Large Compound Databases for Structure-Activity Relationships Studies in Drug Discovery

Thomas Scior[1,*], Philippe Bernard[2], José Luis Medina-Franco[3] and Gerald. M. Maggiora[3,4]

[1]*Departamento de Farmacia, Facultad de Ciencias Químicas, Benemérita Universidad Autónoma de Puebla, Puebla, Pue, México;* [2]*GreenPharma S.A.; 3, Allée du titane, F-45100 Orléans, France;* [3]*BIO5 Institute, University of Arizona, Tucson 85721, AZ, USA;* [4]*College of Pharmacy, University of Arizona, Tucson 85721, AZ, USA*

Abstract: Large libraries of chemical compounds reflect the exponentially growing data-enrichment in drug discovery that trends towards fully automated informatics solutions to study structure - activity relationships by screening docked ligand candidates to biological target structures. We review otherwise disseminated user descriptions of mainly public databases with free access and also our integrated data mining tool *GPDBnet* for phyto-pharmacology.

Key Words: Chemical libraries, data mining, focused library, computational screening, privileged structures, scaffold hopping, ligand – target docking, GPDBnet.

1. INTRODUCTION

In this paper, we review large databases, covering a wide range of therapeutic interests, as valuable sources of structure-activity relationships studies making emphasis on those that are freely accessible. Recent advances in computational tools that have been developed to identify structural leads and activity patterns in large compound collections are also discussed.

The massive advent of combinatorial and parallel synthesis as well as *in vitro* high-throughput screening (HTS) and the access to web-based bioinformatic tools has given rise to a new data-rich environment for the life-sciences dealing with biomolecular targets (DNA, RNA, proteins, enzymes, receptors) and molecular ligands (substrates, inhibitors, agonists) alike. In direct response, a plethora of chemical and pharmacological records are electronically edited, stored, linked and organized for expert retrieval. The key goal becomes evident: by shortening the trial and error cycles in the drug discovery and development process costs are lowered [1]. In early stages of drug research computational studies of structure-activity relationships (SAR) on compound collections are being used to accelerate the identification of promising new candidates with innovative mechanisms of action [2]. Precisely, among the key goals in research and development (R&D) projects are the introduction of fully automated virtual screening (VS) in a knowledge-based manner. In this way, chemical databases have evolved from being just repositories of compiled compounds to being active tools in drug discovery. Recommended reading in this area is M. A. Miller's review about the role of chemical databases in drug design [3].

In pharmaceutical companies certain databases are kept "in house". Other databases, started as commercial ventures, are consulted mostly by pharmaceutical companies which can afford the user license costs. Alternatively, non-commercial projects such as the NIH Molecular Libraries Initiative [4] give the scientific community free access to the records of small molecules. Particularly, the electronic access to data of public sources such as patents, scientific journals and conferences has spurred the creation of annotated chemical libraries, i.e. added expert comments (if available) to the numerical, graphical or text records [5]. The vast amount of information contained in such compound collections can lead to new computational strategies that enhance the reliability of computed SAR- and QSAR- predictions for valid compound selection (Table **1**).

Table 1. List of Goals Achieved by Fully Automated *In Silico* High Throughput Screening Also Called Virtual Screening of Large Compound Databases

Improve efficiency of drug discovery
Describe computationally drug-like molecular properties
Identify promising target candidates (biomolecules, proteins, enzymes, receptors)
Identify promising ligand candidates (small organic synthetic or biosynthetic compounds)
Conduct lead generation
Perform lead optimization
Find innovative new pathophysiological pathways for patents
Reduce the costs of R&D
Revise older *in silico* approaches to drug discovery when probably promising candidates failed to pass the tests due to incorrect assumptions and computation
Revise older *in vivo* and vitro assays to unravel false compound selections
Improve drug potency and specificity (pharmacodynamics)
Predict ADME characteristics of potential drug candidates (pharmacokinetics)
Estimate potential toxicity (ADMET)

*Address correspondence to this author at Departamento de Farmacia, Facultad de Ciencias Químicas, Benemérita Universidad Autónoma de Puebla, Puebla, Pue, México; E-mail: tscior@siu.buap.mx

1389-5575/07 $50.00+.00

However the overwhelming number of molecular descriptors that are frequently used in QSAR studies [6] remains difficult to interpret by medicinal chemists and makes it difficult to guide the next generation candidates for synthesis and testing [7]. Useful experimental and computed molecular descriptors to assess druggability or drug-like profiles are summarized in Table **2**.

Table 2. List of Computed or Experimental Molecular Descriptors Commonly Used During *In Silico* Screening Studies. The Values Indicate Established Thresholds for Drug-Likeliness

Number of hydrogen-bond donors and acceptors
Aqueous solubility, logS > -4 ug/ mL
Octanol/water partition coefficients, logP, logD
Dissociation, ionization, pKa
Calculated topological parameters: polar surface area, PSA < 140 $Å^2$; scaffold and side chain flexibility for conformers
Lipinski's Rule of Five: molecular weight ≤ 500, number of hydrogen-bond donors ≤ 5, Log P ≤ 5, sum of oxygen and nitrogen atoms ≤ 10
Experimental pharmacokinetic paramenters (ADMET): volume of distribution, bioavailability, permeation tests
Permeability, like *in vitro* cell permeation, Caco2 cell culture, etc.
Blood/brain partitioning, logBB
Chemical stability, reactive groups, reactive atom or hetero-atom counts
Predictions based on annotated fragment libraries for reactive or toxic groups, i.e. certain chemical groups associate to certain intrinsic toxicological effects, such as carcinogenicity
In vitro activity tests, like log IC_{50}, ion channel blockage, etc.

