

CHEMICAL DATABASE TECHNIQUES IN DRUG DISCOVERY

Mitchell A. Miller

Chemical databases are becoming a powerful tool in drug discovery. Database searches based on possible requirements for biological activity can identify compounds that might be suitable for further analysis or indicate novel ways to achieve the desired activity. What considerations are involved in the construction and searching of chemical databases?

Chemical databases have progressed over the past 15 years from being a mere repository of the compounds synthesized within an organization, to being a powerful research tool for discovering new lead compounds. By using a query that encapsulates some idea of the requirements for biological activity, previously untested compounds that might have the same type of activity can be identified. In many cases, a compound that is found in a search will not act as the new drug itself — compounds in databases tend to be known to the world and therefore unsuitable for patenting — but it could indicate novel ways to provide a desired activity. The chemist will then create molecules that are close to the database 'hit'. In this way, the database serves as an idea generator.

The growth of chemical databases in drug discovery research has fuelled a growth in commercial software for chemical database management (see ONLINE TABLE 1), and in publicly and commercially available chemical databases¹ (see online links and ONLINE TABLE 2). This review examines the chemical database as a research tool in the current drug discovery environment. We look at the types of database that are in use, how they are generated, what data they contain, how queries are formulated and run, and finally, how the results of searches are processed.

Types of chemical database

In many disciplines, the most commonly used terms are the ones that lead to the most confusion. 'Chemical structure' is one such pervasive yet vague term. To a chemist concerned with synthesizing new compounds, 'chemical structure' would usually mean a two-dimensional sketch

of atoms and bonds that represents the compound produced (FIG. 1a). Implicit in this type of diagram is a set of x,y coordinates for all of the atoms in the structure, allowing them to be seen in a way that allows another chemist to understand the identity of the compound immediately.

However, chemical-structure diagrams are not amenable to computational operations such as database searching, so several types of chemical-structure representation have been developed by theoretical chemists for use in computer systems. The predominant form is the atom–bond connection table. Formally, a connection table records the chemical structure as a graph^{2,3} — a set of vertices (the atoms) linked by edges (the bonds) — which allows mathematical analyses to be applied to classify the structure or calculate its molecular properties^{4,5}. At a basic level, a connection table represents a chemical structure by listing the atoms and bonds that are present in a tabular form (FIG. 1b). Line notations for chemical structures (FIG. 1c), such as SMILES (simplified molecular input line system)⁶, are also commonly used. SMILES representations contain the same information as that which might be found in an extended connection table, and are compact and easy to use.

Another chemist looking at the same compound as a potential drug for a given receptor would look at the overall shape and surface characteristics of the three-dimensional molecule — individual atoms and bonds would be less important. Generally, at least the atoms are present, along with 3D coordinates that provide a full spatial depiction of the molecule, such as that shown in FIG. 1d. More than one 3D structure can be generated for many drugs and DRUG-LIKE molecules, as

LION bioscience,
9880 Campus Point Drive,
San Diego, California 92121,
USA.
e-mail: mitchell.miller@
lionbioscience.com
DOI: 10.1038/nrd745

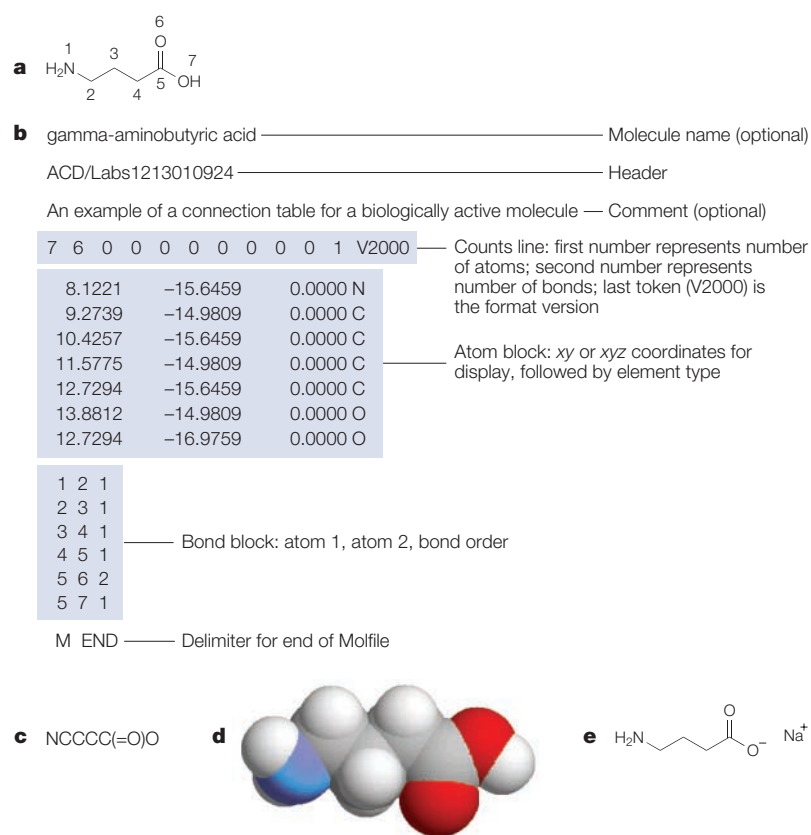


Figure 1 | **Representations of chemical structure, using the neurotransmitter γ -aminobutyric acid (GABA) as an example.** **a** | Two-dimensional chemical-structure diagram. The numbering is that used in the counts line in panel **b**, which shows a simple connection table in the Molfile format, with selected fields highlighted. Some zero-filled fields have been removed for clarity. **c** | SMILES (simplified molecular input line system) representation. **d** | Space-filling model. **e** | A curator's view.

DRUG LIKE

Sharing certain characteristics with other molecules that act as drugs. The set of characteristics — size, shape and solubility in water and organic solvents — varies depending on who is evaluating the molecules.

