be developed into clinical candidates and ultimately marketed drugs.

[0003] A rational approach to this task has to rest on two bases: The knowledge of chemical reactions and accessible structural classes that are innovative enough to allow room for development; and a technology that enables the straightforward identification of the right molecules and in consequence the right reactions for a given target. By now (May 5, 2010) 53,404,695 compounds have been registered in CAS (source: CAS homepage). Bearing in mind that many of these compounds have been published as examples for general chemical synthesis processes that could be applied to a much broader set of starting materials—consequently leading to a multitude of possible products—it is clear that astronomic numbers of accessible structures have to be processed and searched, even if obvious drug-likeness criteria are imposed. Only with powerful chemoinformatic tools this huge set of accessible structures can be searched efficiently.

[0004] The principal task of any computational search process is to reduce the billions of accessible structures to a final number that can be handled manually using human "Medchem intelligence" (some 100). This enables the selection, synthesis and biological testing of a reasonable and affordable number of compounds with a high probability of success. It is clear that the search algorithm and the molecular descriptors it uses to represent the compounds and their properties in silico is the crucial element of this process. An ideal balance between computing speed and accuracy has to be found to obtain high value computational hits within a reasonable time and cost frame.

[0005] The selection process can be broken down into three stages: In a first step the compounds are filtered for simple key data such as molecular weight, lipophilicity, polar surface area etc. which allow a rough indication specific classification. Still many millions of compounds are in the search set. In a second step, some hundreds of these millions are selected by different means for the third step, in silico docking. Finally the highest scoring compounds from step three are refined manually by molecular modeling to obtain candidates for synthesis and biological testing.

[0006] While step one is rather trivial and highly advanced software packages for docking and molecular modeling are available for step three, step two has a great potential for improvement. In this step most of the compounds are eliminated to achieve a reduction to numbers that can be handled reasonably in the laborious step three process. Consequently the method in step two needs to be fast enough to handle millions of compounds and accurate enough to select the most promising 0.1-0.01 percent of these for refinement.

[0007] During the last three decades the number of available 3D protein structures or protein/ligand complexes has grown rapidly. To be able to exploit this knowledge accurately there is a strong need for computational methods that rely on molecular descriptors encoding 3D structural information.

[0008] In contrast to the "data-explosion" in the field of chemical structure elucidation the number of available molecular 3D descriptors is poor, especially if compared to the great number of well defined 2D descriptors that rely on chemical connectivity only.

[0009] The situation gets even worse, when looking for a 3D descriptor that encodes the active binding site of a target protein and further, so far a complementary pair of molecular 3D descriptors that mirrors the geometric and chemical complementarity of a ligand/target interaction does not exist.

[0010] So the current state of chemoinformatics is far away from providing complete solutions that include molecular 3D information in the process of rational drug design.

[0011] This situation is typically caused by some general problems a designer of a molecular 3D descriptor is faced with. First, the processing of 3D information is by its very nature computationally more expensive than methods that only rely on connectivity, i.e. most of the computational code executed in 2D descriptor calculation can be implemented as fast integer operations whereas 3D descriptor calculation depends strongly on more time consuming floating point operations. Second, molecules in 3 dimensional space have an absolute position and orientation, but the 3D descriptor representation has to be independent of these coordinates. This is the so called "requirement of translational and rotational invariance". Third, a molecular 3D descriptor should also allow to encode a "distribution of physico-chemical properties in space", in other words: a pure geometric descriptor is a poor abstraction of a molecule, because a molecule is not just a set of points in space. And last, setting up a model that can describe the immanent complementarity of ligand/target interaction is not trivial. So the design of efficient and accurate molecular 3D descriptors is an art in itself.

[0012] The methods presented here address all these problems in an exact and flexible approach. Further, algorithms are described that are fast, robust and intuitive. Moreover they reflect the natural complementarity of a ligand molecule and the corresponding active site of a target protein. The methods are generally applicable to a broad range of potential targets and the search set of chemical structures is only limited by chemical and computational feasibility. Still there are fields in which their performance is exceptionally high. From the target side these are protein-protein interactions, which require a particularly precise description of complex 3D properties. From the side of the structural search space chemistries based on multicomponent reactions are especially attractive, because they allow the straightforward assembly of highly decorated (substituted) scaffolds that by now have only been poorly exploited in drug discovery efforts.

[0013] The present invention provides a method for identifying (or selecting) compounds (especially useful compounds). This method comprises the steps of:

[0014] (a) providing a set (space) of compounds;

[0015] (b) optionally selecting a sub-set (sub-space) from the set of compounds based on one or more specified compound properties;

[0016] (c) generating a 3D structure of (each of) the compounds provided and/or selected in step (a) or (b);

[0017] (d) encoding each 3D structure;

[0018] (e) providing at least one known compound having at least one desired property and/or providing a target molecule;

[0019] (f) encoding the 3D structure of each of the known compound(s) provided in step (e) and/or encoding the active site of the target molecule provided in step (e);

[0020] (g) comparing said encoded 3D structure(s) of step (d) with the encoded 3D structure(s) of step (f); and

[0021] (h) selecting all compounds falling within a specified similarity range.

[0022] Additionally, this method may further comprise the steps of:

[0023] (i) optionally selecting a further sub-set of the compounds provided in step (h) based on one or more specific compound properties;

[0024] (j) preparing the remaining (selected) compounds and testing the same;

[0025] (k) optionally repeating steps (g) to (j) or (h) to (j).

