CHEMOINFORMATICS

1. What is Chemoinformatics?

Different definitions of chemoinformatics (1) have been given but, within the context of this article we will view it broadly as the management, analysis, and dissemination of data related to chemical compounds. Chemoinformatics results from the application of methods in information technologies to problems in chemistry.

During the last decade, chemoinformatics has become one of the essential tools in the early stages of pharmacological and agrochemical discovery. The reason for its importance is rooted in the emergence of high throughput screening and high throughput chemical synthesis as the dominant technologies for the discovery of starting points for chemical optimization (2). The use of robotics for screening and large chemical libraries has resulted in extremely large volumes of data that require informatics management.

Initially, chemical collections, commonly referred to as libraries, comprised small numbers of distinct chemicals (3). Nowadays, much larger compound collections are routine throughout the chemistry-based industries. The need to evaluate ever expanding libraries requires the development of tools for storage of the information generated, data analysis, the identification of trends in the data, and their eventual correlation to the structural and physicochemical properties of the compounds. In the near future, the challenges in this area will be compounded by the integration of developments in genomics, proteomics, and bioinformatics (4). Chemoinformatics work is and will continue to be multidisciplinary, because it acts at the interface between chemistry and informatics, as well as the multiple disciplines that use it.

Some avenues of research in chemoinformatics evolve from observations made as its tools are applied. An example is provided by diversity analysis. In the past, it was observed that if the compounds evaluated in high throughput screening showed a high degree of structural similarity, the result would be sparsely successful, or have a limited numbers of related hits (5). Consequently, the design of the libraries for screening based on chemical diversity ideas became a crucial step for lead discovery. Methods that provide objective measures of the dissimilarity among compounds to be acquired or synthesized (6,7) are part of the chemoinformatics realm, which was developed to avoid the repeated evaluation of the same chemical classes (8). Because of the wide range of subjects, all of these aspects of chemoinformatics will be discussed only briefly.

The main purpose of chemoinformatics is to provide tools for the efficient management of information, a critical step in any decision making process. Chemoinformatics transforms data into information and subsequently into knowledge, thus greatly facilitating all aspects of chemical research. This field is continually expanding and the number of applications and tools available is very large. Therefore, only some of the many algorithm approaches will be described here, with a particular emphasis on analysis of chemical information.

2. Chemical Information Storage

Perhaps the most important task in the creation of a chemical database is the definition of the fields to be stored. The type and scope of the information to be stored should be pondered carefully at the onset of a project. The database design requires particular attention, because errors or lack of foresight when creating it are painful to correct as the systems are deployed throughout the organization. Depending on the type of information, the project may need to be restarted. Yet, such foresight is challenging because chemical databases used in research are continuously evolving together with the data collected. Even at the earlier stages of the project, input from the end-users is a requisite for the design of any database.

The distinctive feature of a chemical database is that it allows the storage and retrieval of structural information as well as textual or numerical information on a chemical. All datatypes, including chemical, numerical, and textual information, can be combined when querying a chemical database. Simple queries may include combinations of datatypes such as 'Display all compounds with an imine functionality that cost less than a given amount and are currently in inventory', or 'Display all benzimidazoles that have been made between 1971 and 1983 that are still available', or 'Display all thiazoles that show no cytotoxicity at 10 μm concentrations'. Without the chemical structure component, the same searches could not be done within the framework of textual or numerical datatypes alone.

Representation of chemical structures in a computer searchable form requires the adoption of special formats. While multiple formats have been used over the years, the dominant chemoinformatics software provides a relationship between structure and either tables or lines that are intrinsically computer searchable. The two dominant file formats for structural representation are the SMILE strings and the MOL files, discussed below.

2.1. Line Notations.

Alternatives to the valence representation of the molecular structure in the form of lines or strings have been pursued for decades, even before the use of computers in chemical information storage and management. Earlier attempts included the well-known Wiswesser line notation, or Bielstein's ROSDAL (9). Currently, SMILES (Simplified Molecular Input Line Systems) is one of the most popular notations in this class. The SMILES notation was developed by Weininger and co-workers and is commonly associated with Daylight software (10–12).

In a SMILES string, each atom is identified by its element symbol, as well as additional information that is placed into brackets, including chirality and net charge. Single bonds are not made explicit, double bonds are indicated as "=", while the triple bonds are shown as "#". Aromatic bonds are represented by ":", bond alternancy or more commonly the aromaticity of a ring, is indicated by using lower case letters for the atoms in aromatic rings. For salts, the smiles string of the ion and the counterion are connected by a dot. In all cases, hydrogens are not made explicit, unless required to establish isomerism. Examples of these representations are shown in Figure 1.

The representation of rings requires two steps (12). First, one bond per ring is broken in such a way that an acyclic structure results. There is always a way to break one bond per ring in a structure so that the result is an acyclic molecule. Second, broken bonds are numbered and the string for the resulting acyclic structure is written. In the resulting string, the numbers assigned to each broken bond are placed next to the adjacent atoms. Examples are also shown in Figure 1.

Geometric isomerism is indicated by the use of both the forward and backslash. Before and after a double bond, if the same type of slash is used to show bonds attached to the double bond, then the arrangement of the centers is *trans*, while if opposite, the atoms are *cis*. Optical isomerism is indicated by the use of "@" or "@ @"; if the order of the atoms attached to the chiral center are ordered anticlockwise, or clockwise, respectively.

One of the limitations of the SMILE strings is that there could be more than 1 equivalent string for the same molecule, because they depend on the internal numbering system of the structure used. One given structure will not always yield the same representation, but it will depend on the algorithm used. A canonical representation is one where rules are defined to the extent that only one string is the correct representation for any chemical structure. "Unique SMILES" (USMILES) are such representations (12).

2.2. Table Representation.

The most common alternative to the string or line notation is the use of connectivity or connection tables. File formats that incorporate connection tables have been described in detail in the literature, but are commonly associated with the software developed by MDL Inc. An example of a file that incorporates a connection table is shown in Figure 2 (13,14).

MOL (molecule) and SD (structure data) files contain connectivity tables. Both types of files have a "counts′ line", after comments lines that specify the total number of atoms in the file, the number of bonds, atom lists, and information on chirality. In addition, since the format of the files has been mildly modified with time, the version of the file is included. The line is followed by an atom