CYCLIC/ACYCLIC BONDS

If chemical bonds occur in a ring, they are termed 'cyclic'. 'Acyclic bonds' occur in open chain structures.

COUNTERION

A set of one or more bonded atoms, with opposite charge and generally smaller size, that accompanies another charged set of bonded atoms as dictated by the principle of electrical neutrality of substances, solutions and so on.

COMBINATORIAL CHEMISTRY

The generation of large collections, or 'libraries', of compounds by synthesizing all possible combinations of a set of smaller chemical structures.

many such molecules contain *ACYCLIC* single bonds that join two groups of atoms. Such bonds can usually rotate with a low-to-moderate energy barrier that changes the orientation of other groups in the structure and therefore how the structure can interact with other molecules (BOX 1).

To the curator of a corporate compound collection, the 'structure' might include both the connection table and the displayable sketch, along with *COUNTERIONS*, solvent molecules, and so on, in order to completely define the substance (FIG. 1e). A 'structure' could also be viewed as being not just one compound, but an entire set or 'library' that is formed as the result of a *COMBINATORIAL CHEMISTRY* programme^{7–9}. Other definitions of structure, such as the searchable scheme of a chemical reaction, or those found in polymer, formulation and zeolite chemistry, are beyond the scope of this review¹⁰.

The different definitions of 'structure' are important when considering the function of a chemical-structure database, but broadly, chemical database systems can be classified into two categories: those that define the structure on the basis of connectivity (atoms and bonds) with or without 2D coordinates for the atoms, and those that store 3D coordinates for atoms, in addition to bonds. Most commercial packages support both types.

Database building

The process of constructing/populating chemical databases is complicated. Each organization has its own rules, conventions and procedures; however, there are several shared basic concepts, which are discussed below.

Database structure. A few conventions are implicit in the way in which the database is set up. One consideration is which chemical entities are to be tracked. The three basic entities are: compound (a generalization of the chemical structure, which generally corresponds to the synthetic chemist's view that was discussed above); form (the compound plus counterion(s), solvent(s) and isotopically labelled atoms, as in the curator's view); and lot (a sample of the compound, which is often termed 'batch'). In some organizations, the form is suppressed as a distinct entity, but the defining attributes of counterion and solvent are included in the lot. In other organizations, counterions and solvent are explicitly included in the structure of compounds. This merely reflects different ways of organizing the same set of information.

Structure input. The first step in registering a new compound is generally to recognize (or assume) that a new substance has been obtained (as the result of an in-house synthesis or purchase) or detected in the scientific literature. In most cases, the chemist who performed the synthesis has a good idea of the structure, although this idea can be revised in the future as more information becomes available. The chemist sketches the putative structure in a drawing program, such as ISIS/Draw, ChemDraw or ChemSketch, and transfers the sketch to a custom application. A set of basic verifications, including checks for overlapping atoms, multiple fragments, mislabelled groups, and so on, is typically performed, and the basic properties of molecular weight and formula are calculated and added to the display.

Most organizations have a policy that a parent-compound structure is registered only once to the database. So, a uniqueness check is performed when new structures are added to the database. The definition of uniqueness varies between organizations. In some, multiple salts of the same parent compound are considered as unique; in others, as duplicates.

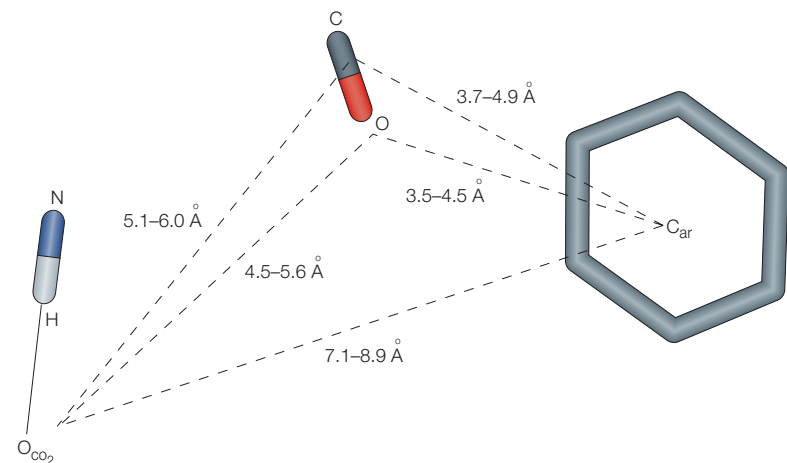
Some organizations allow a user to override a uniqueness check and register a structure that seems to be a duplicate of something in the database (it is possible that the limitations of the software with respect to molecular representation make two different structures appear the same). The structure might also be checked for possible violations of organizational policy, such as multiple fragments (which can arise from a drawing error, or a mistaken attempt to represent salts with explicit counterions, and so on), non-standard orientation of common groups, or charged atoms if these are forbidden.

It should be noted that a given chemical-structure diagram could have many valid computational representations; for example, a different connection table for GABA (γ -aminobutyric acid; FIG. 1a,b) could be