Historically, since the seminal work of Cramer to conduct substructural analysis [8] several computer programs have been developed to this end: they provide new approaches to handle predefined substructure fragments. An early but typical example of parsing a chemical library through general substructure fragments is the work of Bemis and Murcko who analyzed the Comprehensive Medicinal Chemistry Database [9] in terms of ring systems, linker atoms, side chain atoms or scaffolds [10]. Thus, new computer tools have been emerging that are helpful to derive SAR studies defining structural features like rings, side chains, functional groups or pharmacophoric patterns. For instance, SLASH, HookSpace, RECAP or Stigmata are reviewed in [11].

Theoretical outcomes of SAR-studies exploiting compound libraries by aforementioned parser options include: (i) Identification of new classes of active chemotypes (e.g., scaffold hopping [12,13]); (ii) identification of privileged substructures [14,15]; (iii) chemotype-based hierarchical analysis of the distribution of actives associated with biological screening [16]; as well as (iv) exploration of chemotypes which are strongly associated with inactivity (e.g. database shaving [17]). Loading known pharmacologically active agents with annotations about their experimentally identified or computed targets/receptors creates a more or less reliable reference dataset for similarity queries of still undetected compounds in a far larger dataset. Hence annotation cross-linking bio- and cheminformatics forms a prerequisite for the design of ligand or target type focused compound libraries in search of potential drug candidates. In the opposite direction work *in silico* approaches exploiting sets of known pharmacologically active agents to identify novel disease-related biomolecular targets, in analogy to *reverse pharmacology* (see also discussion).

The next segment will highlight large databases - mainly public ones - with annotated pharmacological activity that can be used to conduct SAR-studies dealing with various therapeutic applications. In the following part, we focus on recent developments of computational programs, some of them freely available, in order to parse compound collections and to perform clustering studies. This is followed by a discussion of recent studies to parse compound collections. Related services on the Internet to generate the necessary molecular parameters were reported in two reviews [6,18].

2. DATABASES

2.1. Annotated Compound Collections

Table **3** summarizes the names, developers, and web sites of large, annotated collections of compounds that are freely available. A description of each database with corresponding references is given in the following:

PubChem

It is a public database of chemical structures and their corresponding activities accessed through the National Library of Medicine [4]. The system links the chemical structure records with text of biomedical literature, a protein structure database and to the depositor web sites. It also links small-molecule information to the PubMed Entrez databases [19]. Retrievable items are chemical structures and names, bio-assay descriptions with activities. Another noteworthy feature of PubChem is its tool for fast structural similarity searches.

National Cancer Institute Databases

It contains a remarkable dataset of more than 250,000 molecules. Around 40,000 structures alone deal with anticancer and anti-HIV activities gathered during the NCI's Developmental Therapeutic programs. This dataset has been reviewed elsewhere and is said to be one of the most popular data sets to perform SAR-studies and to test new data-mining approaches [18].

ChemBank

This compound collection belongs to the Initiative for Chemical Genetics (ICG) of the National Cancer Institute [20]. The web-based database stores information on small molecules that have been tested in biological assays in cell cultures or whole model organisms. ChemBank also contains several visualization tools to assist navigation through chemical and biological space [21,22]. ICG's molecules can be filtered either by their names, structures or similarities defined through numerical descriptors, like molecular

Table 3. Publicly Available Compounds Databases Annotated with Biological Activity

Database / Developer Site or Support	Internet Web side
PubChemNational Library of Medicine / NIH	http://pubchem.ncbi.nlm.nih.gov/
Developmental Therapeutic Program / NCI and NIH	http://dtp.nci.nih.gov/ Files can be downloaded at: Chemical Structure Lookup Service / Frederick and Bethesda http://cactus.nci.nih.gov/
ChemBank Initiative for Chemical Genetics / NCI	http://chembank.broad.harvard.edu/
DrugBank / University of Alberta	http://redpoll.pharmacy.ualberta.ca/drugbank/
Chemical Entities of Biological Interest / European Bioinformatics Institute, EBI and EMBL	http://www.ebi.ac.uk/chebi/
World of Molecular Bioactivity (WOMBAT) / Sunset Molecular Discovery	http://www.sunsetmolecular.com/index.php
Binding Database / University of Maryland Biotechnology Institute	http://www.bindingdb.org
PDBbind / University of Michigan	http://www.pdbbind.org/
Mother of All Databases (MOAD) / University of Michigan	http://www.bindingmoad.org/
Ligand-Protein DataBase / The Scripps Research Institute	http://lpdb.scripps.edu/
Protein Ligand Database / University of Cambridge	http://www-mitchell.ch.cam.ac.uk/pld/
ChemMine / University of California, Riverside	http://bioweb.ucr.edu/ChemMineV2/
French National Chemical Library / National Center for Scientific Research, CNRS	http://chimiotheque-nationale.enscm.fr/ http://chimiotheque.ujf-grenoble.fr/induk.html
Therapeutic Target Database / National University of Singapore	http://xin.cz3.nus.edu.sg/group/cjttd/ttd.asp
Kyoto Encyclopedia of Genes and Genomes / Kyoto University, University of Tokyo	http://www.genome.jp/kegg/
ChemIDplus	http://chem.sis.nlm.nih.gov/chemidplus/

weight. The user is informed about more than 30 screening projects divided into more than 440 individual experiments (last web - visit September 2006).