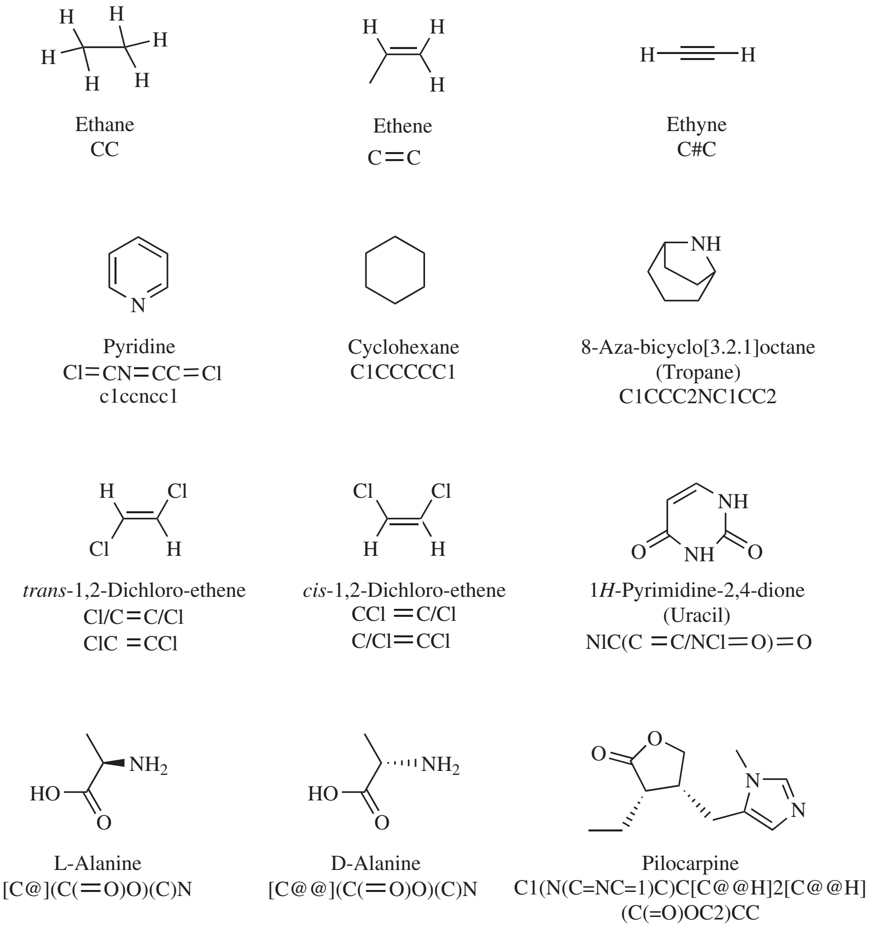
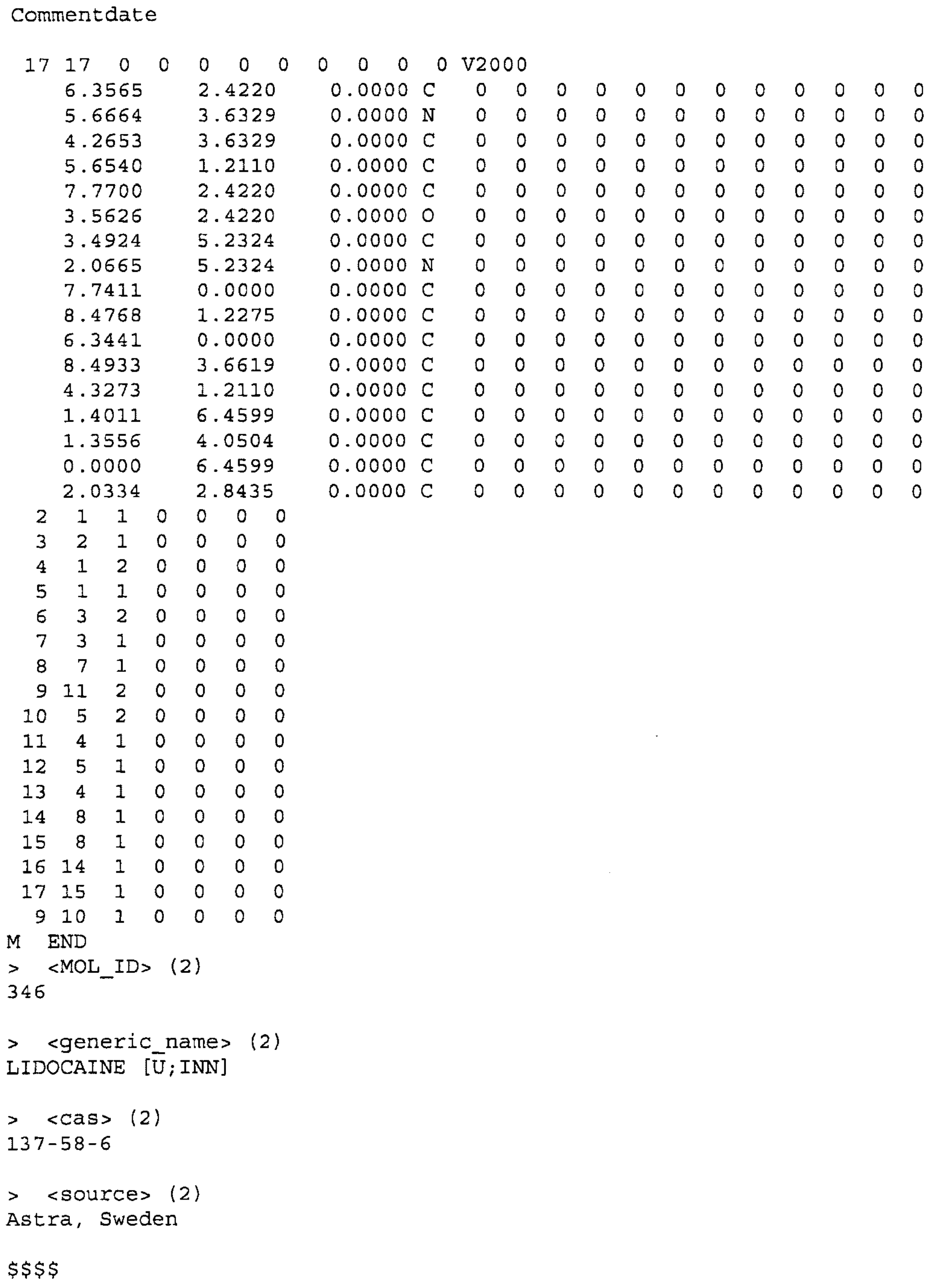


Fig. 1. Examples of compounds with their structure and their SMILES string underneath.

block that contains the atom symbol, charge stereochemistry, attached hydrogens for each atom, and a set of Cartesian coordinates for each atom. In two-dimensional (2D) representations, these coordinates can be used to plot a flat molecular structure. However, the coordinate fields can be used to store a three-dimensional (3D) representation of the molecule, in cases where a spatial arrangement of atoms is available. The ability to store (3D) information in some cases could be an advantage of table representations. The bond block specifies the atoms connected, the bond type, and any stereochemistry or topology associated with the bond. Atom list blocks and a structural text descriptor are also part of the file, though not always explicitly.

The actual connection table follows the atom block, where each bond is represented by each atom and the bond order. Molecular properties including net charge, radical character, isotope, etc, are stored in subsequent lines that



```
Commentdate

 17 17  0  0  0  0  0  0  0  0  0 V2000
    6.3565    2.4220    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.6664    3.6329    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    4.2653    3.6329    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.6540    1.2110    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    7.7700    2.4220    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.5626    2.4220    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.4924    5.2324    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.0665    5.2324    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    7.7411    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    8.4768    1.2275    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.3441    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    8.4933    3.6619    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.3273    1.2110    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4011    6.4599    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.3556    4.0504    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    6.4599    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.0334    2.8435    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  3  2  1  0  0  0  0
  4  1  2  0  0  0  0
  5  1  1  0  0  0  0
  6  3  2  0  0  0  0
  7  3  1  0  0  0  0
  8  7  1  0  0  0  0
  9 11  2  0  0  0  0
 10  5  2  0  0  0  0
 11  4  1  0  0  0  0
 12  5  1  0  0  0  0
 13  4  1  0  0  0  0
 14  8  1  0  0  0  0
 15  8  1  0  0  0  0
 16 14  1  0  0  0  0
 17 15  1  0  0  0  0
  9 10  1  0  0  0  0
M  END
>  <MOL_ID> (2)
346

>  <generic_name> (2)
LIDOCAINE [U;INN]

>  <cas> (2)
137-58-6

>  <source> (2)
Astra, Sweden

$$$$
```

Fig. 2. Example of SD File. The first lines are identical to a MOL file.

start with an M and a word that indicates the property contained in that line. The MOL terminates in a 'MEND' line. While MOL and SD files are identical at this point, SD files also allow the storage of other properties associated with the molecule, as well as multiple molecules in a single file.

Molecular properties provided by the user that are to be stored in an SD file are indicated by a '>' sign followed by the property name in between brackets ('< >') as shown in Figure 2. The information for each molecule is separated by a blank line and "$$$$".

Other file formats centered on the connectivity tables are available. Reaction data (RD) files (14), are similar to the basic SD file but are able to contain structural data for the reactants and products of a reaction, as opposed to individual molecules.

3. Chemical Information Retrieval: Data Searching

The use of computer readable formats to store structural information is the key to generating software that will be capable of searching such data. The search and display of textual and numerical data can be done with Boolean operators (AND, OR, LESS THAN, GREATER OR EQUAL TO, etc). However, the unique feature of a chemical database is in the handling of structural information, and that is where we will focus our discussion.

Different types of searches on structural information can be carried out (15,16). Two-dimensional information is searched differently from 3D. Two-dimensional searches can be done with the purpose of (*1*) identifying an exact chemical structure; (*2*) identifying a molecule or molecules that contain a given structural feature, commonly referred to as a substructure search; and (*3*) to search for molecules that look like those in another used as a query, which is described as a similarity search. Three-dimensional searches can be carried out to identify molecules that have predetermined pharmacophoric features in a correct spatial arrangement, with or without explicit knowledge of the 3D structure of the target.

3.1. Two-Dimensional Searches. *Structural and Substructural Searches.* The problem of identifying two identical chemical structures could be relatively straightforward, if a canonical representation of a molecule is used (12). However, table representations are not canonical, as atoms could be ordered in a certain sequence in the database, and entered in a different way in the query. Such comparison would require going through all possible permutations of the atoms, which is computationally prohibitive even for relatively small molecules. Algorithms that are more efficient have to be implemented to search in real time, and are the centerpiece of the searching for structural information.

Significant gains can be achieved by the use of filters to limit the number of structures to be explicitly compared. In that way, the expensive exhaustive searches would only be needed on smaller numbers of compounds instead of the entire database. A series of computed properties can be stored as the compounds are loaded into a database and can then be used to establish a similarity or identity to a query fragment or structure.