Box 1 | Conformations and activity

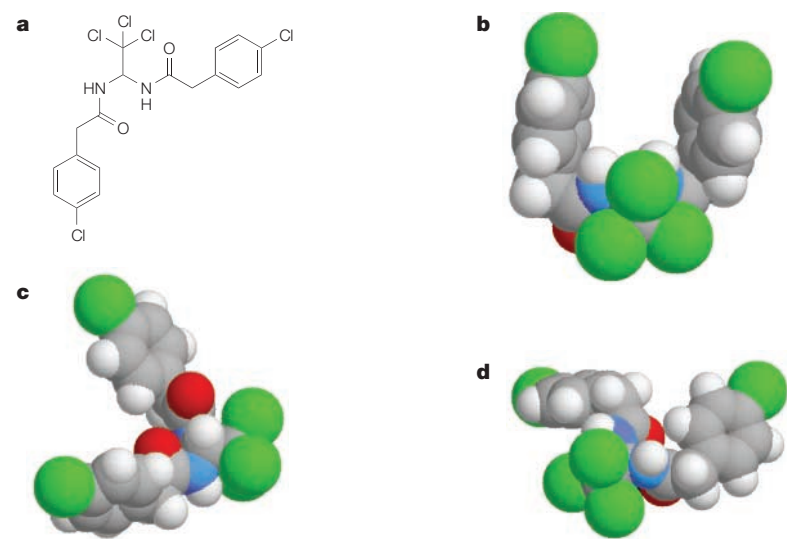
A drug-like molecule can typically adopt many conformations, primarily by twisting around single bonds that are not part of a ring. Most three-dimensional chemical databases store only a limited number of conformations per compound, often just one. Nevertheless, it is of great interest to chemists to pose a 3D question in terms of the entire conformational space of a molecule.

Can any molecule in this database access a conformation that will bring PHARMACOPHORIC groups into the correct distance range for a hypothesis? It is instructive to examine one case. A recent publication by Lavecchia and colleagues⁵⁶ presented a pharmacophore hypothesis for agonists of the κ -opioid receptor. One form of that pharmacophore is shown below:



C_{ar} represents the centre of an aromatic ring; O_{CO_2} refers to a point placed 2 Å from the ammonium hydrogen along the N–H vector to simulate a carboxylate oxygen of aspartate 138; other symbols refer to atoms in the standard manner.

Using the free, publicly accessible National Cancer Institute (NCI) Database (see link to NCI Database), one can pose a query of this type over a set of 250,251 compounds, with ~25 conformations for each generated by using Catalyst (Accelrys). One of the hits, NSC213671, can be examined in detail. The two-dimensional structure is shown in part a of the figure below. Assessment of the possible conformations revealed some, such as those shown in parts b and c, that do not match the query, and one, shown in part d, that does.



This shows that, in a compound with some flexibility, some of the conformations can meet the requirements for a pharmacophore of interest, indicating that the compound might possess the desired activity.

constructed by numbering the atoms differently. It is important for comparison purposes that a unique computational representation is created, and this can be achieved using various algorithms — a process known as canonicalization. So, once the user's input has been verified, the software will generate a terse, canonical representation of the structure (such as a unique SMILES⁶ or a SEMA¹¹ name). The canonical representation is then indexed or hashed to allow rapid exact-match searching, and analysed for designated features that will facilitate rapid SUBSTRUCTURE searching (see below).

3D-coordinate generation. A connection table along with 2D coordinates for display is generally sufficient for the identification of a substance. However, to perform any energy calculations or to determine whether the compound has the potential to bind to a receptor or enzyme of interest, 3D coordinates are necessary. The subject of 3D-structure generation is large and beyond the scope of this review; there are many ways to explore the sets of 3D structures that a given compound can adopt. When building a large database of 3D structures, one needs methods that are fast, albeit approximate. Some of the methods that are in use today on a large scale are CORINA¹², CONCORD^{13,14} and Converter¹⁵. Most drug-like molecules can adopt several 3D CONFORMATIONS by rotation around one or more single, acyclic bonds (BOX 1), and 3D search systems can take this factor into account in the database-building process (by incorporating multiple conformations per compound) or at search time (by fitting candidate molecules from the database onto the query), or both^{16,17}.

Combinatorial libraries. Most of the techniques of chemical database searching that are discussed in this review apply to databases of discrete compounds; that is, structures that represent a single, complete compound, which is generally synthesized and isolated as a pure sample. However, much chemical work in recent years has proceeded through the creation of large numbers of compounds by automated synthesis^{7,8}. The requirements for handling combinatorial libraries are much more complicated than those for discrete structures. The user must be able to input the entire library by creating a set of 'rules,' in the form of a MARKUSH STRUCTURE¹⁸ or a reaction diagram, along with a set of reagents or building blocks that obey the rules. Furthermore, once the library has been created, the system must be able to generate the individual structures on demand. This process, which is termed 'enumeration,' can be performed for the entire library or for structures that meet user-specified criteria⁸. It must be possible to specify a discrete structure and use it to find substructure (or other) matches within the library, without performing an explicit enumeration⁸.

Data cartridges. Recently, the development of chemical data cartridges has enhanced the capabilities of chemical databases. Essentially, data cartridges provide the ability to create custom data types and perform searches on these data types within relational databases, such as

Box 2 | **Rapid pre-screening**

Substructure searches generally include the time-consuming operation of mapping all atoms in the query structure against those of candidate structures from the database. It is advantageous to perform this operation on as few molecules as possible by eliminating molecules that cannot possibly match. Approaches to reducing the number of structures to be compared vary.

One approach is to use keys that code for the presence or absence of specific structural features, such as functional groups of particular interest to medicinal chemists, five-membered rings, or metal atoms. A key is created by defining the structural features of interest, assigning a bit ('1' represents presence, '0' represents absence) to each one of these features, and generating a bitmap for each compound in the database. Keys are generally set when the compounds are registered. At search time, only those structures that have all the keys set by the query structure need to be examined for atom-by-atom mapping.

Like keys, fingerprints are bitmaps that are derived directly from the connection table of a molecule. However, whereas the bits in a key retain a direct mapping to structural features, bits in a fingerprint are subjected to mathematical operations ('folding') that reduce the size of the bit string and scramble the meaning of the individual bits.

A third approach is to build a specialized index for each structure in the database that encodes all possible ways to traverse the atoms and bonds of the structure. A query structure is compared against the atom-bond paths of the index. Only those structures that match need to be considered further^{57,58}.