[0026] Preferably, steps (a) to (k) (as well as any further steps given herein) are carried out in the order given.

[0027] Preparing (or synthesizing) the compounds in step (j) may e.g. be performed manually in a laboratory (e.g. in a chemical laboratory by a chemist). As an alternative, the selected compounds of step (h) and/or (i) may be prepared automatically by a synthesizer for automated chemical synthesis.

[0028] Testing the compounds in step (j) may e.g. be performed manually in a laboratory (e.g. in a biological laboratory by a biologist). As an alternative, the testing may be carried out automatically, e.g. by a screening robot.

[0029] Testing (in step j) is preferably carried out in vitro.

[0030] Especially preferably, the present invention also relates to a method for screening for identifying compounds comprising the above mentioned steps (a) to (k).

[0031] Further preferably the present invention also relates to a method for synthesizing compounds comprising the above mentioned steps (a) to (k).

[0032] In step (e), also a set of known compounds each having at least one desired property may be provided instead of the at least one compound.

[0033] The compound(s) provided in step (e) is/are preferably provided in a 3D form.

[0034] In step (a) of the method for identifying useful compounds of the present invention, a set of compounds is provided. Basically a compound of this set of compounds can be any known compound or hypothetical compound. The hypothetical compound(s) is/are only limited by the possibility of their synthesis by known chemical reactions and/or reaction sequences and known educts for these reactions and/or reaction sequences. As already mentioned above, by now 53,404, 695 known compounds have been registered in CAS. Further, since many of these compounds have been published as examples for general chemical synthesis processes that could be applied to a much broader set of starting materials a multitude of possible (hypothetical, virtual) compounds can be produced. These known and possible compounds (furthermore simply called compounds) form the basis for the set of compounds which can be provided in step (a).

[0035] Preferably, the compounds comprise at least one cyclic scaffold, e.g. at least one aromatic or heteroaromatic ring and/or non-aromatic ring (carbocyclic or heterocyclic). Especially preferably, the compounds have at least one non-aromatic five, six or seven membered ring (carbocyclic or heterocyclic) as scaffolds. In case a ring (aromatic or non-aromatic) is heterocyclic, it is preferred that it contains 1, 2, 3 or 4 heteroatoms selected from O, N and S.

[0036] Preferably the compounds provided in step (a) are products of one or more multicomponent reaction(s) (MCRs).

[0037] Especially preferred are multicomponent reactions providing compounds with a characteristic, three dimensional arrangement(s) of substituents around a scaffold.

[0038] Further preferred are multicomponent reactions yielding one or more non-aromatic five, six or seven membered rings as scaffolds.

[0039] Multicomponent Reactions (MCRs) are convergent reactions, in which three or more starting materials react to form a product, where basically all or most of the atoms contribute to the newly formed product. In an MCR, a product is assembled according to a cascade of elementary chemical reactions. Thus, there is a network of reaction equilibria, which eventually result in an irreversible step yielding the product.

[0040] Multicomponent reactions are e.g. described in: I. Ugi, Pure Appl. Chem., Vol. 73, No. 1, pp. 187-191, 2001; A. Dömling and I. Ugi. Angew. Chem. 112, 3300 (2000); Angew. Chem. Int. Ed. Engl. 39, 3168 (2000); A. Dömling, Chemical Reviews 2006 106 (1), 17-89; C. Kalinski, Molecular Diversity (published online March 2010; http://www.springerlink.com/content/3585832278t0k513) and references cited therein.

[0041] Several hundreds of MCRs are currently known. Of these, especially those MCR products which offer characteristic (e.g. fixed) three dimensional arrangements of substituents around a scaffold are preferred. Thus, all MCRs or MCR based preparative sequences that yield one or more non-aromatic five, six or seven membered rings as scaffolds are especially preferred.

[0042] MCRs may be used to generate a set of compounds in silico. Substituents bound to these scaffolds may e.g. be or resemble amino acid residues and may contain one or more representatives of all general classes of residues (small/big, polar/lipophilic, rigid/flexible, aliphatic/aromatic, presence of H-bond donors/acceptors etc.).

[0043] In optional step (b), a sub-set may be selected from the set of compounds based on one or more specified molecule properties. Preferably in step (b) one compound property for selecting the sub-set is the molecular weight; especially a molecular weight of 300 to 800 Da. Further specified compound properties which may be used for the selection of step (b) are clogP, D&A count, lipophilicity, polar surface area, etc.

[0044] In addition, compounds containing residues which are known to be problematic in pharmaceuticals may be removed from the set of compounds. Examples for such groups are e.g. epoxides, Michael-acceptors, nitro groups, anilines and hydrazines.

[0045] In step (c) a 3D structure of each of the compounds provided and/or selected in step (a) or (b) is generated. This is preferably done by generating all possible isomers (e.g. cis/trans) of the compounds and a representative set of conformers (e.g. 10-100) for each compound. Preferably in step (c) the generation of the 3D structure is carried out by generating a representative ensemble of low energy conformers via molecular modeling. The method preferably utilized for step (c) is based on a modified Genetic Algorithm (GA) allowing for a fast exploration of conformational space and the associated energy defined by a molecular mechanics potential energy function (force field). GAs have proved to be the method of choice when large search spaces like the conformational space of flexible molecules have to be sampled efficiently [Judson, R. *Genetic algorithms* and their use in chemistry. Reviews in Computational Chemistry 1997, 10, 1-73]. The modified GA used to perform step (c) is implemented as a part of a proprietary software package.