DrugBank

DrugBank is a product of Canada's genomics strategy to the end that information on drug safety is collected [23]. Some 4,300 drugs are organized into four major groups: (i) FDA-approved small organic substances; (ii) FDA-approved protein/peptide drugs; (iii) nutraceuticals or micronutrients and finally (iv) experimental agents. More than 6,000 biological targets are linked to these drugs. Web-based full search in DrugBank allows the user to download any interesting text entry - called a DrugCard - or any structural file associated to it.

Chemical Entities of Biological Interest (ChEBI)

ChEBI focuses on small chemical compounds and includes both synthetic and natural products. It is a cornerstone in the database system of the European Bioinformatics Institute (EBI) [24]. There are two main compound sources: (i) an Integrated Relational Enzyme database from EBI or (ii) another small molecule database from the Kyoto Encyclopedia of Genes and Genomes (see below). Molecules in the database can be browsed using several web-based filter options such as SMILES, formula, registry numbers, IUPAC names. Chemical structures and biological information can be downloaded, too.

World of Molecular Bioactivity (WOMBAT)

WOMBAT constitutes an annotated database distributed by Sunset Molecular Discovery. The free database holds information from the scientific literature, specifically from papers published in medicinal chemistry journals. The most recent version contains information of 154,236 entries, totaling over 307,700 biological activities on 1,320 unique targets, for instance, G-protein coupled receptors, ion channels, kinases, serine proteases. Molecules are further annotated with calculated logP or descriptors associated with Linpinski's rule of five [25].

Binding Database

The Center for Advanced Research in Biotechnology at the University of Maryland Biotechnology Institute operates this online database [26-28]. It has gathered nearly 19,700 binding affinities of synthetic ligands on more than 230 targets through either enzyme inhibition or isothermal titration calorimetry. Key word search options include target names, compound names, substructures or SMILES strings. The

user may retrieve both the molecular structures and the inhibition measurements as annotated SDF files. The web site contains links to other storage services regarding molecular interactions such as the PDBbind database [29,3] and the Mother of All Databases (MOAD) [31,32]. Examples of further World Wide Web accessible databases that gathers structural information linked to experimental binding data is the Ligand-Protein DataBase [33] and the Protein Ligand Database [34].

ChemMine

The University of California, Riverside web site, contains a suite of tools for free compound searching, structure-based clustering, descriptor calculations and search options for published biological activity together with target protein descriptions [35]. The annotated database includes over 5,800,000 public and commercial synthetic or natural compounds. The structures and annotations can be searched by chemical properties, substructure matches, structural similarities or biological activities.

French National Chemical Library

The French National Chemical Library is a federal initiative of French academic laboratories. They share vast collections of synthetic products in a database with some 26,800 substances. Its scaffold-diversity analysis capacities have been demonstrated in [36].

Therapeutic Target Database

The National University of Singapore gives access to this web-based resource. The database currently hosts 1,535 proteins and nucleic acid targets, and 2,107 drugs or ligands. The targets represent 125 different human diseases. The database stores knowledge taken from literature and connects its entries to corresponding sequences, 3D-structures, functions, drug-ligand binding properties, drug usages or effects held in other databases. The query items are target names, disease names, drug therapeutic classifications and so forth [37].

Kyoto Encyclopedia of Genes and Genomes (KEGG)

A combined effort in bioinformatics on the part of Kanehisa Laboratories at the Bioinformatics Center of Kyoto University and the Human Genome Center at the University of Tokyo, KEGG is actually more a collection of databases for better understanding cell function of complex systems and entire organisms in view of genome analysis [38]. Its Compound database administers 14,000 entries of known metabolic compounds and to a lesser extent pharmaceutical and environmental substances.

ChemIDplus

It was conceived as a web-based source containing information on the chemistry of substances [39]. ChemIDplus has been developed by the National Library of Medicine. It keeps records on pharmaceutical, industrial and environmental substances reaching a total of roughly 160,000 structures and 360,000 corresponding names. ChemIDPlus constitutes a member the TOXNET database family [40].

Electronic Orange Book

FDA's Electronic Orange Book for approved drug products with its orange colored pages enjoys great popularity amidst health care professionals [41].

Table **4** lists widely used commercial databases dotted with annotations which came to our attention but we do not make any claim of being a complete listing.

2.2. Very Large Compound Collections for Database Mining

In addition to the aforementioned databases with database mining capabilities there are others providing very large compound collections, for instance the *Available Chemicals Directory*, the *Cambridge Database,* or the *Chemical Abstracts Registry* have already been reviewed in [1,3].

In recent times, Monge *et al*. compiled and analyzed a set of 32 libraries gathering 2.6 million unique compounds [42]. As an instance out of the larger commercial databases we present *iResearch Library* from ChemNavigator, which holds more than 26 million chemical samples [43]. The commercial Dictionary of Natural Products, which has become quite popular, holds chemical, physical and biological records on nearly 200,000 compounds [44].

The *ZINC* database developed at the University of California, San Francisco, is another large compilation of more than 4.6 million entries. The molecules and structures are freely available for database mining purposes [45]. Their records are enriched by property annotations such as molecular weight, calculated LogP, and number of rotatable bonds. Each molecule in the library contains vendor and purchasing information and is directly amenable to receptor docking.

Ligand.Info can be found on the Internet as a free service aimed at handling challenges of today's database mining operations. It compiles various publicly available databases issued by the BioInfoBank Institute in Poland [46]. More than one million small molecules have been obtained so far from commercial and public providers. This database can be screened using a Java-based search tool.