SEMA NAME

A stereochemical extension of the Morgan algorithm. A compact, canonical representation of a connection table.

SUBSTRUCTURE

One chemical structure is said to be a substructure of another if the first structure can be located within the second. (The second is said to be the superstructure of the first.) All structures are substructures of themselves. A substructure search scans a database for all substructural matches.

CONFORMATIONAL SPACE

The ensemble of three-dimensional shapes that a molecule can adopt without breaking any bonds.

MARKUSH STRUCTURE

Markush structures represent a set of chemical structures as a common core that contains marked substitution sites, and a set of possible fragments for each substitution point. They can be used to represent a set of compounds that are analysed to determine the effect of varying substituents on compound activity; to represent a set of compounds that are produced using combinatorial techniques; to produce a fine-tuned substructure query; or to represent a set of structures in a chemical patent or patent database.

Oracle or IBM Informix. For chemical database users, this means the ability to store a complete chemical database within a relational database, and perform various types of chemical search that are discussed below. This opens up a wide range of possibilities when designing chemically aware applications, because these applications can make use of existing modes of communicating with relational databases. Furthermore, because a large amount of biological-assay and -property data also reside in the relational database, the ensemble of chemical, biological and property data can be queried and browsed in context.

Methodology of database searching

Chemical structures differ greatly from other entities that are commonly stored in databases, such as text or numbers, and so there are many differences between search methods for chemical databases and those for textual or numerical databases. Nevertheless, some parallels can be drawn between chemical database searches and searches on, for example, words. An exact-match search can be thought of as looking up a complete word in a dictionary. A substructure search is analogous to a wild-carded text search, and a similarity search resembles a 'sounds-like' search. Let's explore these search types in more detail.

Exact-match searching. The simplest kind of chemical searching is an exact-match search, in which a user looks for a given, fully specified chemical compound in a database. This type of search is generally well defined in the mind of the user — does the compound I have drawn exist in this database? Exact-match searches are performed to find out whether a proposed new structure already exists in a database, to determine the overlap between two databases by using all of the compounds in one database as queries in the other, or to find a

reagent in an in-house inventory system or online ordering system.

An exact-match search might yield no hits, even though the compound is present in the database. Depending on the flexibility of query specification and the query engine, this can happen if: the structure in the database has no marked **STEREOCHEMISTRY**, whereas the query does; the structure in the database is one **TAUTOMERIC** form, whereas the user has drawn the other; the structure in the database is an explicit salt (the counterion is drawn as a set of atoms), whereas the query structure is just the parent compound; or because of other differences in representation. It might be possible to overcome the problem in these cases by performing a similarity search (see below) with a high cutoff.

Substructure searching (2D). One of the most common structural searches is a substructure match, in which a user draws, copies or pastes a set of 'pieces' of a chemical structure, and requests that the system return a set of compounds that contain the pieces. This type of search is typically well defined in the mind of the user. That is, every experienced user has a similar expectation of what hits will result from such a search, and can generally tell, while browsing hits, how each answer satisfies the search question. (There are some issues with respect to stereochemistry, aromaticity and the limitations on atomic representation in a particular software package that could lead to 'surprises' for the user who performs a substructure search, but these can generally be resolved by closer study of the software package.)

On the database side, the implementation of substructure searching in arbitrarily large databases in a reasonable time frame is far from trivial. The crucial step in determining whether a molecule in a database matches a query involves examining atoms and bonds in detail, and can be time consuming. A necessary step is to pre-screen the database to eliminate from consideration those structures that cannot possibly match the query. Most systems compute keys or **fingerprints** that encode features or fragments of chemical structures (BOX 2) to allow rapid pre-screening.

Pharmacophoric searching. Pharmacophoric or 3D substructure searching involves a (generally sparse) set of atoms/bonds/groups that is combined with specific 3D constraints, such as distances and angles. This process is generally much slower than a 2D substructure search, as it requires the examination of *xyz* coordinates for atoms of the candidate structures to compute 3D constraints. Pharmacophoric searching can provide an indication of whether a set of structures can bind to a receptor or enzyme. This means that hits might be very valuable in the drug discovery and design process^{19–21}.

There are two general routes to generating a pharmacophoric search query. In the first case, the user has a high-quality 3D structure for a receptor or enzyme (protein) with a known agonist or antagonist (small molecule) attached. Here, one makes some assumptions about which groups of atoms on the small molecule are involved in binding, examines the spatial relationship of

PHARMACOPHORE

The ensemble of steric and electronic features that is necessary to ensure optimal interactions with a specific biological target structure and to trigger (or to block) its biological response.

STEREOCHEMISTRY

The spatial arrangements of atoms in molecules and complexes.

TAUTOMER

One of two or more structural isomers that exist in equilibrium and are readily converted from one isomeric form to another.

HYDROGEN BOND

A weak attraction (much weaker than a covalent or ionic chemical bond, but much stronger than van der Waals forces) between an oxygen, nitrogen or fluorine atom in one molecule and a hydrogen atom in a neighbouring molecule. Hydrogen-bond donors are groups with electron-hungry hydrogen atoms. Hydrogen-bond acceptors are atoms with electrons to share.

BIT STRING

A contiguous set of characters that consists entirely of 1s and 0s. A bit string can be used to encode a good deal of information in a compact way, and is easily and rapidly interpreted by computer systems.

AND

The combination of two input bits such that the result is 1 if both bits are 1 and 0 otherwise.