[0046] In step (d) the 3D structures of the compounds are encoded. Preferably the encoding of the 3D structures comprises the following steps:

[0047] (d1) taking only non-hydrogen atoms of the compound into account;

[0048] (d2) determining the center of mass of the compound;

[0049] (d3) determining the relative position of each non-hydrogen atom with respect to the center of mass;

[0050] (d4) determining the non-hydrogen atom farthest away from the center of mass and defining a vector s_j pointing from the center of mass to said atom;

[0051] (d5) defining a spatial area SA_j around said vector s_j; preferably the spatial area is a conic spatial area around vector s_j, especially with s_j being the rotational axis and the center of mass being the top of the cone;

[0052] (d6) associating all non-hydrogen atoms falling within said spatial area SA_j with said vector s_j;

[0053] (d7) repeating steps (d4) to (d6) with the remaining non-hydrogen atoms until no further non-hydrogen atoms are left; and

[0054] (d8) assigning all hydrogen atoms to the non-hydrogen atoms of the compound.

[0055] In the following paragraphs the preferred encoding of the 3D structures of the compounds will be described in detail:

[0056] In the context of the present invention the term "DPSM Descriptor" (DPSM=Distorted Polyhedral Super Molecule) relates to the encoded 3D structure.

[0057] A molecular graph is defined by a set of atoms (nodes) and a set of bonds (edges), where N is the number of atoms and N_b is the number of bonds:

$$A=\{a_1, a_2, \ldots a_N\}$$

$$B=\{b_1, b_2, \ldots b_{Nb}\}$$

[0058] The atomic coordinates are given by:

$$\vec{r}_i = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} \quad i = 1 \ldots N$$

[0059] The center of mass of a molecule is:

$$\vec{c} = \begin{pmatrix} x_c \\ y_c \\ z_c \end{pmatrix} = \frac{1}{\sum\limits_{i=1}^{N} m_i} \begin{pmatrix} \sum\limits_{i=1}^{N} m_i x_i \\ \sum\limits_{i=1}^{N} m_i z_i \\ \sum\limits_{i=1}^{N} m_i z_i \end{pmatrix}$$

[0060] For calculating the DPSM representation only non-hydrogen atoms are taken into account.

[0061] In the first step the relative atomic positions $\vec{u}_i$ with respect to the center of mass of the molecule are calculated. This already separates the 3 translational degrees of freedom from the descriptor representation.

$$\vec{u}_i = \begin{pmatrix} x_i - x_c \\ y_i - y_c \\ z_i - z_c \end{pmatrix} \quad i = 1 \ldots N$$

[0062] This results in a set of vectors pointing from the center of mass to the individual atoms:

$$U=\{\vec{u}_1, \vec{u}_2, \ldots, \vec{N}_N\}.$$

[0063] The elementary step consists in reducing this set of vectors to a basic set of so-called "shape vectors" which will point into the principal directions of molecular extent. For this purpose first a part of the molecule consisting of a sub-set of atoms is defined:

$$S_j=\{a_{i(j)}\}\ j=1 \ldots M$$

[0064] There is one atom in each sub-set $S_j=\{a_{i(j)}\}$ which is denoted as "base atom" $a_{i(j)}*$. This is the atom with the greatest distance to the center of mass:

$$|\vec{u}_{i(j)}*|>|\vec{u}_{i(j)}|$$

[0065] The position vector of the base atom defines the shape vector of the molecular part:

$$\vec{s}_j=\vec{u}_{i(j)}* \text{ shape vector}$$

$$\sigma_j=|\vec{s}_j| \text{ length of the shape vector}$$

[0066] The initial shape vector is given by the atom farthest away from the center:

$$\vec{s}_1=\vec{u}_{i(1)}*$$

[0067] This shape vector defines the first principal direction of molecular extent. Then a conic spatial area is specified round $\vec{s}_1$, with $\vec{s}_1$ being the rotational axis and the center of mass building the top of the cone. Each atom $a_{i(1)}$ falling inside this spatial area is then associated with molecular part S_1 and its shape vector $\vec{s}_1$ respectively (FIG. **1**).

$$S_1=\{a_{i(1)}\}, \vec{s}_1$$

[0068] With the remaining set of atoms $A_2=A_1-S_1$ the procedure is repeated. Again the atom farthest away from the center now defines the second shape vector $\vec{s}_2$ and all atoms falling inside the conic area are associated with S_2:

$$S_2=\{a_{i(2)}\}, \vec{s}_2$$

[0069] The procedure works in a recursive manner ($A_{i+1}=A_i-S_i$) and an example for a corresponding algorithm is shown below. It terminates, when there are no further atoms left to process, i.e. all atoms are associated with a shape vector. Computational experiments have shown that for drug-like molecules the number of shape vectors typically ranges between 4 and 8.