Table 4. List of Popular Commercial Databases Annotated with Biological Activity Stating Database, Vendor and Web Sites

Database / Vendor	Internet Web side
Comprehensive Medicinal Chemistry (CMC) MDL	http://mdl.com/products/knowledge/medicinal_chem/index.jsp
MDL Drug Data Report (MDDR) MDL	http://www.mdli.com/products/knowledge/drug_data_report/index.jsp
Derwent World Drug Index (WDI) Thomson Scientific	http://scientific.thomson.com/products/wdi/

2.3. Datasets Used to Develop Computational Techniques

The interplay between large compilations of data sets and new implementations of *in silico* methods has boosted the general use of databases to develop new ideas and computational approaches. The most frequently applied resources are examined in [18]. Andreas Bender compiled a superset of 44 data sets organized into nine categories that have been used in works by other scientists in the field of cheminformatics [47]. Also, the European Bioinformatics Institute has provided free access to a number of datasets [48]. The *Collection of Bioactive Reference Analogues* (COBRA), which has incorporated 4236 sample molecules from the literature [49], has been used to evaluate a modified version of the *k*-means clustering algorithm [2].

3. TOOLS FOR SAR-STUDIES BASED ON STRUCTURAL ANALYSIS OF COMPOUND LIBRARIES

The increasing number of commercial and public chemical databases has boosted innovation in the field of cheminformatics, especially in data management [3]. Today's technological advances offer fully automated structural analysis of compound collections, albeit with certain pitfalls (Table **5**).

Table 5. List of Challenges Encountered in Fully Automated Solutions Aimed at Handling Molecular Computations

Integrate heterogeneous programming languages, software and informatics tools
Convert data types and ensure compatibility between different types of data sources
Collect chemical structures, and experimental or already computed data from different electronic sources or even print media
Evaluate large amounts of molecular data in a consistent and unrestricted protocol
Recognize mesomeric systems, tautomers, ionization, dimerizations etc.
Multiple binding modes of ligand
Incorporate protein target flexibility and higher energy conformation states of ligand in conformational analysis in case of induced fit into unfavorable conformations, observed in over 50 % of complex structures
Protein target (receptor, enzyme) selectivity of ligand
Detect and manage untreatable molecules and outliers to a rule (e.g. cell uptake below)
Predict pharmacokinetics: roughly 30 % of all known drugs may be substrates for active transporters across cell membranes. Hence Log P does not always reflect passive diffusion processes
Take decisions if problems concerning the structural models are met
Handle all possible exceptions in an unattended way
Missing data and uneven data quality management

Due to the critical examination provided by Roberts *et al*. of *SLASH, HookSpace, RECAP* or *Stigmata* [11] our emphasis in this review is placed upon other approaches in the following:

LeadScope is a commercial program that performs systematic substructural analysis of compound collections using structural features [50]. There are 14 major structural classes (e.g., structural patterns) that are predefined in a template library and are related to common building blocks in medicinal chemistry. Among these structural classes are amino acids, functional groups, aromatics, heterocycles, etc. The user interface of LeadScope assists not only the visualization of all the structural classes in the data set, but also helps to apply filtering criteria and identify statistically significant features [11]. An excellent example provided by this approach is the analysis of the NCI database [11,51].

Molecular Equivalence Index (MEQI) was developed by Johnson and Xu [52,53]. The program parses a chemical structure into five major categories: complete 2D-structure, cyclic system, rings, side chains and functional groups. Each chemotype is identified by a code of four or five characters that uniquely identifies that chemotype. This approach creates chemotypes that are equivalence classes at a given level of structural resolution, thus overlapping classes do not occur. A chemical structure can also be parsed at different levels of structural resolution. MEQI is free for academics [54]. It utilizes several post-processing capabilities, e.g. selections of most or least frequent chemotypes and generation of chemotype-based fingerprints. Recent applications of MEQI are the construction of an annotated compound library directed to nuclear receptors [5] and a chemotype-based hierarchical classification of the NCI-AIDS database [16]. A related approach based on chemotypes has recently been published by Wolohan *et al*. [55]. In this so-called *Structural Unit Analysis* the underlying algorithm splits the molecules into fragments and then analyzes what "structural subunits" are associated with activity. The authors applied this approach to explore the NCI database.

ClassPharmer is a commercial software product [56] that classifies structures into molecule classes based on the principle of maximum common substructures (MCS) [57]. The size of the MCS is customizable by modifying a homogeneity setting. In addition to the MCS, the chemical environment is also considered as a means for sorting out the molecules. In this approach the same molecule may appear in various classes depending on the class definition. If necessary a customize option adjusts the level of accepted redundancy. Newer studies with ClassPharmer include a scaffold diversity analysis of 17 commercially available screening collections [58] and a substructure analysis of ligand sets from five target families [59].

Distill, a software from Tripos [60] parses and classifies all molecules through a similarity filter based on their common substructure. Prior to the output of hierarchical clustering each molecule is evaluated by scores calculated from deviations in the observed number of atoms, bonds, ring bonds, hetero-atoms or branched atoms in the common substructure. Although the construction of the hierarchy is independent of the processed property data, the resulting nodes in the dendogram can be color-coded by averaged property values. Thus, a visual analysis can be performed between

structures and properties. Distill also enables the construction of queries from selected compounds to search other data sets.