Structural keys, such as MOLSKEYS, are properties evaluated as compounds are loaded into the database (16) that could be used to reduce the number of pairwise comparisons. Structural keys are binary strings (set of zeros and ones) that indicate whether a given characteristic is present in the molecule. For example, if the molecule is charged, a predetermined bit in the string will be set, a different one will indicate the presence of aromatic amines, while yet another bit could be set by the presence of a carbonyl group. In the end, a vector is created that shows the presence or absence of predefined features. The resulting strings can be compared efficiently by multiple algorithms. For example, look-up tables are created that list all compounds having a particular bit set. Compounds present in all the lists that depend on the query can subsequently be pairwise compared. While not unique for a compound, the comparison of fingerprints greatly reduces the number of structures that have to be compared pairwise. Software, such as Chemfinder (17) or MDL's ISIS relies on the use of structural keys.

Alternatives to the structural keys are molecular fingerprints (12). Contrary to the structural keys, in the case of the fingerprints, there is no preassigned meaning to each bit. Fingerprints are also bit strings but are deprived of a direct meaning as found with structural keys. The process of generating fingerprints is initiated by an exhaustive enumeration of all linear patterns in a molecule, from a list of atoms to paths up to a determined length, which is typically seven bonds. The number of conceivable paths could be extremely large, which makes the assignment of each path a position in a predetermined bit string prohibitive. Instead, each pattern serves as a seed to a pseudonumber generation, the output of which is a set of bits. In more technical terms, the pattern is hashed. The set of bits is then composed in a series of Boolean operations that result in the actual fingerprint. Fingerprints can be handled like structural keys with Boolean operations for structure related searches (12).

The search for a substructure is another common problem in chemoinformatics. The substructure search is the process of finding particular fragments or patterns in a molecule. For table notations, the problem is similar to the identification of complete structures because they also use structural keys, but in this case, only the list of features present (bits set to one) need to be analyzed. The software retrieves compounds from the look-up tables that are associated with all of the features. The fingerprints can be used to match part of the structure as well. A connectivity table can be generated for the fragment and can verify if it is contained in other objects in the database.

Line notations have the additional challenge that the queries are only a part of a structure. SMILES strings are designed for complete structures. Consequently, to construct a query (molecular fragment) a different notation is needed. In the case of SMILES strings, similarly built SMARTS provide the query language (12).

Similarity Searches. Often times, searches are not for specific compounds or compounds containing a given molecular fragment exactly. Searches could be for compounds that are similar to others or are variants of others in different ways. Such a search requires that the similarities between two molecules be quantified, which is not a simple endeavor. For example, molecules could be similar biologically, have a similar arrangement of key functional groups in space, or be similar in physicochemical characteristics. For database searches, similarity is understood to be structural. In this section, we will be limited to the 2D similarities. Later in this article we will consider other issues on similarity search.

As already described, a compound can be represented by a collection of qualitative properties that describe general aspects of the structure, in the form of a structural key, such as MDLs MOLSKEYS. The similarity between two molecules can be reduced to measuring the similarity between the two binary strings that represent each molecule. Mathematical methods exist that permit such evaluation. Perhaps the most commonly used parameter to measure similarity between binary representations is the Tanimoto Coefficient (18,19). The Tanimoto coefficient is the ratio of bits set (ie, equal to 1) in both molecules to the total number of bits set in either structure. Figure 3, should help us to understand this definition. An alternative to the use of the Tanimoto coefficient is the XOR (exclusive OR) operator, which is simply a count of the number of positions at which the bitstrings for both molecules differ. This operator is also known as city block or Hamming distance, and can be shown to be identical to the square of the Euclidean distance between the two binary strings.

A generalization of the Tanimoto coefficient is the Tversky similarity (20). The coefficient is defined as

$$(\mathrm{Q\ AND\ M})/(\alpha\ \mathrm{Q\ AND\ (NOT\ M)} + \beta\ \mathrm{M\ AND\ (NOT\ Q)} + (\mathrm{Q\ AND\ M}))$$

If α and β are set equal to 1, it reduces to the Tanimoto coefficient. If $\alpha > \beta$ means that the features of the query are weighted more heavily and this is commonly referred to as a “superstructure-likeness” search. In the opposite case, where $\alpha < \beta$ produces a “substructure likeness” search, in that case the completely embedded structures have a higher similarity. Super- and substructure likeness searches are part of the major software that runs chemical databases.

3.2. Three-Dimensional Searches: Virtual Screening. The understanding of small molecule protein interactions is key to pharmaceutical and modern agrochemical research because nearly all compounds interact with proteins to elicit their biological activity. Methods to predict those interactions have become of paramount importance. Virtual screening is the process that permits the selection of the compounds that are most likely to interact with a potential target, from a much larger set that is either available or computationally created. Three-dimensional searches are a key component of the virtual screening process (21) because the interaction of a small molecule with a protein depends on the spatial arrangement of their functional groups. Three-dimensional searches can be done based exclusively on the structure of the ligands, without explicit knowledge of the protein target, or they could be done based on the structure of the protein target or direct drug design (22,23). The former is commonly denoted as indirect drug design (24,25), because the process requires the development of hypotheses about the preferences of the protein target of unknown structure based on the types of ligands it binds and those that it does not bind.

The spatial arrangement of functional groups in small molecules is critical in either case: direct or indirect design. The storage of molecular structure becomes important to carry out searches. The creation and maintenance of databases of 3D structures of compounds is essential for chemoinformatics work.

	P1	P2	P3	P4	P5	P6	P7
A	1	2	3	4	5	6	7
B	1	2	1	4	5	6	3
C	2	1	7	7	7	6	5
D	7	1	1	1	2	3	4
E	2	1	3	4	5	6	7
F	7	6	5	4	3	2	1
G	6	7	4	5	2	3	1
H	7	7	6	6	1	1	1
J	5	5	5	6	2	2	2
K	3	4	2	2	4	4	6
L	4	3	2	2	6	6	4
M	1	3	2	5	4	7	6
N	5	1	3	4	6	7	7
O	5	7	6	7	7	1	2
P	7	4	4	3	3	3	2
Q	6	6	6	5	5	5	1

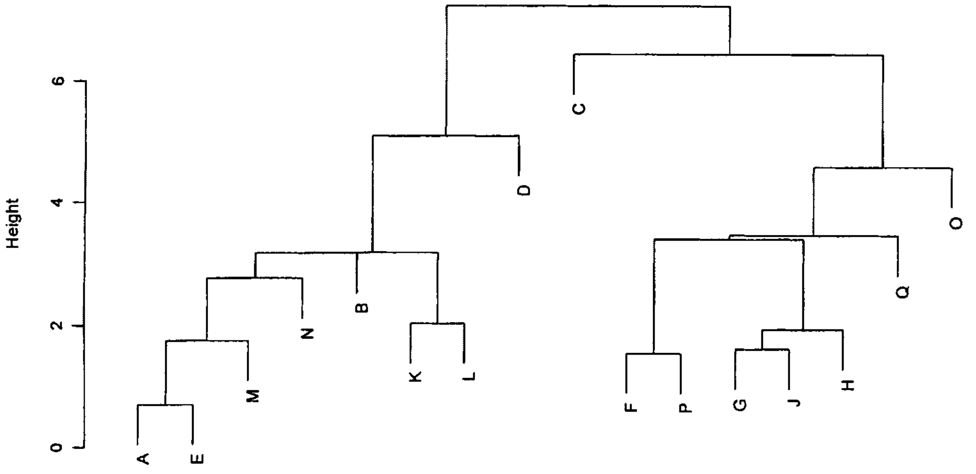


Fig. 3. Examples of bitstrings for a molecule M and a query Q. Tanimoto coefficient is 7/10 or 70% similarity between Q and M, while XOR is 10.

3.3. Construction of Three-Dimensional Chemical Databases.

Structural information about the small molecule is necessary to build a 3D chemical database. The information can be obtained from experimental sources, such as the Cambridge Crystallographic Database. The Cambridge Structural Database (26) (CSD) contains crystal structure information for >230,000 organic and metallorganic compounds. All of these crystal structures have been analyzed using X-ray or neutron diffraction techniques. Alternatively, computational techniques can be used to convert 2D representations of a molecule (a SMILES string or a MOL or SD file) to a 3D structure.

The conversion programs apply a knowledge base to construct the 3D structure. The pioneer program for the rapid conversion of 2D to 3D structures is CONCORD (27), which combines rules with energy estimation procedures in an attempt to produce the lowest energy 3D conformation for each structure. The procedure uses a fragment-based approach, where different portions of each molecule are constructed separately, and pieced together to form a complete molecule. Tables are used for acyclic bond lengths and angles; torsions are assigned to generate optimum interactions between atoms that are not directly linked. For rings, bond lengths and angles are calculated using preassigned rules, and assignment of gross conformations for each ring provides a framework for the torsions. A strain minimization function removes clearly problematic areas in the resulting structure. The structures can be further optimized using molecular mechanics with other programs.