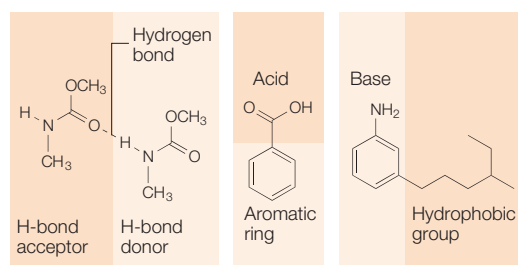


Figure 2 | Examples of atomic groups that are commonly used in three-dimensional queries.

these groups, and encodes this information into a 3D query, generally using a graphical tool. In the second case, no protein structure is available, but the scientist has a set of molecules that are known to bind to produce a desired pharmacological effect. 3D structures of the molecules are compared, generally using an automated overlay procedure, and areas of commonality thus identified are used to define the pharmacophoric query.

3D database queries are often different from 2D queries in one important way: the atoms in the query tend to be generalized types rather than specific chemical elements. Commonly used types are HYDROGEN-BOND acceptors, hydrogen-bond donors, acids, bases, aromatic rings and hydrophobic groups²² (FIG. 2). There might also be extra features in 3D searches, such as excluded volumes²³.

3D search systems generally allow the user to input extra parameters to control how tightly the 3D constraints must match — whether to consider conformational flexibility in the candidate matches, whether to check for van der Waals contacts, and so on. In particular, the issue of conformational flexibility is important, as most drug-like structures have more than one low-energy conformation (BOX 1). Typically, databases include a single 3D structure for each compound. In general, the more extra parameters that are used, the longer the search will take, but the more accurate (and therefore valuable) the hit list will be. Features and

algorithms in 3D structure searching are areas of active research to improve speed and discrimination^{23–25}.

During the search process, the software will generally perform a screening procedure, as in 2D substructure searching, then proceed to atom/bond mapping and constraint verification. Consideration of conformational flexibility generally takes place ‘on the fly’ as 3D constraints are mapped onto the candidate molecule. The procedure moves on to the next candidate when one valid conformation is found. Computational complexity due to any single database entry (highly symmetrical structures, for example) might lead to their inclusion in a hit list if search time limits expire.

Similarity searching. Whereas user expectations with respect to substructure matching are very well defined, similarity searching is often termed ‘fuzzy’ matching. The user wants compounds that resemble the compound of interest on the basis of a chemist’s intuitive thinking, but that do not necessarily reflect an exact or substructure match. The hope is that the biological system will respond to the molecules in a similar way, even though they represent different substances.

Similarity searching is also implemented very differently by the various software packages. Two factors are notable in the evaluation of similarity: first, the molecular property (or set of properties) evaluated for each molecule, and second, the coefficient, which is computed on the basis of this property to quantify how similar two compounds are^{26–28}.

Similarity properties. Most commercial programs calculate similarity on the basis of the keys or fingerprints that are used in the first step of 2D substructure searching (BOX 2). These keys or fingerprints are generally in a format that is relatively easy for a computer to work with — BIT STRINGS that can be ANDed together easily. Similarity searching is therefore a facile operation in most software.

Compound similarity has also been computed using atom pairs^{29,30}, sets of four consecutive atoms (also known as ‘topological torsions’³¹), sets of three or four disconnected atomic points²², and other molecular properties³². The basic aim is to find patterns within a molecule that provide a basis for assessing its degree of ‘likeness’ to another molecule, at the same time transcending the obvious substructural features. The relative merits of the keys, fingerprints, atom pairs, and so on, depend on the purpose at hand and the biases of the user^{33–40}.

Similarity coefficients. Once a property has been selected to describe each molecule, some way of turning that property into a numerical value that tells us how similar two molecules are must be generated. The most common similarity metric is the Tanimoto coefficient⁴¹, which is defined in BOX 3. Other coefficients include the cosine coefficient, the Euclidean distance coefficient and the Tversky coefficient⁴¹.

A user who wishes to perform a similarity search provides a structure (generally, a complete compound

Box 3 | Similarity coefficients

Once a set of features has been accepted for describing a molecule, a formula is needed to evaluate the degree of similarity between two molecules. By far the most common mathematical formula is the Tanimoto coefficient (EQN 1):

$$S_{A,B} = c/[a + b - c] \quad (1)$$

$S_{A,B}$ = similarity of structures A and B; c = number of features in common between the given property in the two structures (in the case of structure keys or fingerprints (BOX 2), this means the number of ON bits when the two bit strings are logically ANDed); a = number of features ON in structure A; b = number of features ON in structure B.

For example, for a hypothetical designation of five features that are present (1) or not (0) in two molecules, A and B:

Feature number	1	2	3	4	5
Molecule A	1	1	0	1	1
Molecule B	0	1	1	1	0

the similarity assessed by the Tanimoto coefficient is given by EQN 2:

$$S_{A,B} = 2/(4 + 3 - 2) = 0.4 \quad (2)$$

of interest), selects a database and also supplies a cutoff percentage, which limits retrieval to only those compounds that are more than this percentage similar to the input structure. At search time, the software compares the query structure with each structure in the

database (using a very fast bitmap comparison), computes the similarity coefficient, and compares this coefficient with the user's cutoff. Compounds that have equal or greater similarity are considered as 'hits', and are made available to the user for browsing.

Box 4 | Searching a chemical database

These instructions can be used to search for analogues of baclofen, a drug that acts at GABA (γ -aminobutyric acid) receptors (structure shown in panel a), in the freely accessible National Cancer Institute (NCI) Database:

Retrieve compound by name

- Go to the web address <http://cactus.nci.nih.gov/ncidb2/> or the german mirror, <http://www2.chemie.uni-erlangen.de/services/ncidb2/> (make sure all the text boxes on the right side of the screen are empty). Should you have any problems with these web addresses, please contact Marc Nieklaus at mn1@helix.nih.gov.

- In the top 'Query type' pull-down menu on the left side of the screen, select 'Name search...' On the right side, type 'baclofen' into the corresponding text box. Select 'Any exact name' from the pull-down menus that are immediately beneath the text box, and then click on the 'Start search' button.
- In the tabular results page, note that baclofen (NSC329137) is the only hit. Select it by clicking on one of the '329137' links.
- From the frame entitled 'Operations with this Structure (NSC329137)', click on the 'Transfer to Java Editor' button.
- Note that changes can be made to a query structure using this applet. For a first pass, make no changes. When ready, click on the 'Transfer to Query Form' button.