Fragmenter can be used to sort out molecules of a data set with virtually any fragmentation engine. It includes the Retrosynthetic Combinatorial Analysis Procedure (RECAP) method of molecule fragmentation following simplified retrosynthetic rules [61]. The program also analyzes libraries in terms of side chains and the specific side chain position. *Jklustor* performs clustering of large databases using MCS among other descriptors. Fragmenter and JKlustor are integrated in ChemAxon's JChem software suite [62].

The next step is logically the random recombination of RECAP fragments to generate virtual combinatorial libraries of synthetically reasonable chemical structures, a new solution to assist virtual *de novo* drug discovery and proposed by MOE, a modeling package for cheminformatics and bioinformatics [63]. It can generate diverse combinatorial libraries for functionalized scaffolds either by systematic combinations of fragments with scaffolds to yield all possible products (*fully-enumerated QuaSAR-CombiGen*) or by random combinations using scaffold and fragmental substituent databases (*non-enumerative QuaSAR-CombiDesign*). The latter is also capable of statistically sampling diverse subsets from extremely large virtual libraries. The user-made data sets can be tested by HTS against either molecular property filters/ descriptors or against computed models of pharmacophore, linear or binary correlations from QSAR/QSPR studies as well as similarity/fingerprints according to the working hypothesis. Its latest 2006 version comes with a database of over half a million compounds for lead search against known biomolecular targets for screening and ligand docking. To this end, it also provides a user interface for FlexX ligand docking into 3D-structures of receptors and enzymes [64].

Examples of proprietary software that has been developed to conduct SAR-studies of large databases are *Visuali-SAR* [65] and *Hits Analysis Database* [66], both developed by pharmaceutical companies. Hits Analysis Database provides insight into the structural classes of molecules and conducts cluster analysis applying MCS protocols.

In a recent example of a research project aimed at identifying structural features associated with biological activity, Ertl *et al*. explored the heteroatomic ring distribution in active molecules taking two active compound sources: the *World Drug Index* (WDI) as well as the MDDR database (Table **4**). The authors compared the structures with the molecules in the *Dictionary of Natural Products* and commercial molecules. They summarized previous studies analyzing ring systems in active molecules [67].

The work by Koch *et al*. also reflects structural analysis of large compound libraries supplied by the *Dictionary of Natural Products*. Here, the authors undertook a structural classification of natural substances in terms of scaffolds in a hierarchical way [68].

Lameijer *et al*. analyzed the NCI database identifying so-called "chemical clichés" as most frequent structural fragments and pairs of fragments [69]. Another recent structural analysis of the NCI-AIDS database is a chemotype-based hierarchical classification of active molecules [16].

The *GreenPharmaDataBase-net* project (GPDBnet) was developed at Greenpharma, France. It constitutes a platform that is particularly useful for treating dispersed knowledge along the wide discovery front of phytopharmaceutical and ethnopharmacological science. In fact, due to the heterogeneity of sources it often becomes extensively time consuming to retrieve relevant pieces of information. In order to overcome these difficulties, a knowledge management strategy is encouraged. GPDBnet embraces an internal database coupled with molecular modeling tools and becomes accessible from the Greenpharma's intranet *via* research contracts. Its built-in database gathers records on plants from internal or external literature: their traditional uses, their biological properties, their metabolites including structures, biological assays, and the targets in human cells. So it becomes possible to conduct cross-queries with traditional use of the compounds, biological activity and biomolecules in a straightforward manner. Isolated phyto-pharmacological agents can also be exploited to discover novel targets, i.e. *reverse pharmacognosy*. The workflow of GPDBnet is presented in Fig. **(1)**. In a more generalized view, data mining architectures like GPDBnet, are knowledge management tools which focus on generating either new ideas or validating them. To optimize the use of this database and allow predictions, molecular modeling software such as the virtual screening program, Selnergy™ [70,71] has been coupled to GPDBnet.

Nowadays *in silico* screening has become a well-established method for hit discovery, yet results still have to be experimentally validated. GPDBnet's asset is twofold: on one hand, virtual screening results are validated with the existing experimental data in order to get robust predictions, i.e. increasing hit rates, while on the other hand, selected substances can further be evaluated on Selnergy or other prediction tools for building new research hypotheses, *e.g.* ADMET models Fig. (**2**). Several applications support the successful implementation of this strategy [70-72].

4. DISCUSSION

In a competitive environment such as the pharmaceutical industry, being first in a market is desirable. Despite long and massive investments on promising technologies such as HTS and combinatorial chemistry, a recent survey [73] showed that 15 years and $880 millions are the average time and the cost from the target validation to the regulatory approval, respectively. Several authors have suggested reconsidering natural products as a source of bioactive entities [74-77], as nature has still a lot of lessons to teach us. Meanwhile, to absorb the risks of pharmaceutical research, a strategy of merging within pharma-industry has been accelerating for a decade or so, exacerbating communication problems between research groups, especially those that are separated geographically. Internet technologies are obviously an appropriate tool for improving communication and accessing useful information. But in reality, one is confronted by an overwhelming amount of data. As a consequence of the avalanche of information compound databases play a pivotal role in the drug discovery process. The number of annotated chemical databases is increasing many of which are freely available. Certain compound collections embrace additional web-based resources that help in retrieving specific information. The structural analysis methods currently