CORINA (28), like CONCORD, is also a rule- and data-based algorithm, but with a superior ring handling technique, as well as organometallics. CORINA has been reported to have a higher conversion rate than CONCORD and other similar software for this purpose. A different approach to the problem is provided by Converter (21). It uses a distance geometry approach, coupled with upper and lower interatomic distance bounds, together with topological rules to generate the 3D structure. The procedure is somewhat slower than the purely rule-based algorithm, and may be less suitable for the conversion of large datasets or if structural information has to be generated as it will be used. That will be the case for software based on-line notations that do not store spatial information but generate it as required.

Conformational Flexibility, Conformational Searches, and Flexible Searches. Conformational flexibility of the molecules presents an additional problem for the 3D searches in chemical databases (25). If only one structure is stored, the searches will be incomplete, because conformations other than the one stored may satisfy the search criteria. A solution to the problem is to store multiple conformations for every compound.

The explicit storage of conformations requires a plan for sampling the small molecule conformational space. The problem of conformational searching in computational chemistry has attracted significant attention for many years and has been the subject of multiple, excellent reviews (29). The approaches devised have relied, for the most part, on the use of molecular mechanics to sample the conformational space of a molecule, by means that include genetic algorithms, stochastic searches, and distance geometry, among others. Since storing all conformations of a molecule is impractical, which conformations should be retained is also a significant issue for all but the most rigid compounds. Many of the conformations generated could be redundant, or essentially identical to others already stored.

One approach commonly used to decide which conformers should be retained has been the generation of extensive conformational libraries, followed by their clustering into related families. The clustering is done using the root-mean-square (rms) deviation in the Cartesian coordinates of the different conformers. Typically, one representative from each cluster is selected.

A more efficient alternative for sampling, which alleviates the time that is required to carry out a full conformational search, is poling (30). This method avoids oversampling of regions of conformational space that have been explored. Still, only individual conformations are stored or scrutinized. Poling represents a significant improvement in terms of the time and amount of conformational space that is sampled compared to other procedures. However, both poling and clustering store only representatives of a much larger set. The set of representative conformers selected may not include the exact spatial arrangements in a query, even though they may be accessible to the molecules under scrutiny in conformations not stored.

With different variants, torsional (21,31) fitting is a technique to deal with the issue of missed conformations that may satisfy a 3D query. Different implementations of the technique, such as the directed tweak algorithm, involve the optimization of rotatable bonds so that they would meet the elements of the query if possible. For a large chemical database, torsional fitting is a process that is computationally demanding, as it involves an optimization step. Filters are applied to limit the number of molecules that must undergo the torsional search. Filters applied depend on the software implementation, or on the protocol defined by the user. Two-dimensional screens ensure that the molecule contains the correct functional groups and are advantageous as a preliminary step. Determination of upper and lower bounds for the distances accessible to the critical atom pairs are common strategies to reduce the amount of sampling required. Inclusion of torsional flexibility increases the number of hits that are retrieved in the search procedure, but with added computational cost.

Building Queries: Pharmacophore Generation and Validation. A query needs to be properly defined before carrying out a search. The type of query and the search strategy will depend on the knowledge that is available on the structure of the target. Indirect drug design techniques can be used to identify pharmacophores when no knowledge about the protein structure is available. Pharmacophores are the collection of relevant groups in the small molecule that can be responsible for the observed biological response. Beyond drug or agrochemical discovery, structure–property relationships, ie, in material science research, can be utilized.

Pharmacophore patterns could serve as queries to identify more molecules that satisfy them, and possibly the pharmacological activity that the pharmacophore summarizes. Their spatial arrangement is called a pharmacophoric pattern, whereas the position of complementary groups on the protein would be designated as a protein or pharmacophore map. A variety of techniques, from simple SAR to computation of quantum mechanical properties, can be used to define a pharmacophoric pattern (24,25). Once the pattern is found, the search of 3D databases can be undertaken, using software such as ISIS (32), UNITY (27), APEX-3D (25), or ALADDIN (12).

Excellent reviews in the area of pharmacophore design and validation exist (33,34). For the most part, automated pharmacophore generation techniques are in use, such as DISCO (35), Catalyst (25), MolMod (36), and FlexS (37) to name but a few. Pharmacophoric points are identified, but in addition, other features can be part of a pharmacophore, such as receptor points or excluded volumes.

The process of identifying a pharmacophore requires consistent biological information determined under identical conditions. The set should contain active and inactive compounds to serve as controls. If the compounds are flexible, a conformational study of each of them is also required. After the conformational libraries are available for each active compound, conformations are sought that maximize the similarities among the compounds of the same biological activity. Those features found common to the active compounds should be absent in the inactive compounds that are structurally related, which greatly reduces the total number of possibilities. From the study, one or a number of models could result that can be used for database searching. In the initial stage, the properties used to define pharmacophores are distances between functional groups within the molecules, which can be grossly classified as hydrophobic centers, hydrogen-bond donors, hydrogen-bond acceptors, charged centers, etc, and are referred to as pharmacophore elements. When similar spatial arrangements of pharmacophore elements are found in any of the low energy conformations of the active compounds that are absent in the inactive compounds, the structures of the active compounds can be overlaid providing a pharmacophore map. The conformation that provides such common spatial arrangement of the pharmacophore elements in each molecule is its bioactive conformation.

In some instances, the identification of pharmacophore points on the molecules themselves is not possible. While the functional groups in a series of different molecules may not be occupying the same relative portion within the molecule, it may be possible to orient them in such a way that they point to a putative external point in a similar manner. These points, external to the molecule, could also be part of a pharmacophore, and are commonly referred as receptor points. These points reflect that the pharmacophore elements are those that by indirect evidence appear to be interacting with the protein target.

Additional information about the active compounds could be incorporated as part of the pharmacophore, and can be used during the search in chemical databases. The most straightforward is to look for patterns in computed partition coefficients that may discriminate the active from the inactive compounds. In addition, other shape or electronic properties can serve to differentiate active from inactive compounds, and can be incorporated during the searches.

After a pharmacophore has been established, it is possible to carry out statistical studies to determine the relative importance of the different properties in the pharmacophores. Analysis such as CoMFA (38) or CoMSIA (39) or Hopfinger's molecular shape analysis (40) do provide information that could be used to rank order the hits that come out of the database. The hits can also be ranked based on their goodness of fit to the pharmacophore features (25).

Most of the work carried out to define pharmacophores has been done using small sets of compounds, in part, due to the need to ensure the homogeneity of the biological data imposed a constraint. Most automated pharmacophore determination software assumes that the compounds under study are binding or activating the target using a similar mechanism and site of action. The advent of high throughput screening as a dominant force in drug discovery has meant that much larger datasets are now available for analysis, but those sets are not necessarily homogeneous. The methods that were used to develop pharmacophores are not suitable to address the issues posed by large heterogeneous datasets. One solution has been the selection of homogeneous subsets for which detailed pharmacological information is possible. Methods that take into account the heterogeneous nature of the data as well as larger sample sets have been developed, using new means to carry out the analysis (41,42).

Docking and Target Structure-Based Database Searches. Whenever the structure of the target protein is known, it can be used in the process of searching chemical databases. The search is done by attempting to fit the small molecule into the known protein structure, commonly referred to as docking. In the first step, docking aims to predict how a small molecule can interact with a macromolecular target, and subsequently attempts to score how well the small molecule complements the binding site. The ultimate goal is to predict if and how a given small molecule can favorably interact with a protein.

A large number of different algorithms have been proposed for the automated docking of molecules (43,44), including DOCK (45,46), AutoDOCK (47,48), FLExX (27,49), and GOLD (26,50). While originally, the methodologies for docking were rigid-body matches, the most common implementations optimize the small molecule in the cavity of the site, and the way in which that is achieved constitutes one of the most significant differences among the different algorithms. A variety of optimization techniques including shape matching to genetic algorithms, evolutionary programming, or simulated annealing are used to that end.

While the above methods use complete molecular structures to carry out the searches, fragment-based methods are also common (51). Whole molecule methods are based on molecules that are part of the 3D chemical database, but fragment-based methods build new molecular structures in the site from substructures. Fragment-based methods can place a seed molecule in a cavity, and attach other groups in a stepwise manner, building up the desired structure. Another possibility is to place key functional groups complementing the features of the protein, and attempt to connect them into a single structure. LUDI is the most common among such build up procedures (52). The program connects fragments that dock into specific sites of a receptor, such as hydrogen-bond donating or accepting or hydrophobic residues. Fragments come from predefined libraries that can easily be customized. Once the fragments are positioned, they may be linked together using linear groups. If a seed fragment is placed in the binding site, the program can be used to add functionality that complements the site. The program may even perform a preliminary evaluation of synthetic accessibility of the linkage required, one of the major problems of drug design. Several other methods have been described.