Set up a substructure search

- If you followed the steps listed above, you should now have the SMILES (simplified molecular input line system) for baclofen on your query form.
- If you skipped this step, you can paste the SMILES (NCC(CC(O)=O)C1=CC=C(Cl)C=C1) directly into the text box.
- If you'd like to practise drawing structures, you can use the 'Editor' button to invoke the structure drawing applet and draw baclofen, as shown in panel a above.

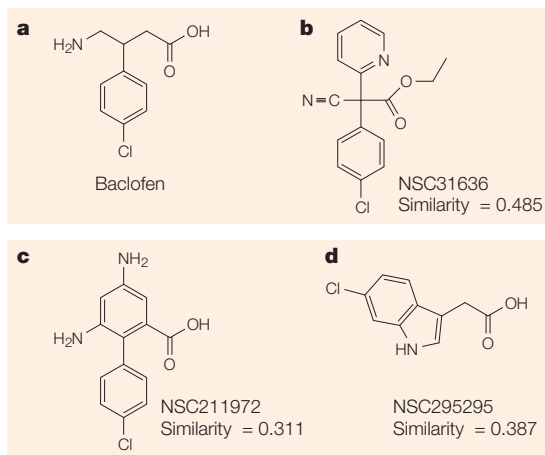
Perform the substructure search and view the results

- Make sure the drop-down menu on the left of the SMILES text box shows 'Substructure and/or 3D Search'. Make sure that no other search criteria are filled in.
- Click on a 'Start Search' button.
- Browse the results as desired.

Some selected results are shown in panels b, c and d, with their NSC number, and similarity to baclofen evaluated using the externally calculated Tanimoto coefficient (BOX 3).

Variations

- Perform a similarity search on baclofen or another one of the molecules retrieved.
- Modify baclofen in the structure editor and perform another substructure search. Deleting more atoms will increase the number of compounds in the hit list.
- Compare the properties of hits to those of baclofen.
- View the three-dimensional conformations for baclofen or another of the compounds in the hit list.



Molecular docking. One form of chemical database search that has gained momentum in recent years is 'docking', or placing a series of candidate molecules from a database into the active site of a protein to evaluate how well the compounds might bind to the receptor or enzyme. There are two basic problems to be solved in the docking process: how best to fit the small molecules (ligands) into the active site of the protein, and how to compare and rank the best 'poses' (or fittings) of a set of molecules in order to compare them.

Originally, small molecules were docked into active sites using a single rigid conformation⁴²⁻⁴⁴. Now, faster algorithms and better hardware allow the consideration of conformational flexibility^{9,45}. The methods of providing ligand flexibility include most of the accepted ways of evaluating conformations of 3D structures, and providing an overview of these is beyond the scope of this review. The functions that are used to score dockings are varied; a study of docking and scoring methods found that different scoring functions work best in different situations, and that a combination ('consensus') of scores might be the best way to rank molecules from a docking study⁴⁶.

The term 'virtual screening' is often applied to computational processes that select molecules that are likely to have activity against a biological target of interest^{9,24,45,47}. Docking is perhaps the computational technique most worthy of this moniker, as the molecules that are identified in a docking scan have been compared most directly with the requirements of the target.

Post-search processing of results

Realizing the value of a chemical database search begins after the search is complete. The user might face a list of compounds that is too large to be examined or tested using available resources. Some strategies include: filtering — essentially imposing secondary search criteria to eliminate compounds; clustering — taking a representative subset of a larger set; and human inspection of the compound structures (with or without extra data).

Filtering. The set of compounds can be pruned by eliminating those with properties that are deemed to be 'undesirable' or not drug like. One famous set of rules to determine which molecules are most likely to behave pharmaceutically was developed by Lipinski and colleagues⁴⁸. Lipinski's profile includes: molecular mass <500 g mol⁻¹; log of octanol/water partition coefficient (LOG P) <5.0; no more than five hydrogen-bond donors; no more than ten hydrogen-bond acceptors.

A chemist may also remove from consideration all compounds without available samples if the compound is part of an in-house database, or compounds that cost too much if the database represents a set of commercial catalogues. Molecules with groups that are

LOG P

The octanol/water partition coefficient is the ratio of the solubility of a compound in octanol to its solubility in water (also known as K_{ow}). The logarithm of this partition coefficient is called log P. It provides an estimate of the ability of the compound to pass through a cell membrane.

Box 5 | Database techniques in lead discovery: a case history

A recent study by Pang and colleagues⁴⁷ shows the usefulness of database techniques in the discovery of new lead compounds. The biological target in this study was the enzyme farnesyltransferase, the inhibition of which might lead to anticancer activity. Structures were obtained from MDL Information Systems' Available Chemicals Directory (ACD) database, and were pre-screened to remove compounds with molecular masses outside the desired range of 300–700 Da, charged species and those without aromatic rings. This yielded a set of 67,928 compounds, which were docked into farnesyltransferase using the in-house flexible docking programme EUDOC⁵⁹ in two stages: low resolution (30° rotational increments and 3.0-Å translational increments) and high resolution (10° rotational increments and 1.0-Å translational increments).

The docking yielded a set of 313 compounds, which was further reduced using estimated solvation energies to eliminate those considered to be too hydrophilic or too hydrophobic. Visual inspection was used to remove compounds with overreactive chemical functionality to yield 27 compounds, which were then purchased. Chemical analysis showed that six of these compounds were highly impure, so they were removed from further consideration. Of the final 21 compounds that were submitted for testing, four were found to inhibit the target enzyme in 25–100- μ M concentrations. As a control, 21 other compounds were randomly selected from ACD. None of these had activity in the 25–100- μ M range.