[0070] In this way molecular geometry is described as a kind of "super-molecule" consisting of a "central atom" (center of mass) and a set of "super-substituents" represented by the shape vectors and the atoms associated with them. The shape vectors point into the principal directions of spatial extent of the molecule and they typically describe a distorted polyhedral coordination sphere round the "central atom". (FIG. **2**: Distorted octahedral orientation of "super-substituents")

[0071] One of the most important requirements a molecular descriptor should satisfy is the independency of its numerical representation from the size (number of atoms) of the molecule, because only then it is possible to compare different molecules based on their descriptor representation. To finally introduce uniqueness and rotational invariance, a static vector representation of the super-molecule is calculated:

[0072] The first vector coordinate stores the number of atoms:

$$g_1=N$$

[0073] The second coordinate is given by the number of shape vectors:

$$g_2=M$$

[0074] The third coordinate is defined by the sum of the lengths of all shape vectors:

$$g_3=\sum_{k=1}^{M}|\vec{s_k}|$$

[0075] The higher vector components are all calculated as triples of 3 statistical measures, i.e. the mean, the variance and the skewness of selected geometric or physical properties p_k of the super-substituents:

$$g_3=\overline{p}=\frac{1}{M}\sum_{k=1}^{M}p_k \qquad \text{Sample mean}$$

$$g_4=\sigma=\frac{1}{M}\sum_{k=1}^{M}(p_k-\overline{p})^2 \qquad \text{Sample variance}$$

$$g_5=\frac{\frac{1}{M}\sum_{k=1}^{M}(p_k-\overline{p})^3}{\sigma^{3/2}} \qquad \text{Sample skewness}$$

[0076] When the lengths of the shape vectors are used as a geometric property, the mean gives a measure of the general size of the super-molecule, the variance describes how strong the super-substituents differ in spatial extent and the skewness characterizes the symmetry of the super-molecule.

[0077] However, when a physical property like the number of π-electrons associated with a super-substituent is used, these statistical measures provide information how this properties are distributed over the principal directions in space.

[0078] In one approach the lengths of the shape vectors, the approximate van der Waals volumes, the number of π-electrons and the number of branches associated with the super-substituents are used. Finally this leads to a vector representation of the molecule with a constant vector dimension (e.g. 15 in this case).

$$\overrightarrow{\text{DPSM}}=(g_1, g_2, g_3, \ldots g_{3(P+1)})\ \dim(\overrightarrow{\text{DPSM}})=3(P+1)$$

[0079] P is the number of geometric and physical properties included.

[0080] When constructing the shape vectors additionally the condition that a super-substituent must represent a connected molecular graph (substructure) is imposed, i.e. each atom must be connected to at least one other atom of the super-substituent. This enables to include arbitrary physicochemical properties that can be calculated for a conventional molecule, e.g. logP, number of H-bond donors/acceptors, Van der Waals Volume, Van der Waals Surface, Solvent Accessible Surface (SAS) to mention only a few.

[0081] To additionally include a specific measure of folding or puckering of the molecular shape also the ratio between solvent accessible surface and molecular volume should preferably be included:

$$p_j=\frac{SAS_j}{(V_{vdw})_j}$$

[0082] Pseudo code of the algorithm to construct a DPSM representation of a molecule:

```
0 -  take into account only non-hydrogen atoms A
1 -  given the set of atoms A, search for the atom with greatest distance
     from the center
2 -  make this atom the base atom $a_{i(j)}$*of $S_j$
3 -  define a spatial area $SA_j$ around the shape vector $\vec{s}_j$ of $S_j$
4 -  associate with $S_j$ all those atoms $a_{i(j)}$ falling inside this spatial area
5 -  with the remaining set of atoms $A_{j+1} = A_j - S_j$ repeat the
     procedure starting at step 1 again
6 -  GOTO 1 UNTIL all atoms are processed
```

[0083] Preferably the active site of a target molecule is encoded by a method comprising the following steps:

[0084] (s1) taking only non-hydrogen atoms of the target molecule into account;

[0085] (s2) defining the center of the active site;

[0086] (s3) defining a sphere of radius R_C around this center;

[0087] (s4) determining all non-hydrogen atoms falling inside the sphere defined in (s3);

[0088] (s5) calculating the distance vector u_j between each atom determined in (s4) and the center of the active site;

[0089] (s6) defining a spatial area SU_j around each vector u_j;

[0090] (s8) calculating the reduction of volume of SU_j caused by intersecting atom spheres;

[0091] (s9) repeating steps (s5) to (s8) until no further non-hydrogen atoms are left;

[0092] (s10) creating a ranking of all u_j based on their effective volume; and

[0093] (s11) using the N best u_j as shape vectors for a comparison with the encoded 3D structures in step (g).

[0094] The basic idea behind the description of an active site of a target molecule via the DPSM concept is to encode geometric characteristics that are complementary to the DPSM representation of a compound. Whereas for the latter the basic algorithm shown above searches for principal directions of molecular extension in space, the procedure applied to an active site searches for principal directions of "ligand accessible space". These regions of "empty space" may indicate the existence of pockets inside the active side which can be occupied by the corresponding residues of a potential ligand. Like in the case of the molecular DPSM descriptor the goal is to construct a set of shape vectors that define these principal directions in space. The basic strategy is again to define a center or reference point in space, but then to systematically "scan" the space around this center for regions of "empty space".

[0095] Because the DPSM descriptor of the active site of a specific target is preferably calculated only once, there is no basic requirement to make the procedure incredible fast. From this point of view the construction of a set of DPSM shape vectors may also be carried out manually, e.g. by visually analysing the molecular surface of an active site via a molecular modeling tool and then to define a set of vectors pointing from the reference point into directions where the binding pockets are assumed to be located.

[0096] Nevertheless this simple approach is sometimes not adequate enough, especially if it comes to tasks like providing a whole set of possible DPSM descriptors for one active site, or if some DPSM descriptor based statistics of the active sites of target proteins stored in a database like the PDB (Protein Data Bank) must be performed.

[0097] A precondition for the algorithm presented below, is that at least an approximative location of the active binding site can be defined.