Once the small molecule has been docked, the goodness of its fit should be determined. The prediction of binding affinity, or at least a correct rank order, is currently one of the most challenging problems in ligand design (53). The purpose is to prioritize the hits obtained from a computer program from a 3D search of structures in a database. This is another area where the different methods diverge, as several scoring methods are described (54,55). Some scoring methods such as free energy perturbation provide relative scores that are quite accurate but very computationally demanding, and therefore ill suited for virtual screening where hundreds of thousands or millions of compounds are typically computationally screened. Most scoring is done using force field calculations or empirical free energy scoring functions. Another possibility is the use of consensus scoring, where different scoring techniques are computed simultaneously, and a weighted average is taken as the parameter to rank order the goodness of binding (44).

For the most part, only limited efforts have been placed in dealing with the flexibility of the protein or the binding site. Some attempts have been made to take snapshots during a molecular dynamics trajectory, and to use the different structures, or averaged structures, for docking. The treatment of scoring functions and protein flexibility are major shortcomings at this time, which are attracting active research.

Examples of Applications and the Success of Virtual screening. High throughput screening methodologies in the pharmaceutical industry became so dominant due to a perceived lack of success in the area of structure-based drug design (56). Since then, the use of more sophisticated techniques facilitated their success, particularly when it was closely coupled to structural information (49,57,58).

Human immunodeficiency virus HIV protease and neuramidase inhibitors were derived from the use of computational tools and structural information. Various other enzyme inhibitors were also successfully designed by using a combination of structure-based and computer-aided drug design as well. Pharmacophore-based approaches resulted in the design of metalloprotease, tyrosine kinase inhibitors, and integrin receptor antagonists.

Nowadays, all the tools of computational chemistry, molecular design, and chemoinformatics are integrated into the process of designing new products. Soon it will be difficult to identify compounds largely derived based on those techniques, as they will be entwined with other discovery technologies. The acceptance of the tools of chemoinformatics and virtual screening is pervasive, and most discovery projects use them to the extent that is required.

3.4. Diversity Searches. The need to carry out diversity searches emerged in the pharmaceutical and agrochemical industry because of the advances in automation in biological screening and high throughput chemical synthesis. Those technologies posed a new set of questions to be asked from a chemical database. Initially, chemical collections for screening were obtained from internal libraries in the pharmaceutical industry. Those libraries had been created over the years by medicinal chemistry efforts, and consisted of large numbers of analogues on a few chemical families. Lack of success in the approach (5) was attributed to the relatively small size of the libraries studied, and to the lack of variety prosessed by those chemical collections. The need to detect redundancies in a chemical library led to the concept of diversity analysis.

With few exceptions (59), diversity analysis of chemical collections evolved from the methodologies of structure-based drug design. Consequently, diversity analysis has been heavily dependent on the computation of physicochemical or structural properties (8,60). Chemical diversity is mostly associated with methods that allow the determination of how well libraries could represent portions of chemical property space. The scattering or clumping of representatives in a library can be surrogate indicators of the probability that that library would provide multiple, singular, or no hits for a set of targets. The challenges associated with the determination of chemical diversity are quite varied. The most significant issues are the selection of properties and the algorithms that are to be used to determine and select diverse compound sets.

Properties for Diversity Analysis. Properties in use for diversity analysis include the use of computed physical properties for the compounds. A compound can be described by a collection of global properties, such as the octanol–water partition coefficient, its pK_a parameters related to molecular size and shape, counts of hydrogen-bond donor and acceptor centers, among others (61). These properties can be combined with topological descriptors, such as molecular connectivity indexes that encapsulate information about the 2D structure of the molecule, its structural complexity, and some simple measure of its electronic character (62,63). From them, shape and flexibility parameters can be generated.

While the structural keys were originally designed to make searches of chemical databases more efficient (8,61), they also play a significant role in the analysis of chemical diversity, and chemical similarity. Structural keys are discrete valued descriptors contrary to the global physicochemical properties, and therefore require different tools for their analysis. Some of those methods have been implemented in the Catalyst software (64), etc.

Three-dimensional properties can also be used and provide another view of the diversity of a compound collection. The structures necessary can be generated by the same methods indicated above for the creation of 3D databases. Three-dimensional properties for diversity analysis include the BCUT parameters (65). These parameters involve three types of matrices, where the diagonal elements are based on the atomic charges, polarizability, and hydrogen-bond donor or acceptor capabilities for each atom, respectively. The off-diagonal elements are based on 2D or 3D information, including functions of interatomic distance, overlaps, computed bond-orders, etc. All these parameters can be computed using semiempirical molecular orbital programs. The lowest and highest eigenvalues that result from the diagonalization of these three matrices are considered to reflect most aspects of the molecular structure. Methods must be developed for rationally deciding which BCUT values (eigenvalues) would be best for representing the chemical diversity of a given population of compounds. The analysis is part of the DiversitySolutions (27) software, whose efficiency has been reported in the literature. However, even at the 2D level, BCUT parameters can be satisfactory for diversity analysis.

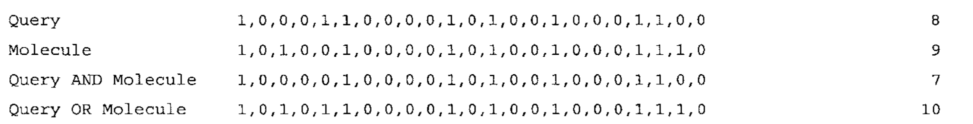
Counts on the possible spatial arrangement of chemical groups (66,67) are alternatively used to determine the pharmacophores accessible to a molecule. Pharmacophoric centers commonly associated with intermolecular interactions are typically included, such as hydrophobic centers, charged centers and hydrogen-bond donor and acceptor centers. Once the machine recognizes those groups, the distances between each of the centers are recorded. Distances among the pharmacophoric centers assume continuous values, but when variations in distance are small, they may be considered equivalent. For this reason, binning of pharmacophoric patterns is used by the major commercial software for this approach. In this case, the pharmacophore-based representation of a molecule is still a binary string that indicates the presence or absence of a certain combination of pharmacophoric centers at a certain range of distances.

Other 3D descriptors are based on the Comparative Molecular Field Analysis (CoMFA), which is particularly useful for series of related compounds (8). The scores from docking compounds against a set of random proteins have also been utilized to check the diversity of compounds (68). The set of scores determined for each protein constitutes a descriptor. The approach was inspired in a previous attempt to utilize experimental binding data for diversity assessment (59,69).

Algorithms for Diversity Analysis (70). Regardless of the set of properties considered, molecules are represented by vectors either of continuous values, or as bitstrings in the case of structural keys or pharmacophore analysis. Hence, each molecule in the set is represented by a vector, or as a point in a high-dimensional space, or a "chemical space". The similarity between two molecules can be measured calculating a distance between the two points that represent each molecule. If the properties are binary, then the distance can be computed using the Tanimoto coefficient or the XOR distance, if the properties are continuous, an Euclidean distance can be used.

The distances generated for a set of molecules can be used as a similarity matrix, when the distance between every pair of molecules is determined. A similarity matrix can then be used to select a subset of compounds that are as diverse as possible for the set of properties under consideration. However, the relation between compounds will critically depend on the chemical representation adopted because compounds that appear to be different for a set of properties can be very much alike under a different set of descriptors (71). The selection of a property space is therefore an important issue when analyzing for diversity and requires careful consideration. The selection of properties is mostly based on the ability of the set chosen to segregate compounds of different pharmacological profile (60).

Clustering techniques provide the means to group compounds in sets that have similar properties (72). There are different types of clustering algorithms, but one of the most commonly employed in many arenas and also in chemoinformatics work, is the hierarchical agglomerative. Step after step the method clusters successively more distant compounds. In the first step, the two most similar compounds are grouped together forming a cluster. The next set of closely related points or clusters are linked together, and the procedure continues until all points are part of a single cluster. The representation of the clustering process is a dendrogram (a classification tree) that goes from the individual points or singletons to all compounds in a single cluster, a representation that is common in other disciplines. An example is shown in Figure 4. The number of clusters is



	Bitscreen	Bits Set
Query	1,0,0,0,1,1,0,0,0,0,1,0,1,0,0,1,0,0,0,1,1,0,0	8
Molecule	1,0,1,0,0,1,0,0,0,0,1,0,1,0,0,1,0,0,0,1,1,1,0	9
Query AND Molecule	1,0,0,0,0,1,0,0,0,0,1,0,1,0,0,1,0,0,0,1,1,0,0	7
Query OR Molecule	1,0,1,0,1,1,0,0,0,0,1,0,1,0,0,1,0,0,0,1,1,1,0	10

Fig. 4. Dendrogram: A table of properties (P1 to P7) for a series of compounds (A–Q). When a cutoff for similarity (height) of 2 is, eg, selected, compounds A, E, and M belong to a cluster, while G, J, H form a different one. Some compounds have clusters of their own, denoted as singletons. Compounds B, C, and D are examples of that category.

determined by the degree of similarity that is considered significant. Once a set of clusters is defined, a compound can be selected from each cluster, resulting in a diversified set.