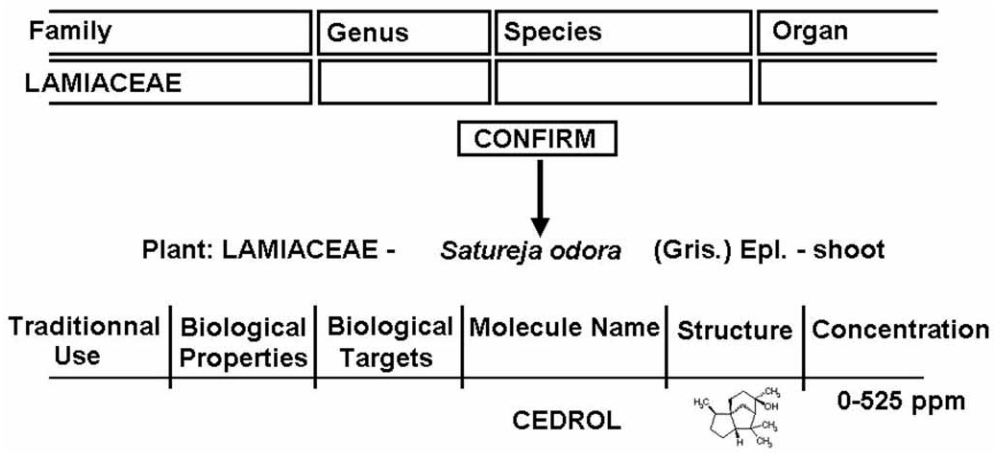

Fig. (1). Illustration of a database search in *GPDBnet*: a specific plant becomes interesting and the query is launched with the following results (see Fig. **2**).

in use originate from several cheminformatics approaches that try to reveal structure-activity relationships among compounds in large virtual databases.

Managing large compilations of molecular records involves identifying the most common substructures associated with observed activities. Direct outcomes of such studies are strongly linked to concepts like scaffold hopping and privileged structures. The NCI database and other free sources are frequently consulted to test and validate novel computational approaches. But, even today, not all of the molecular data-handling challenges have been solved (Table **5**). New programs or enhanced versions must tackle these problems. Hence, the future is an integrated platform or net of various concepts, not just electronic storage facilities. Such a package should include predictive modeling software to compute and store molecular profiles as complements to any experimental data. It should also be able to carry out large numbers of ligand docking calculations against biological targets of either known or computed structure. *In silico* methods estimate *in vitro* or *in vivo* properties associated with ADMET, too.

On the other hand *in vitro* experiments remain paramount for verifying computed models or for determining results since *in silico* predictions do not hold in all cases. Especially, HTS in cell cultures determines ADMET properties, detects metabolic pathways, stability and toxicity, or identifies candidates that interact with more than one biomolecular target.

In contrast to the aforementioned virtual databases for *in silico* approaches to drug discovery the identification of

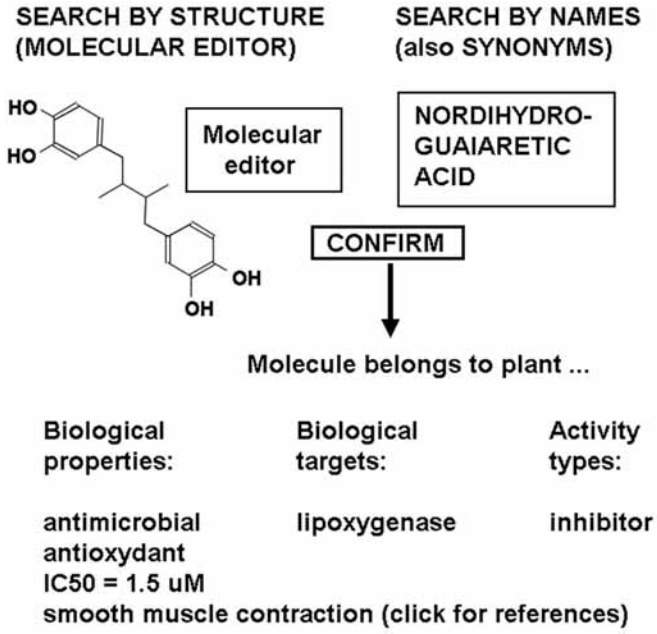

Fig. (2). Illustration of query and results in *GPDBnet*: The molecular editor (*JME sketcher*) starts the substructure search. A 3D-representation can be displayed with *Chime plug-in*.

novel drug targets can be conducted experimentally by *reverse pharmacology*. The latter is sometimes referred to as *chemical genetics* using real libraries of synthetic or natural compounds with pharmacological effects to find new disease-related molecular targets in cells. A recent review published in this Journal explains the latest developments of such *in vitro* HTS methods [78]. Integrating data management with a user-friendly interface combining automated data storage with 3D- virtual screening tools already emerge as innovative informatics solutions to boost knowledge search in a heterogeneous data environment. In this respect, our own solution in the field of phyto-pharmacology, GPDBnet, illustrates such an integrated data mining tool to explore molecular resources of plants.

CONCLUSIONS

Large libraries of chemical compounds reflect the exponential growth in amount of data in the field of drug discovery. For the last two decades molecular R&D concepts have tended towards fully automated informatics solutions to study structure - activity relationships, especially in combination with screening of docked ligand candidates to biological target structures. The *in silico* selection of natural or synthetic compounds aims at improving effectiveness throughout the discovery pipeline. The focus of this review lies mainly on freely accessible public virtual databases. They cover a wide range of known molecular targets and in some cases include potentially new therapeutic cellular pathways. They also constitute valuable sources for docking simulations. Novel compound collections designed for database mining are also described.