[0098] In the first step the center $\vec{c}$ of the active site is defined. All atoms of the target protein falling inside a sphere of radius R_C around the center are assumed to represent the set of atoms A interacting with a potential ligand:

$$|\vec{r}_i - \vec{c}| < R_c \Rightarrow {}_{ai} \in A$$

$$A = \{a_l, a_j, \ldots a_N\} \; j = 1 \ldots N$$

Then, as in the case of the molecular DPSM, the relative atomic positions with respect to the center are calculated:

$$\vec{u_j} = \begin{pmatrix} x_j - x_c \\ y_j - y_c \\ z_j - z_c \end{pmatrix} \; j = 1 \ldots N$$

$$U = \{\vec{u_1}, \vec{u_2}, \ldots \vec{u_N}\}$$

[0099] Again U is reduced to a basic set of shape vectors. First a spatial area SA_j is defined around each vector $\vec{u}_j$. For reasons of mathematical simplicity a cylinder is prototypically used, whereas the vectors $\vec{u}_j$ define the rotational axis. The volume of a cylinder is:

$$V_j = SA_j = r_j^2 \pi \cdot |\vec{u}_j|$$

$$r_j \approx 2.0$$

[0100] The volume will be reduced as soon as there is an atomic sphere a_k intersecting with the cylinder. Because an intersecting atom a_k (note: this is given by $\vec{u}_k$) represents a "spatial barricade" along direction $\vec{u}_j$, the length of $\vec{u}_j$ is simply reduced instead of calculating the reduction of volume explicitly:

$$\vec{u}_j^{\#} = \frac{1}{|\vec{u}_j|} \cdot \vec{u}_j \cdot |\vec{u}_k| \cdot \cos[\angle(\vec{u}_j, \vec{u}_k)]$$

[0101] So the shortened vector $\vec{u}_j^{\#}$ points into the same direction as $\vec{u}_j$ does, but it indicates, that there is "empty space" along this direction until $|\vec{u}_j^{\#}|$ is reached, where the atomic barrier is "located". The reduced volume is therefore:

$$V_j^{\#} = r_j^2 \pi \cdot |\vec{u}_j^{\#}|$$

[0102] Processing in this way all vectors $\vec{u}_j$ against all intersecting atoms a_k and finally sorting the $\vec{u}_j^{\#}$ according to their length (or volume) results in a set:

$$U^{\#} = \{\overrightarrow{u_1^{\#}}, \overrightarrow{u_2^{\#}}, \ldots, \overrightarrow{u_N^{\#}}\}$$

[0103] From this set, the first M vectors are selected to define the shape vectors $\vec{s}_j$ of the DPSM descriptor.

$$S = \{\vec{s}_1, \vec{s}_2 \ldots \vec{s}_M\} = \{\overrightarrow{u_1^{\#}}, \overrightarrow{u_2^{\#}}, \ldots \overrightarrow{u_M^{\#}}\}$$

[0104] Pseudo code of the algorithm to construct a DPSM representation of an active site:

```
 0 - only take into account non-hydrogen atoms
 1 - define the center C of the acive site
 2 - define a sphere of radius $R_C$ around this center
 3 - define all atoms falling inside this sphere the set of active site atoms A
 4 - FOREACH atom do
     5 - calculate the distance vector $\vec{u}_j = \vec{r}_j - \vec{c}$ between the atom and the center C
     6 - define a spatial area $SA_j$ around $\vec{u}_j$
     7 - calculate the reduction of volume of $SA_j$ caused by intersecting atom spheres
 8 - END FOREACH atom
 9 - create a ranking of all $\vec{u}_j$ based on their effective volume
10 - use the M best $\vec{u}_j$ as the shape vectors $\vec{s}_j$ of the DPSM descriptor
```

[0105] In this way a shape vector $\vec{s}_j$ of an active site DPSM descriptor represents a region of ligand accessible space (LAS) and may indicate a pocket that can be occupied by a super-substituent of a DPSM of a ligand molecule.

[0106] FIG. **3** shows regions of ligand accessible space calculated for the active site of mdm2.

[0107] FIG. **4** shows shape vectors of the DPSM descriptor of the active site of mdm2.

[0108] Like in the case of the DPSM descriptor for compounds, several atoms of the active site can be associated with a shape vector $\vec{s}_j$ and so the physico-chemical properties of the pocket can be included into the descriptor.

[0109] To rationally handle Protein-protein-interactions (PPI) the knowledge about the binding site of a target protein is indispensable. Since this knowledge is not always available from scratch, the following fast and robust computational method has been developed that is able to identify potential binding sites as soon as 3D structural information of a target protein is available.

[0110] This method starts from the hypothesis that a binding site builds a kind of cavity that is more or less embedded into the molecular surface of a protein. Such a cavity is characterized by two major spatial regions. First there is an outer shell occupied by atoms of the target protein that constitute the molecular surface inside the cavity. Second there is an inner region that provides enough space for a ligand molecule to “reside” inside the cavity. This lead to a simple model of a cavity and a method for calculating a probability score for a definite protein region to be an active site.

[0111] The basic strategy of the method for identifying the binding site of a target molecule is to systematically scan the space occupied by a target protein for potential cavities that may be more or less embedded in the molecular surface. Starting from the simple picture, that a perfectly embedded cave can be abstracted as a “closed” sphere, the less the cave is embedded the more “open” the sphere will be.