A clustering strategy that has been widely used in chemoinformatics is the Jarvis Patrick algorithm (73). In simple terms, the algorithm creates a list of its nearest neighbors for each point. Two points belong to the same cluster if they are in each other's list of nearest neighbors, and they share a number of common neighbors. Multiple technical reasons make it a preferred choice for the problem of compound selection. Beyond their commonly reported speed and efficiency, it can automatically deal with nonspherical clusters efficiently. Hierarchical methods can also deal with nonspherical clusters, but prior knowledge about the distribution is required.

Nonclustering methods are also used for diversity selection. Cell-based methods are part of the DiverseSolutions software (27), where the space defined by the chemical properties selected is partitioned into cells. The occupancy of each cell is determined based on the properties of the compounds (74,75). The advantage of cell-based techniques is that they provide a uniform sampling and the areas of property space that are not represented in the library can readily be identified, providing a simple representation of the completeness of the chemical library. D-optimal design is also a technique that was applied to this problem (61). However, its use has been less significant because of its tendency to select compounds unequally from the entire chemical space, and show a bias for points at the edge distribution.

Software such as C2-Diversity (Accelrys, San Diego) provides a broad assortment of properties and a variety of methods for selecting such diversity (64). Pipeline Pilot offers an alternative of innovative software architecture that computes processing, analysis, and mining of large volumes of data through a user-defined computational protocol (76).

Contrary to similarity comparisons, where success is measurable by the number of compounds that share the desired profile possible from among those chosen using the metric, the goodness of diversity algorithms are harder to characterize. If the goal is simply to remove redundancy from a chemical library, even the simplest of algorithms can fit the requirement. If, on the other hand, the goal is to increase the hit rate for a library of related targets, or for any random set of targets, the best strategy to utilize could be different. The overlap and similarity of the software currently available for library design can be a major challenge, since most different packages are fragmented. The fragmentation is the result of two trends. On one hand, there is the acquisition of software from academic sources or by formal mergers and acquisitions that the major chemoinformatics software vendors have undergone. On the other hand, there is the commercial tendency to fragment the software, to customize to specific needs, which in practice results in redundancy as frequently more than one package is required.

Design of Chemical Libraries. Few, if any, hits from massive and truly random screening libraries could be evolved into starting points for product development. Additional constrains have been imposed to the value ranges of the properties being considered, which are consistent with their intended use, effectively limiting the chemical space taken into account. Lists of exclusions are common when the compounds are intended for screening where reactive groups are removed from the searches. The same types of properties have also been used to tailor a library for drug discovery, in an attempt to look for drugs in an area of chemical space that is relevant to medicinal chemistry (77,78). For the most part, these efforts have been focused around the structural characteristics of drugs (79–81) and the design of libraries for screening has now centered on the use of compounds similar to drugs, or "drug-like" molecules.

The nature of the constraints to be imposed on chemical space was derived by the statistical analysis of the properties of marketed drugs or of compounds that have undergone human testing. The types of chemical functionalities found in them, and their physicochemical characteristics, serve to define the acceptable range of properties that are characteristic of drugs. The argument has been that the properties of compounds that have been in late stage trials reflect what is biologically compatible. The study of successful compounds has been part of the attempt to predict their absorption, metabolism, distribution, excretion (ADME), as well as their toxicology. Prediction of those properties is currently receiving significant attention in the chemoinformatics field since it represents a major bottleneck in drug discovery and development processes.

Within this realm, the "rule of 5" (82) has gained acceptance (83). These rule was developed by a simple analysis of databases of compounds that had undergone clinical trials. It was concluded that poor permeation or absorption were more common when: there were >5 hydrogen bond donors; >10 hydrogen-bond acceptors; the calculated Log P was >5, and the molecular weight was >500. The cutoffs for each of four parameters are multiples of five. Thus its name. The rule does not cover compounds that are actively transported. This simple chemometric exercise has affected the design of ligands for target proteins. However, the approach is not without its criticisms (84,85).

Statistical rules to predict solubility, oral availability, and permeability are part of the repertoire of chemoinformatics tools [86–88]. Those properties can be quickly computed with packages such as QikProp (89), or derived based on global molecular properties, such as those provided by the ACD Labs package (90) and that are straightforward to implement. Many other relations are commonly used to predict blood-brain barrier permeation, cell permeability, or to estimate stability and pharmacokinetic parameters.

Important efforts have been made in the area of using experimental information to predict oral absorption. The iDEA (In Vitro Determination for the Estimation of ADME) simulation system is a computational model developed to predict human oral drug absorption based on its solubility and permeability, which is empirically determined (91).

The determination of metabolism is an extremely complex issue, where the different isozymes of cytochrome P450 play an all-encompassing role. Prediction of the sites of metabolism (regioselectivity) for this enzyme can be done by evaluating the electronic tendencies for oxidation of all the potential sites within the substrate molecule (92–95). However, this is only a preliminary approach, and complex simulations are still required to carry out metabolism prediction accurately. Biological processes are quite complex and cannot be simulated entirely in-silico. The real value in chemoinformatics resides in the derivation of simple rules. Whenever those rules are possible, they result in biases in compound selection toward candidate compounds that are more likely to succeed.

3.5. Toxicology Prediction.

The prediction of the toxicological character of a compound is of foremost importance throughout the chemical industry, and extends from the agrochemical and pharmaceutical industry to the environmental and food chemistry. A large number of approaches have been adopted and are employed to study the compound toxicity problem (96).

One of the most popular packages for that purposes is TOPKAT (64,97) which is a self-contained computational toxicology package that uses 2D descriptors and statistical models to generate reliable toxicological profiles of organic chemicals, one at a time.

An alternative to the statistical analysis of properties are knowledge-based systems that are computer programs to organize relevant experimental data to help a user make decisions about concrete issues. They require the use of a systematic database of information from which rules are derived, which allows the prediction of the property to be scrutinized. HazardExpert predicts different toxicity effects of compounds such as carcinogenic, mutagenic, teratogenic, membrane irritation, and neurotoxic effects (93). The knowledge base was developed based on the list of toxic fragments reported by more than 20 lead experts. This software also predicts bioaccumulation as well as bioavailability based upon predicted physicochemical values. It is a rule-based system using known toxic fragments collected from in vivo experimentation. DEREK is also a rule-based approach that can make predictions about a large set of toxicological properties including carcinogenicity, irritancy, lachrymation, neurotoxicity and thyroid toxicity, teratogenicity, respiratory and skin sensitization, and mutagenicity (98).

MULTICASE and CASE programs (99,100) can automatically identify molecular substructures that have a high probability of being relevant or responsible for the observed biological activity of a learning set comprised of a mix of active and inactive molecules of diverse composition. New, untested molecules can then be submitted to the program, and an expert prediction of the potential activity of the new molecule is obtained.

4. Chemical Databases

Chemical information itself is abundant, and the Chemical Abstract Service (CAS) has maintained the most comprehensive resource, in this field (101). SciFinder provides a desktop research tool that allows the exploration of research topics, with little training, containing information on >33 million substances. The STN service provides specialized information on >200 different subjects.

CrossFire Beilstein (32) has extensive information on bioactivity and physical properties that makes it particularly useful when undertaking biological research. The database also provides information on the ecological fate of compounds.

Other databases are also worth noting because they contain information that is more specific. There are vast numbers of commercial databases with different focus. One of the most common is the ACD (Available Chemicals Directory) (32), which contains information on price and availability on >300,000 compounds. This information includes not only a 2D representation of the molecule, but also 3D models that make it useable for pharmacophore searches or for docking purposes. Many chemical vendors also provide catalogs of rare chemicals in 2D format, but that can be readily converted to a 3D database using the methods described previously.