This study is important, because it shows the usefulness of virtual screening techniques in dealing with an enzyme with an active site that is large and contains metallic ions, and because a control was used to rule out the possibility that active compounds were identified merely by chance.

deemed to be 'overly reactive', such as acid chlorides, which react violently with water, are also excluded from further consideration.

Clustering. In many cases, even after applying filters, there are still too many compounds to test, or the compounds in the hit list resemble one another to such an extent that there is no point in testing all of them⁴⁹. In these cases, a chemist can perform a clustering of the database to group similar compounds^{32,50–52}. A representative compound from each cluster is sent for biological assay. Various computational methods exist for performing clustering, including Wards⁵³, Jarvis-Patrick⁵⁴ and Guénoche⁵⁴. The metrics that are used as input are often those used for similarity searches (see above)^{35,55}. Clustering can also be applied to understand the underlying set of chemical scaffolds that are represented in a large hit list; this can be a useful prelude to analysis of structure–activity relationships.

Human inspection. Having a person look over the results of a database search is by far the most time-consuming process in the consideration of what might be a large set of results. Inspection might involve scanning a matrix of structures on a page, or a more in-depth analysis of a small number of compounds in the context of further information. This further information might include measured or calculated physical properties, analytical test results, molecular spectra, biological-assay results, and so on. The process requires a good deal of effort, but it can yield valuable results because of the insights that can be drawn by seeing a set of structures in the wider context of the research process.

Conclusion

As we have seen, chemical structures are entities that are substantially different from text, numbers and other more common data. They have their own representation requirements and unique searching modes. For those who would like some 'hands on' experience of chemical database searching, a guide to searching the [National Cancer Institute Database](#), which illustrates some of the principles discussed above, can be found in BOX 4. A recent case history that highlights the potential for chemical database searches in lead discovery is described in BOX 5.

Many vendors now offer basic software for registration, searching, visualization and analysis of chemical structures (see ONLINE TABLE 1). What will differentiate these products? The usefulness of chemical database searches for lead discovery often depends on factors beyond the ability to perform the searches themselves. First, users must have the ability to perform searches on chemical structures in the context of related data, including chemical properties, biological activity and protein structural data. There must be tools to allow the researcher to wade through the often large hit sets, to select the most promising compounds for further evaluation. The user must have the ability to browse the results within the context of related research information in order to draw the proper inferences from the data. The specifics of these requirements vary from organization to organization, but adaptability of the software is crucial, as is the ability to interface with other analytical and visualization tools. Products that can fulfil these requirements will succeed, whereas those that merely have good technology will fail.

1. Voigt, J. H., Bienfait, B., Wang, S. & Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **41**, 702–712 (2001).

An excellent analysis of publicly and commercially available chemical databases.

2. Trinajstić, N. (ed.) *Chemical Graph Theory* (CRC, Boca Raton, 1983).

3. Balaban, A. T. Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci.* **25**, 334–343 (1985).

4. Dalby, A. et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **32**, 244–255 (1992).

5. Dury, L., Latour, T., Leherite, L., Barberis, F. & Vercauteren, D. P. A new graph descriptor for molecules containing cycles.

Application as screening criterion for searching molecular structures within large databases of organic compounds *J. Chem. Inf. Comput. Sci.* **41**, 1437–1445 (2001).

6. Weininger, D. SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31 (1988).

7. Warr, W. A. Combinatorial chemistry and molecular diversity. An overview. *J. Chem. Inf. Comput. Sci.* **37**, 134–140 (1997).