In conclusion, we discuss recent advances in computational tools that have been developed for the medicinal chemist to identify structural leads and activity patterns in large compound collections.

ACKNOWLEDGEMENTS

This work was financially supported by Dr. *Pedro Hugo Hernandez T*., VIEP, *Benemérita Universidad Autónoma de Puebla* (BUAP), Mexico. Thanks to M.C. *Julian A. Yunes Rojas*, www.buap.mx, for improving the figure quality.

ABBREVIATIONS AND GLOSSARY

2D- / 3D-	=	Two-dimensional / Three-dimensional
ADMET	=	Absorption, Distribution, Metabolism, Elimination and Toxicity (pharmacokinetics)
AIDS	=	Acquired Immune Deficiency Syndrome
Annotation	=	Assignment of biological or chemical information for pattern recognition using textual or numerical descriptors, chemical signatures, structural features or keys like fingerprints or pharmacophore models to search molecules in database queries
CSD	=	Cambridge Structural Database
Chemotype	=	Structural pattern (e.g., ring, functional group, side chain, etc.)
ChEBI	=	Chemical Entities of Biological Interest
Clustering	=	Sorting out a collection of compounds with assigned descriptors in an effort to reduce unwanted redundancy in data size, structural similarities and other correlations (bias)
Dataset	=	Pieces of information sorted by some filtering scheme or record type
Databank	=	A repository of collected pieces of information that can be digitalized for local (CDs) or remote retrieval (online web servers)
Database	=	A computer storage of records that can be searched and sorted to answer questions of the user through a query macro language or database management system
EBI	=	European Bioinformatics Institute
FDA	=	Food and Drug Administration (of USA)
Fingerprint	=	Set of structural properties (e.g. cyclic, aliphatic, with atoms X, etc.) as a molecular surrogate to compare chemical similarity of compounds
Focused	=	Annotated compound libraries *focused* on a specific set of target molecules, e.g. enzyme class
Functionalizing	=	Attachment of chemical groups or functions on a scaffold
HTS	=	High Throughput Screening (fast filtering and sorting to test against targets or criteria)
ICG	=	Initiative for Chemical Genetics
IUPAC	=	International Union of Pure and Applied Chemistry (chemical naming of substances)
KEGG	=	Kyoto Encyclopedia of Genes and Genomes
MCS	=	Maximum Common Substructure
MEQI	=	Molecular Equivalence Index
NCI	=	National Cancer Institute (of USA)
NIH	=	National Institute of Health (of USA)
PDB	=	Protein Database / Protein Data Bank
Pharmacophore	=	Ligands' 3D-fragments or properties which interact with the receptor's binding site and are essential for activity
QSAR	=	Quantitative Structure - Activity Relationships
RECAP	=	Retrosynthetic Combinatorial Analysis Procedure

R&D = Research and Development (of drugs in pharmaceutical industry)

SAR = Qualitative Structure - Activity Relationships

Scaffold = Main or central fragment common to a group of molecules

SMILES = Simplified Molecular Input Line Entry System

VS = Virtual Screening

WDI = World Drug Index

WHO = World Health Organization

WOMBAT = World of Molecular Bioactivity

REFERENCES

[1] Richardson, R. In *Exploiting Chemical Diversity for Drug Discovery,* Bartlett, P. A.; Entzeroth, M. Ed.; Royal Society of Chemistry, Cambridge, UK, **2006**; pp. 112.

[2] Bocker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. *J. Chem. Inf. Model.,* **2005**, *45*, 807.

[3] Miller, M. A. *Nat. Rev. Drug Discov.,* **2002**, *1*, 220.

[4] Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. *Science,* **2004**, *306*, 1138.

[5] Cases, M.; García-Serna, R.; Hettne, K.; Weeber, M.; van der Lei, J.; Boyer, S.; Mestres, J. *Curr. Top. Med. Chem.,* **2005**, *5*, 763.

[6] Dragon 3.0 (Milano Chemometrics). Todeschini, R.; Consonni, V.; Mauri, A.; Pawan, M. Milano, Dragon Software, Chemometrics, Italy. http://www.disat.unimib.it/chm/

[7] Cosgrove, D. A.; Willet, P. *J. Mol. Graph. Model.,* **1998**, *16*, 19.

[8] Cramer III R. D.; Redl, G.; Berkoff, C. E. *J. Med. Chem.,* **1974**, *17*, 533.

[9] MDL Comprehensive Medicinal Chemistry. MDL Information Systems, Inc.

[10] Bemis, G. W.; Murcko, M. A. *J. Med. Chem.,* **1996**, *39*, 2887.

[11] Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. *J. Chem. Inf. Comput. Sci.,* **2000**, *40*, 1302.

[12] Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. *Angew. Chem. Int. Ed.,* **1999**, *38*, 2894.

[13] Zhang, Q.; Muegge, I. *J. Med. Chem.,* **2006,** *49,* 1536.

[14] Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S. L.; Lotti, V. J.; Cerino, D. J.; Chen, T. B.; Kling, P. J.; Kunkel, K. A.; Springer, J. P.; Hirshfieldt, J. *J. Med. Chem.,* **1988**, *31*, 2235.

[15] Patchett, A. A.; Nargund, R. P. *Annu. Rep. Med. Chem.,* **2000**, *35*, 289.

[16] Medina-Franco, J. L., Petit, J., Maggiora, G. M. *Chem. Biol. Drug, Des.,* **2006**, *67*, 395.

[17] Schreyer, S. K.; Parker, C. N.; Maggiora, G. M. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 470.