Chemical databases also deal with reactivity information. As described above, reaction databases have special characteristics that arise from the need to handle reactants and products. Multiple databases exist for this purpose. SpresiReact database is available from InfoChem GmbH and contains 2.5 million different reactions (12,32). It contains 1.8 million individual molecules that appear as components of the reactions, journal references, yields, and reaction conditions information. RefLib (32) is another broad collection of novel organic synthetic methodologies that covers functional group transformations, metal-mediated chemistry, and asymmetric syntheses, as well as reactions from Theilheimer's *Synthetic Methods of Organic Chemistry*. An electronic version of the entire series of *Organic Syntheses*, ORGSYN, contains general synthetic methods and proven compound preparations. Similarly, Derwent's *Journal of Synthetic Methods*, has been condensed in the RX-JSM database. *Methods in Organic Synthesis* (MOS) is a selective current awareness database derived from a bulletin of the same name, published by the Royal Society of Chemistry. This database focuses on important new methods in organic synthesis and comprises >3300 reactions per year, dating back to 1991. BioCatalysis is a selective, thematic database that focuses on chemical synthesis using biocatalysts, including pure enzymes, whole cells, catalytic antibodies, and enzyme analogues. The Failed Reactions database (63) is a unique compilation of reactions with unexpected results, which may involve an unexpected product, an immediate further reaction, or simply no reaction.

Combinatorial chemistry has played a significant role in the making of compounds for material sciences as well as for the agrochemical and pharmaceutical industries. The SPORE (32) and Solid-Phase Synthesis (63) databases include data particular to solid-phase organic synthesis, such as information on polymeric materials, linkers, solid supports, and protecting groups. The Protecting Groups database (32) provides information on methods for protection, deprotection with the ability to search generically, based on functional groups, protected groups, tolerated groups, and reaction conditions. Bunin's book *The Combinatorial Index* has also been put into electronic form. Other databases on chemical reactivity also exist

Apart from chemistry resources, there are a large number of content databases with information specific about different areas (102,103) all linked by their chemical structure, which include material sciences, agrochemical, physicochemical, and biological activity. These databases are provided by a variety of solution providers, offering different products, which may complicate the search for an appropriate one.

5. Data Analysis and Presentation

Data mining is crucial when large amounts of data are generated and is a trend observed in many industries, including the pharmaceutical industry (104). The idea is to integrate a number of visualization, statistical analysis tools through graphical user interfaces.

One of the most popular packages for this purpose is Spotfire (105), where chemical structure data can be combined with data from different sources, in order to provide insights into the property of interest, biological or physical. These software tools allow the users to manipulate variables and large quantities of data, and integrate them with simple statistical tools such as graphs, decision trees, and scatter plots. The software allows the user to merge data from diverse sources into a single screen with the ability to visualize trends. The software is in use in the pharmaceutical, but also in the energy, specialty chemical, and semiconductors industries.

DIVA (63) is a package with a similar purpose that was developed specifically for the pharmaceutical industry. DIVA allows users to retrieve and work with chemical structures, assay results, and other chemical and biological data in one convenient spreadsheet. Powerful easy-to-use tools for data integration, visualization, analysis, and reporting save time and allow researchers to get more value from their data.

6. Economics of Chemoinformatics

The economic impact of chemoinformatics is two sided, as it is an industry that produces software, but it also greatly affects the productivity of all chemistry research and development. On one hand, during the year 2000, the overall market for chemoinformatics, bioinformatics, and simulation software was ~1.3 billion. The number is in circulation and is based on assuming a spending of ~7% of the R&D budget of $15 billion in informatics services for the composite of pharmaceutical, specialty chemical, and agrochemical markets. About 90% of that amount is spent in-house, giving an estimate of ~$150 million for third parties. The numbers are poised for significant growth on a yearly basis.

However, the market is remarkably fragmented. The reasons for the fragmentation involve the nature of the business, where technological innovation is key and there are low barriers to establish new ventures. New players with an interesting application can create a niche from which they can grow. Thus, the established vendors face competition from nonprofit organizations, in-house solutions, and other technology providers, such as IBM, SGI, or Agilent.

No publicly traded company can be labeled truly chemoinformatics pure play. In many cases, the company has other business associations or forms part of a major conglomerate that makes the analysis more complex. Two companies that have a strong chemoinformatics component, Tripos and Pharmacopeia, also have associated molecular modeling software and chemistry research. Another equally important player in the area is MDL, Inc., a subsidiary of the major publishing conglomerate Elsevier.

Pharmacopeia's software revenue for the 2001 third quarter rose 20% compared to its 2000 third quarter of $21.5 million, which included the effects of acquisitions. For the 9-month year-to date period, Tripos recorded $32.5 million in revenues compared to $16.2 million in 2000, an increase of 101%. However, this number is an aggregate of financial transactions and other nonsoftware business. The results for both companies reflect the growth rate of the industry, which is picking up pace. As of the end of the third quarter 2001, the market capitalization for Tripos was 127.7 million versus 334.4 million for Pharmacopeia.

Two revenue models exist in the industry. On one hand, there are the companies that simply sell software and services into the drug discovery market, and on the other hand, there are companies with research collaborations with major pharmaceutical companies, with upside potential if royalties are retained. For software providers, the preferred model is that of software licensing and maintenance fees. In some cases, and due to the steep licensing fees, yearly license payments have been adopted. However, this limited the perceived value of those companies and imposed restrictive market caps. In an effort to improve their valuation, companies moved to provide other services as well. Through mergers and acquisitions during the late 1990s, Pharmacopeia, a chemistry services provider for the pharmaceutical industry acquired MSI, Inc., while Tripos, a software provider, acquired Receptor Research a small chemistry services provider based in the United Kingdom.

Consolidation is not new to the industry. Accelrys, the software division of Pharmacopeia, is the result of a number of mergers and acquisitions, the most recent being that of the Oxford Molecular Group, based in the United Kingdom in 2000 for $22 million. The acquisition of Trega by Lion Biosciences has resulted in a different model where there is forward integration and where a company in the area of genomics software acquired a provider of tools and content.

The developments in genomics and proteomics are likely to produce a new wave of software tools that more closely integrate the tools of bioinformatics and chemoinformatics. Structural proteomics is also giving rise to a new crop of companies that aim at structural chemistry and drug discovery, many with in-silico components. This is an arena where the landscape is rapidly changing, with a yearly growth of up to 40%.

From a different angle, it is different to gauge the importance and productivity gains due to chemoinformatics. However, robotics and automation, as well as multiparametric analysis, would not be possible without the intensive use of computers in chemistry.

BIBLIOGRAPHY

1. F. K. Brown, *Annu. Rep. Med. Chem.* **33**, 375, (1998).
2. J. Drews, *Science* **289**, 1960 (2000).
3. L. M. Kauvar and E. Laborde *Curr. Op. Drug Disc. Divers* **1**, 66 (1998).
4. J. Bajorath, *Drug Discov. Today* **6**, 989 (2001).
5. R. Lahana, *Drug Discov. Today* **4**, 447 (1999).
6. L. A. Thompson and J. A. Ellman, *Chem. Rev.* **96**, 555 (1996).
7. J. C. Hogan, Jr., *Nature Biotech.* **15**, 328 (1996).
8. Y. C. Martin, R. D. Brown, and M. G. Bures, in E. M. Gordon and J. F. Kerwin, Jr., eds, *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*, Wiley-Liss, 1998. p. 369.
9. J. M. Barnard, in: P. von R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer, and P. R. Shreiner, eds, *Encyclopedia of Computational Chemistry*, Wiley, Chichester, U.K. Vol 4, (1998), p. 2818.