8. Leland, B. A. Managing the combinatorial explosion. *J. Chem. Inf. Comput. Sci.* **37**, 62–70 (1997).
9. Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening — an overview. *Drug Discov. Today* **3**, 160–178 (1998).
- This article provides an excellent overview of the 'hows and whys' of using computers to select molecules for testing.**
10. Schultz, J. L. & Wilks, E. S. Dendritic and star polymers: classification, nomenclature, structure representation, and registration in the DuPont SCION database. *J. Chem. Inf. Comput. Sci.* **38**, 85–99 (1998).
11. Wipke, W. T. & Dyott, T. M. Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **96**, 4834–4840 (1974).
12. Gasteiger, J., Rudolph, C. & Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comp. Methodol.* **3**, 537–547 (1990).
13. Pearlman, R. S. CONCORD: rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Autom. News* **2**, 1 (1987).
14. Rusinko, A. Using CONCORD to construct a large database of three-dimensional coordinates from connection tables. *J. Chem. Inf. Comput. Sci.* **29**, 327–333 (1989).
15. Crippen, G. M. & Havel, T. F. Stable calculation of coordinates from distance information. *Acta Cryst.* **A34**, 282–284 (1978).
16. Hahn, M. Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci.* **37**, 80–86 (1997).
17. Paris, C. G. Chemical structure handling by computer. *Annu. Rev. Inform. Sci. Technol.* **32**, 271–338 (1997/1998).
- An excellent overview of the issues in storing, searching and analysing molecules using a computer.**
18. Barnard, J. M. & Downs, G. M. Computer representation and manipulation of combinatorial libraries. *Persp. Drug Discov. Des.* **7/8**, 13–30 (1997).
19. Martin, Y. C., Bures, M. G. & Willett, P. Searching databases of three-dimensional structures. *Reviews Comput. Chem.* **1**, 213–263 (1990).
20. Good, A. C. & Mason, J. S. Three-dimensional structure database searches. *Reviews Comput. Chem.* **7**, 67–117 (1995).
21. Nicklaus, M. C. *et al.* HIV-1 integrase pharmacophore: discovery of inhibitors through three-dimensional database searching. *J. Med. Chem.* **40**, 920–929 (1997).
22. Pickett, S. D., Mason, J. S. & McLay, I. M. Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* **36**, 1214–1223 (1996).
23. Greenidge, P. A., Carlsson, B., Bladh, L.-G. & Gillner, M. Pharmacophores incorporating numerous excluded volumes defined by X-ray crystallographic structure in three-dimensional database searching: application to the thyroid hormone receptor. *J. Med. Chem.* **41**, 2503–2512 (1998).
24. Olender, R. & Rosenfeld, R. A fast algorithm for searching for molecules containing a pharmacophore in very large virtual combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **41**, 731–738 (2001).
25. Mason, J. S. *et al.* New four-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privilege substructures. *J. Med. Chem.* **42**, 3251–3264 (1999).
26. Downs, G. M. & Willett, P. Similarity searching in databases of chemical structures. *Rev. Comput. Chem.* **7**, 1–66 (1995).
27. Singh, S. B., Sheridan, R. P., Fluder, E. M. & Hull, R. D. Mining the chemical quarry with joint chemical probes: an application of latent semantic structure indexing (LaSSI) and TOPOSIM (Dice) to chemical database mining. *J. Med. Chem.* **44**, 1564–1575 (2001).
28. Hefferlin, R. & Matus, M. T. Molecular similarity for small species: refining the isoelectronic index. *J. Chem. Inf. Comput. Sci.* **41**, 484–494 (2001).
29. Sheridan, R. P., Miller, M. D., Underwood, D. J. & Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **36**, 128–136 (1996).
30. Hull, R. D. *et al.* Latent semantic structure indexing (LaSSI) for defining chemical similarity. *J. Med. Chem.* **44**, 1177–1184 (2001).
31. Nilakantan, R., Bauman, N., Dixon, J. S. & Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **27**, 82–85 (1987).
32. Brown, R. D. & Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **37**, 1–9 (1997).
33. Schuffenhauer, A., Gillet, V. J. & Willett, P. Similarity searching in files of three-dimensional chemical structures: analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **40**, 295–307 (2000).
34. Rhodes, N., Willett, P., Dunbar, J. B. Jr & Humblet, C. Bit-string methods for selective compound acquisition. *J. Chem. Inf. Comput. Sci.* **40**, 210–214 (2000).
35. Xue, L., Stahura, F. L., Godden, J. W. & Bajorath, J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **41**, 746–753 (2001).
36. Butina, D. Unsupervised database clustering based on Daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **39**, 747–750 (1999).
37. Xue, L., Stahura, F. L., Godden, J. W. & Bajorath, J. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *J. Chem. Inf. Comput. Sci.* **41**, 394–401 (2001).
38. Matter, H. & Pötter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **39**, 1211–1225 (1999).
39. McGregor, M. J. & Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **39**, 569–574 (1999).
40. Xue, L. & Bajorath, J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **40**, 801–809 (2000).
41. Willett, P., Barnard, J. & Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).
- A very comprehensive discussion of similarity searching.**
42. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* **161**, 269–288 (1982).
43. Shoichet, B. K., Bodian, D. L. & Kuntz, I. D. Molecular docking using shape descriptors. *J. Comput. Chem.* **13**, 380–397 (1992).
44. Meng, E. C., Shoichet, B. K. & Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13**, 505–524 (1992).
45. Bissantz, C., Folkers, G. & Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **43**, 4759–4767 (2000).
46. Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **42**, 5100–5109 (1999).
47. Perola, E. *et al.* Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **43**, 401–408 (2000).
48. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
49. Higgs, R. E., Bemis, K. G., Watson, I. A. & Wikel, J. H. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.* **37**, 861–870 (1997).
50. Brown, R. D. & Martin, Y. C. Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **36**, 572–584 (1996).
- Applications of database clustering.**
51. Xue, L., Godden, J., Gao, H. & Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal components analysis. *J. Chem. Inf. Comput. Sci.* **39**, 699–704 (1999).
52. Mason, J. S. & Pickett, S. Partition-based selection. *Persp. Drug Discov. Des.* **7/8**, 85–114 (1997).
53. Barnard, J. M. & Downs, G. M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* **32**, 644–649 (1992).
54. Guénoche, A., Hansen, P. & Jaumard, B. Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *J. Classification* **8**, 5–30 (1991).
55. Barnard, J. M. & Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **37**, 141–142 (1997).
56. Lavecchia, A., Greco, G., Novellino, E., Vittorio, F. & Ronsisvalle, G. Modelling of κ -opioid receptor/agonist interactions using pharmacophore-based and docking simulations. *J. Med. Chem.* **43**, 2124–2134 (2000).
57. Rughooputh, S. D. D. V. & Rughooputh, H. C. S. Neural network based chemical structure indexing. *J. Chem. Inf. Comput. Sci.* **41**, 713–717 (2001).
58. Ozawa, K., Yasuda, T. & Fujita, S. Substructure search with tree-structured data. *J. Chem. Inf. Comput. Sci.* **37**, 688–695 (1997).
59. Pang, Y. P., Perola, E., Xu, K. & Prendergast, F. G. EUDOC: a computer programme for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.* **22**, 1750–1771 (2001).

Acknowledgements

M. M. would like to thank his colleagues in the scientific architecture group for useful feedback on the manuscript.

 Online links

FURTHER INFORMATION

CACTVS: <http://www2.ccc.uni-erlangen.de/software/cactvs>

ChemIDPlus: <http://chem.sis.nlm.nih.gov/chemidplus/>

ChemWeb: <http://www.ChemWeb.com/>

Daylight Chemical Information Systems:

<http://www.daylight.com/>

Fingerprints — Screening and Similarity:

<http://www.daylight.com/dayhtml/doc/theory/theory.finger.html>

National Cancer Institute Database:

<http://cactus.nci.nih.gov/ncidb2/>

SMILES:

<http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>

Access to this interactive links box is free online.