[0112] In the present method a cavity is described by two concentric spheres, where the inner sphere provides space for a potential ligand and the region between the inner and the outer sphere defines an area where the “surrounding” atoms of the active site are located. An inner radius r_{inner} of 4-6 angströms is used to approximate the size of a virtual ligand molecule. The radius of the outer sphere r_{outer} is calculated by adding to the inner radius 3 times the van der Waals radius of a carbon atom $r_{outer}=r_{inner}+3r_{vdW}(C)$ which result in about 11-13 angströms.

[0113] The algorithm to predict potential active sites is a “brute force” systematic search. First a cuboid enclosing the protein is defined. Within this cuboid a cartesian grid with a distance of grid points $\Delta x \approx 1.0$ angstroms is created. Each grid point defines the center of a probe cavity. For each probe cavity the number of atoms falling inside the inner sphere N_{inner} and the number of atoms falling inside the region between the inner and the outer sphere N_{outer} of the probe cavity is determined. The score is calculated by:

$$s = \frac{N_{outer}}{1 + N_{inner}}$$

[0114] In this way the highest scores are produced by probe cavities that are well embedded into the protein ($N_{outer} >> 0$) but that miss atoms inside the inner ligand sphere $N_{inner} \to 0$.

[0115] Pseudo code of the corresponding algorithm:

```
only take into account non-hydrogen atoms
define a cuboid enclosing the target protein
inside the cuboid generate a regular cartesian grid with a point distance of $\Delta x \approx 1.0$
FOREACH grid point DO
    define 2 concentric spheres around the point with $R_{Inner} \approx 5.0$ and $R_{Outer} \approx 12.0$
    calculate number of atoms $N_{Inner}$ falling inside the inner sphere
    calculate number of atoms $N_{outer}$ falling inside the shell between the inner and outer sphere
    calculate a fitness-score according $s = N_{Outer}/1 + N_{Inner}$
    store this core if it occupies a rank within the M highest scores found so far
END FOR EACH
the M highest scores provide a list of potential cavities or active sites
```

[0116] The images of FIGS. **5** and **6** show the co-crystal structures of mdm2/nutlin-3 and c-met/su1127 with the active sites predicted by this algorithm:

[0117] FIG. **5**: Active site of mdm2 predicted by this algorithm

[0118] FIG. **6**: Active site of c-met predicted by this algorithm

[0119] The similarity or distance of two molecular DPSM descriptors can simply be calculated on the basis of well-known metrics like the Euclidean or the Manhattan distance:

$$d(\overrightarrow{DPSM}_A, \overrightarrow{DPSM}_B) = \frac{1}{L}\sqrt{\sum_{j=1}^{L}(g_{j,A} - g_{j,B})^2} \quad \text{Euclidean distance}$$

$$d(\overrightarrow{DPSM}_A, \overrightarrow{DPSM}_B) = \frac{1}{L}\sum_{j=1}^{L}(g_{j,A} - g_{j,B}) \quad \text{Manhattan distance}$$

$$L = dim(\overrightarrow{DPSM}_A) = dim(\overrightarrow{DPSM}_B)$$

[0120] As soon as structural information about a validated ligand or the active side of a target (or both) is available, it can be encoded via the corresponding DPSM descriptor and a 3D database of potential peptido-mimetics can be searched. The result of a similarity search is always a ranking based on the calculated distance measure and it provides a set of compounds that can further be processed in docking simulations and finally may lead to promising candidates for synthesis in the laboratory.

[0121] Preferably, in step (h) the similarity range is defined. Because of the difficulties in normalizing the concrete numeric values of the calculated distances, the similarity range is not defined explicitly. Instead of this a maximal number of ranks is used to limit the number of results of the similarity search.

[0122] Preferably, in optional step (i) the sub-set is selected on the basis of results of in silico docking. For the latter a GA based method is used, which is also implemented as a part of a proprietary software package. For each potential ligand molecule a small set of energetic minima of intermolecular ligand-target-interaction is searched. The energy of intermolecular interaction is assumed to provide an approximative measure of ligand-target-complementarity, i.e. a low energy conformation of a ligand molecule is assumed to define a possible binding mode.

[0123] MCRs provide an excellent spectrum of chemical scaffolds that can mimic the interacting amino acid residues of a native PPI ligand, because many of them constitute of a conformationally restrained central unit C (usually a small ring system) and a set of highly variable residues R1, R2 . . . R4 extending into different directions of space.

[0124] The method of the present invention can be applied in drug discovery projects under different starting conditions: If only ligands are known (scaffold hopping), for de novo generation of small molecule modulators starting from target information only, or ideally based on a combination of both.

[0125] Protein-protein-interactions (PPIs) are highly attractive targets for a variety of indications and could become successors of kinases as prime targets for a whole era. The method of the present invention is particularly suited for addressing PPIs, due to the following considerations:

> **[0126]** PPIs employ binding motifs that contain three to four amino acids. An example is Mdm2, where Phe-Trp-Leu is known as binding triad (e.g. P. Chene, Molecular Cancer Research, Vol. 2, 20-28, Jan. 2004; S. Shangary, PNAS, Mar. 11 2008, Vol. 105, no. 10, 3933-3938). In nature there are 22 proteinogenic amino acids. This means that the number of possible sequence variations is limited: 22*22*22=10648.

[0127] The accessible diversity of binding motifs is multiplied by a rich number of conformations these sequences can take in proteins due to secondary and tertiary structures.

[0128] Multicomponent reactions (MCRs) are perfectly suited for the easy and straightforward assembly of three to four highly variable rests. A large number of MCRs deliver scaffolds that could be regarded as "peptide similar" in terms of spatial arrangement of substituents.