10. D. Weininger, *J. Chem. Inf. Comput. Sci.* **28**, 31 (1998).
11. C. A. James, D. Weininger, and J. Delany, Daylight Theory Manual, Daylight Chemical Information Systems, Inc. Santa Fe, N. Mex.
12. http://www.daylight.com.
13. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer, *J. Chem. Inf. Chem. Sci.* **32**, 244 (1992).
14. http://www.mdli.com/downloads/ctfile/ctfile_subs.html.
15. G. M. Downs and P. Willett, *Rev. Comput. Chem.* **7**, 1 (1996).
16. ISIS/Base 2.1.4; MDL Information Systems, Inc., San Leandro, Calif.
17. http://www.cambridgesoft.com/products.
18. P. Willett and V. A. Winterman, *Quant. Struct-Act. Relat.* **5**, 18 (1986).
19. S. L. Dixon and R. T. Koehler, *J. Med. Chem.* **42**, 2887 (1999).
20. P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman & Co. San Francisco, Calif. 1973.
21. A. C. Good and J. S. Mason, *Rev. Comput. Chem.* **7**, 67 (1996).
22. P. J. Gane and P. M. Dean, *Curr. Opin. Struct. Biol.* **10**, 401 (2000).
23. D. A. Gschwend, A. C. Good, and I. D. Kuntz, *J. Mol. Recog.* **2**, 175 (1996).
24. G. H. Loew, H. O. Villar, and I. Alkorta. *Pharm. Res.* **10**, 475 (1993).
25. Y. Kurogi and O. F. Guner, *Curr. Med. Chem.* **8**, 1035 (2001).
26. http://www.ccdc.cam.ac.uk.
27. http://www.tripos.com/software.
28. J. Sadowski and J. Gasteiger, *Chem. Rev.* **93**, 2567 (1993).
29. http://cmm.info.nih.gov/modeling/guide_documents/conformation_document.html.
30. A. Smellie, S. L. Teig, and P. Towbin, *J. Comp. Chem.* **16**, 171 (1995).
31. T. Hurst, *J. Chem. Inf. Comput. Sci.* **34**, 190 (1994).
32. http://www.mdli.com.
33. A. K. Ghose and J. J. Wendolowski, *Persp. Drug Discov. Design* **9**, 253 (1998).
34. G. Klebe, in H. Kubinyi, ed, *3D QSAR in drug design: theory methods and applications*, Escom, Leiden, 1993, p. 173.
35. Y. C. Martin, M. G. Bures, E. A. Dandur, J. De Lazier, I. Lico, and P. A. Pavlik, *J. Comput. Aided Drug Des.* **7**, 83 (1993).
36. D. H. Harris and G. H. Loew, *Bioorg. Med. Chem.* **8**, 2527 (2000).
37. C. Lemmen, T. Lengauer, and G. Klebe, *J. Med. Chem.* **41**, 4502 (1998).
38. R. D. Cramer, D. E. Patterson, and J. D. Bunce, *J. Am. Chem. Soc.* **110**, 5959 (1988).
39. G. Klebe and U. Abraham, *J. Comput. Aided Drug Des.* **13**, 1 (1999).
40. A. J. Hopfinger, S. Wang, S. Tokarshi, B. Jin, M. Albuquerque, P. Madhav, and C. Duraiswami, *J. Am. Chem. Soc.,* **119**, 10509 (1997).
41. A. J. Hopfinger and J. S. Duca, *Curr. Op. Biotechnol.* **11**, 97 (2000).
42. X. Chen, A. Rusinko III, A. Tropsha, and S. S. Young, *J. Chem. Inf. Comput. Sci.* **39**, 887 (1997).
43. R. E. Babine and S. L. Bender, *Chem. Rev.* **97**, 1359 (1997).
44. C. Bissantz, G. Folkers and D. Rognan, *J. Med. Chem.* **42**, 4759 (2000).
45. I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge and T. E. Ferrin, *J.Mol. Biol.* **161**, 269 (1982).
46. http://www.cmpharm.ucsf.edu/kuntz/dock.html.
47. D. S. Goodsell and A. J. Olson, *Proteins* **8**, 195 (1990).
48. http://www.scripps.edu/pub/olson-web/doc/autodock.
49. M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, *J. Mol. Biol.* **261**, 470 (1996).
50. G. Jones, P. Wilett, R. C. Glen, A. R. Leach, and R. Taylor, *J. Mol. Biol.* **267**, 727 (1999).
51. H. J. Bohm, *Persp. Drug Discov. Design* **3**, 21 (1995).
52. G. Schneider, O. Clement-Chomiene, L. Hilfinger, P. Schneider, S. Kirsch, H. J. Bohm, and W. Neidhart, *Angew. Chem. Int. Ed. Engl.* **39**, 4130 (2000).

53. I. Muegge and M. Rarey, *Rev. Comput. Chem.* **17**, 1 (2001).
54. T. Hansson, J. Marelius, and J. Aqvist, *J. Comput. Aided Mol. Des.* **12**, 27 (1998).
55. M. Schapira, M. Trotov, and R. Abagyan, *J. Mol. Recog.* **12**, 177 (1999).
56. R. M. Snider, *Science* **251**, 435 (1991).
57. H. Kubinyi, *J. Recept. Signal Transduct. Res.* **19**, 15 (1999).
58. H. Kubinyi, *Curr. Op. Drug Disc. Dev.* **1**, 4 (1998).
59. S. L. Dixon and H. O. Villar, *J. Chem. Inf. Comput. Sci.* **38**, 1192 (1998).
60. R. D. Brown and Y. C. Martin *J. Chem. Inf. Comput. Sci.*, **37**, 1 (1997).
61. E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong, and W. H. Moos, *J. Med. Chem.* **38**, 1431 (1995).
62. L. H. Hall and L. B. Kier, *Rev. Comput. Chem.*, **2**, 367 (1991).
63. R. A. Lewis, J. S. Mansonand and I. M. Mc Lay, *J. Chem. Inf. Comput. Sci.* **37**, 599 (1997).
64. http://www.accelrys.com.
65. R. S. Pearlman and K. M. Smith, *J. Chem. Inf. Comput. Sci.* **39**, 28 (1999).
66. H. Matter and T. Potter, *J. Chem. Inf. Comput. Sci.* **39**, 1211 (1999).
67. J. S. Manson, I. Morize, P. R. Menard, D. L. Cheney, C. Hulme, and R. F. Labaudiniere, *J. Med. Chem.* **42**, 3251 (1999).
68. H. Briem and I. D. Kuntz, *J. Med. Chem.* **39**, 3401 (1996).
69. J. N. Weinstein, T. G. Myers, P. M. O'Connor, S. H. Friend, A. J. Fornace Jr., K. W. Kohn, T. Fojo, S. E. Bates, L. V. Rubinstein, N. L. Anderson, J. K. Buolamwini, W. W. van Osdol, A. P. Monks, D. A. Scudiero, E. A. Sausville, D. W. Zaharevitz, B. Bunow, V. N. Viswanadhan, G. S. Johnson, R. E. Wittes, and K. D. Paull, *Science* **275**, 343 (1997).
70. D. C. Spellmeyer and P. D. J. Grootenhius, *Annu. Rep. Med. Chem.* **34**, 287 (1999).
71. H. O. Villar and R. T. Koehler, *Mol. Div.* **5**, 13 (2000).
72. P. Wilett, *Pers. Drug Disc. Design* **7–8**, 1 (1997).
73. R. A. Jarvis, E. A. Patrick, *IEEE Trans. Comput.*, *C-22*, 1025 (1973).
74. D. Schnur, *J. Chem. Inf. Comp. Sci.* **39**, 36 (1999).
75. P. R. Menard, J. S. Mason, I. Morize, and S. Bauerschmidt, *J. Chem. Inf. Comp. Sci.* **38**, 1204 (1998).
76. http://www.scitegic.com.
77. E. J. Martin and R. E. Critchlow, *J. Comb. Chem.* **1**, 32 (1999).
78. A. K. Ghose, A. K. Viswanadhan, and J. J. Wendolowski, *J. Comb. Chem.* **1**, 55 (1999).
79. W. P. Ajay Walters, and M. A. Murcko, *J. Med. Chem.* **41**, 3314 (1998).
80. J. Sadowski and H. A. Kubinyi, *J. Med. Chem.* **41**, 3325 (1998).
81. R. T. Koehler, S. L. Dixon, and H. O. Villar, *J. Med. Chem.* **42**, 4695 (1999).
82. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeny, *Adv. Drug Delivery Res.*, **23**, 3 (1997).
83. R. A. Lipper, *Modern Drug Disc.* **2**, 55 (1999).
84. T. I. Oprea and J. Gottfries, *J. Mol. Graph. Model.* **17**, 26 (1999).
85. T. I. Oprea, *J. Comput-Aided Mol. Design* **14**, 251 (2000).
86. F. Yoshida and J. G. Topliss, *J. Med. Chem.* **43**, 2575 (2000).
87. P. Sternberg, K. Luthman, and P. Artursson, *J. Control. Release* **65**, 231 (2000).
88. D. E. Clark, *Comb. Chem. High Throughput Screen.* **4**, 477 (2001).
89. http://www.schroedinger.com.
90. http://www.acdlabs.com.
91. http://www.lionbioscience.com/solutions/idea.
92. http://www.camitro.com.
93. http://www.compudrug.com.
94. G. H. Loew and D. L. Harris, *Chem. Rev.* **100**, 407 (2000).
95. P. R. Chaturvedi, C. J. Decker, and A. Odinecs, *Curr. Op. Chem. Biol.* **5**, 452 (2001).

96. D. F. V. Lewis, **3**, 173 (1993).
97. K. Enslein, V. K. Gombar, and B. J. Black, *Mutat. Res.* **305**, 47 (1994).
98. http://lhasa.harvard.edu.
99. G. Klopman, *Quant. Struct.-Act. Rel.* **11**, 176 (1992).
100. http://www.multicase.com.
101. http://www.cas.org.
102. http://www.scivision.com.
103. http://www.pharmacopeia.com/corp/IR/inv_day_2001/Accelrys.pdf.
104. R. Wedin, *Mod. Drug Disc.* **2**, 39 (1999).
105. http://www.spotfire.com.

Hugo O. Villar
Triad Therapeutics