[0129] Literature PPIs as upcoming attractive target class: O. Sperandio, Drug Discovery Today, Volume 15, Numbers 5/6, March 2010; J. Fuller, Drug Discovery Today, Volume 14, Numbers 3/4, February 2009; J. Wells, NATURE, Vol 450, 13 Dec. 2007;

[0130] The term "useful" or "useful compounds" relates to compounds having desired properties. Preferably the compounds show a specific desired biological activity (e.g. the compounds may act as enzyme inhibitors). Especially preferably the compounds modulate (e.g. inhibit) protein-protein interactions.

[0131] According to a preferred embodiment, the present invention relates to a method for identifying compounds having a desired biological activity.

[0132] According to an especially preferred embodiment, the present invention relates to a method for identifying compounds that modulate (e.g. inhibit) protein-protein interactions.

[0133] The method of the present invention is preferably carried out in silico on a computing machine, e.g. on a computer. The results may e.g. be displayed on a display device (e.g. a monitor). Data may be fed to the computing machine by means of a keyboard and/or by means of a storage device, e.g. a harddisk.

[0134] Especially preferably, the method of the present invention is computer-implemented.

[0135] The method of the present invention especially provides the following advantages:

[0136] 1. It provides for a drug discovery engine merging together several new concepts to approach the challenging field of identifying small ligand molecules e.g. for targets involved in protein-protein interactions (PPI). Current drug discovery engines are usually not capable in this area.

[0137] 2. It uses novel molecular 3D descriptors emphasizing the principal directions of molecular extent in space. Current 3D descriptors mostly rely on viewing a molecule as a set of points in space and take into account interatomic distances only (J. Chem. Info. Comp. Sci. (1995), 35, 373-382).

[0138] 3. It uses a novel active site 3D descriptor emphasizing the principal directions of space accessible for ligands. Current 3D descriptors mostly rely on a negative print of the active site and are computationally expensive to calculate.

EXAMPLES

[0139] Ligand Based Similarity Search Using the DPSM Descriptor

[0140] To perform a validation of the DPSM descriptor implementation, first an appropriate search set of drug like molecules has been constructed.

[0141] The following sub-sets of three well known 3D structure databases were selected:

ChemBank sub-set	2,344 entries
ChemPDB sub-set	4,009 entries
Drug-likeness NCl sub-set	192,323 entries

[0142] (The corresponding SDF-files are available at: http://ligand.info/).

[0143] These 3 sub-sets were merged into a single 3D structure database. Then the following filters were applied:

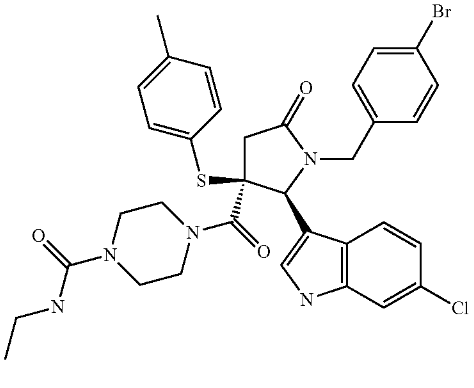
[0144] only use molecules consisting of atoms in the "organic sub-set", i.e.

[0145] atom Type ∈ {H, B, C, N, O, F, P, S, Cl, Br, I}

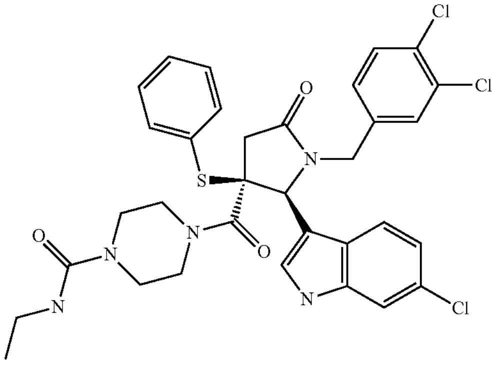
[0146] only use molecules with a molecular weight m>100

[0147] This resulted in a set of 188128 compounds. To the latter the 3D structures of four validated mdm2-inhibitors have been added:

PXN_727 (Priaxon)

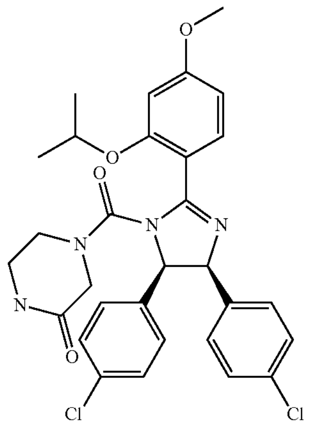


PXN_822 (Priaxon)

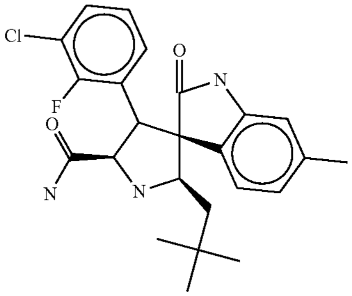


-continued

Nutlin-3



Mi-63



[0148] The final search set then encompasses 188133 molecular 3D structures. For searching this set, the following parameters were used:

[0149] a) Reference Structure: PXN_727

[0150] b) Distance Metric: Manhattan Distance

[0151] c) Similarity-Descriptor: DPSM

[0152] d) DPSM-Parameters:

$g_1 = N$	Number of atoms
$g_2 = M$	Number of shape vectors
$g_3 = \sum_j \sigma_j$	Sum of lengths of shape vectors
g_4, g_5, g_6	Lengths of shape vectors
g_7, g_8, g_9	Widths of shape vectors*
g_{10}, g_{11}, g_{12}	Atomic van der Waals volumes
g_{13}, g_{14}, g_{15}	Number of π-electrons
g_{16}, g_{17}, g_{18}	Number of branches
g_{19}, g_{20}, g_{21}	Number of halogens
g_{22}, g_{23}, g_{24}	Number of chalcogens
g_{25}, g_{26}, g_{27}	Number of nitrogens

[0153] * the "width" of a shape vector is calculated as the mean distance of the associated atoms from the line defined by the base vector

[0154] The 3D similarity search was carried out on a HP Intel 15 Quad Core machine. Because the current implementation did not support parallelism, only one of the CPU cores was used, which corresponds to only 25% of the overall CPU-power. Nevertheless searching the set of 188133 molecular structures needed only 16 seconds of computation time! This already demonstrates that DPSM is a quite fast computational method.

[0155] On the basis of the results obtained from other runs (with varying the DPSM parameters) the following conclusions could be drawn:

[0156] using geometric parameters only (lengths, volumes) usually leads to poor results

[0157] including chemical and topological parameters (π-electrons, branches etc.) dramatically improves the similarity rankings

[0158] including primary "1D filters" (number of atoms, sum of lengths of shape vectors) performs out molecules that show a similar distribution of properties in space but differ significantly in size from the query molecule

[0159] In the similarity search, all three other validated mdm2-inhibitors (PXN_822, Nultin-3, Mi-63) are ranked within the first twenty molecules of highest similarity to reference structure PXN_727.

[0160] PXN_822 is most similar to PXN_727 as can be seen easily from the formula. The structure of Nutlin-3 shows quite the same orientation of chemically similar substituents. The similarity between PXN_727 and Mi-63 is not as obvious as for Nutlin-3 at the first glance, but this is one of the advantages of the DPSM descriptor—it does not take into account only geometric features like e.g. the USR molecular shape descriptor [P. J. Ballaster, W. G. Richards, Proc. R. Soc. (2007), 463, 1307-1321], but also includes physico-chemical properties, that must by nature be similar for the same class of inhibitor molecules.

1-9. (canceled)

10. A method for identifying compounds comprising the steps of:

(a) providing a set of compounds;

(b) optionally selecting a sub-set from the set of compounds based on one or more specific compound properties;

(c) generating a 3D structure of each of the compounds provided in step (a) or optionally selected in step (b);

(d) encoding each 3D structure;

(e) providing at least one known compound having at least one desired property or providing a target molecule;

(f) encoding a 3D structure of each known compound provided in step (e) or an active site of the target molecule provided in step (e);

(g) comparing each encoded 3D structure of step (d) with each encoded 3D structure of step (f); and

(h) selecting all compounds falling within a specified similarity range.

11. The method of claim **10**, further comprising the steps of:

(i) optionally selecting a further sub-set of the compounds provided in step (h) based on one or more specific compound properties;

(j) preparing the selected compounds of step (h) or optionally selected compounds of step (i) and testing the prepared compounds for activity;

(k) optionally repeating steps (g) to (j) or (h) to (j).

12. The method of claim **10**, wherein the compounds provided in step (a) are products of one or more multicomponent reactions.

13. The method of claim **12**, wherein the one or more multicomponent reactions provide one or more products with a characteristic, three dimensional arrangement of substituents around a scaffold.

14. The method of claim **12**, wherein the one or more multicomponent reactions yield a non-aromatic five, six or seven membered ring as scaffold.

15. The method of claim **10**, wherein in step (b) the specific compound property for selecting the sub-set is a molecular weight of 300 to 800 Da.

16. The method of claim **10**, wherein in step (c) the generation of the 3D structure is carried out by generating a representative ensemble of low energy conformers via molecular modeling.

17. The method of claim **10**, wherein encoding of the 3D structures in step (d) comprises the steps of:

(i) taking only non-hydrogen atoms of the compound into account;

(ii) determining a center of mass of the compound;

(iii) determining a relative position of each non-hydrogen atom with respect to the center of mass;

(iv) determining the non-hydrogen atom farthest away from the center of mass and defining a vector S_j pointing from the center of mass to said non-hydrogen atom;

(v) defining a spatial area SA_j around said vector S_j;

(vi) associating all non-hydrogen atoms falling within said spatial area SA_j with said vector S_j;

(vii) repeating steps (iv) to (vi) with the remaining non-hydrogen atoms until no further non-hydrogen atoms are left; and

(viii) assigning all hydrogen atoms to the non-hydrogen atoms of the compound.

18. The method of claim **10**, wherein the active site of the target molecule in step (f) is encoded by a method comprising the steps of:

(i) taking only non-hydrogen atoms of the target molecule into account;

(ii) defining a center of the active site;

(iii) defining a sphere of radius R_c around the center of the active site;

(iv) determining all non-hydrogen atoms falling inside the sphere defined in (iii);

(v) calculating a distance vector u_j between each atom determined in (iv) and the center of the active site;

(vi) defining a spatial area SU_j around each vector u_j;

(vii) calculating a reduction of volume of SU_j caused by intersecting atom spheres;

(viii) repeating steps (v) to (vii) until no further non-hydrogen atoms are left;

(ix) creating a ranking of all u_j based on an effective volume; and

(x) using an N best u_j as shape vectors for a comparison with the encoded 3D structures in step (d).

* * * * *