[18] Tetko, I. V. *Mini Rev. Med. Chem.,* **2003**, *3*, 809.

[19] http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi

[20] Strausberg, R. L.; Schreiber, S. L. *Science,* **2003**, *300*, 294.

[21] Tolliday, N.; Clemons, P. A.; Ferraiolo, P.; Koehler, A. N.; Lewis, T. A.; Li, X.; Schreiber, S. L.; Gerhard, D. S.; Eliasof, S. *Cancer Res.,* **2006**, *66*, 8935.

[22] Lipinski, C.; Hopkins, A. *Nature,* **2004**, *432*, 855.

[23] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. *Nucleic Acids Res.,* **2006**, *34,* D668.

[24] Brooksbank, C.; Cameron, G.; Thornton, J. *Nucleic Acids Res*., **2005**, *33*, D46.

[25] Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. In *Chemoinformatics in Drug Discovery.* Oprea, T. I. Ed.; Wiley-VCH, New York, **2004**; pp. 223.

[26] Chen, X.; Liu, M.; Gilson, M. K. *Comb. Chem. High-Throughput Screen.,* **2001**, *4*, 719.

[27] Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. *Bioinformatics,* **2002,** *18*, 130.

[28] Chen, X.; Lin, Y; Gilson, M. K. *Biopolymers Nucleic Acid. Sci.,* **2002**, *61*, 127.

[29] Wang, R.; Fang, X.; Lu, Y.; Wang, S. *J. Med. Chem.*, **2004**, *47*, 2977.

[30] Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. *J. Med. Chem.*, **2005**, *48*, 4111.

[31] Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. *Prot. Struct. Func. Bioinformatics,* **2005**, *60*, 333.

[32] Smith, R. D.; Hu, L.; Falkner, J. A.; Benson, M. L; Nerothin, J. P.; Carlson, H. A. *J. Mol. Graphics Model.,* **2006**, *24*, 414.

[33] Roche, O.; Kiyama, R.; Brooks, C. L., III. *J. Med. Chem.,* **2001**, *44*, 3592.

[34] Puvanendrampillai, D.; Mitchell, J. B. O. *Bioinformatics,* **2003**, *19*, 1856.

[35] Girke, T.; Cheng, L.-C.; Raikhel, N. *Plant Physiol.,* **2005**, *138*, 573.

[36] Krier, M.; Bret, G.; Rognan, D. *J. Chem. Inf. Model*., **2006**, *46*, 512.

[37] Chen, X.; Ji, Z. L.; Chen, Y. Z. *Nucleic Acids Res.,* **2002**, *30*, 412.

[38] Kanehisa, M.; Goto, S.; Kawashima, S.; Okuno, Y.; Hattori, M. *Nucleic Acids Res*., **2004**, *32*, D277.

[39] http://chem.sis.nlm.nih.gov/chemidplus/

[40] Wexler, P. *Toxicology,* **2001**, *157,* 3.

[41] http://www.fda.gov/cder/ob/

[42] http://www.univ-orleans.fr/icoa/eposter/eccc10/monge/

[43] http://www.chemnavigator.com/

[44] Dictionary of Natural Products. Chapman & Hall/ CRC: London. http://www.crcpress.com

[45] Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Comput. Sci.,* **2005**, *45*, 177. http://blaster.docking.org/zinc/

[46] von Grotthuss, M.; Koczyk, G.; Pas, J.; Wyrwicz, L. S.; Rychlewski, L. *Comb. Chem. High-Throughput Screen.,* **2004**, *7*, 757. http://ligand.info/

[47] http://cheminformatics.org/

[48] http://www.ebi.ac.uk/FTP/

[49] Schneider, P.; Schneider, G. *QSAR Comb. Sci.,* **2003**, *22*, 713.

[50] http://www.leadscope.com/

[51] Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. *J. Chem. Inf. Comput. Sci.,* **2002**, 42, 393.

[52] Johnson, M. A., Xu, Y.-J. In *Chemical Data Analysis in the Large: The Challenge of the Automation Age*; Hicks, M. G., Ed.; **2001**. http://www.Beilstein-institut.de/bozen2000/proceedings

[53] Xu, Y.-J., Johnson, M. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 912.

[54] http://www.pannanugget.com/

[55] Wolohan, P. R. N.; Akella, L. B.; Dorfman, R. J.; Nell, P. G.; Mundt, S. M.; Clark, R. D. *J. Chem. Inf. Model.,* **2006**, *46*, 1188.

[56] http://www.bioreason.com/

[57] McGregor, J. J.; Willett, P. *J. Chem. Inf. Comput. Sci.,* **1981**, *21*, 137.

[58] Krier, M.; Bret, G.; Rognan, D. *J. Chem Inf. Model.,* **2006**, *46*, 512.

[59] Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. *J. Med. Chem.,* **2006**, *49*, 2000.

[60] http://www.tripos.com

[61] Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. *J. Chem. Inf. Comput. Sci.,* **1998**, *38*, 511.

[62] http://www.chemaxon.com

[63] Molecular Operating Environment software MOE 2006.08. http://www.chemcomp.com

[64] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.,* **1996**, *261*, 470. http://www. BioSolveIT.com

[65] Wild, D. J.; Blankley, C. J. *J. Mol. Graph. Model.,* **1999**, *17*, 85.

[66] Shen, J. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 1668.

[67] Ertl, P.; Jelfs, S.; Muhlbacher, J.; Schuffenhauer, A.; Selzer, P. *J. Med. Chem.,* **2006**, *49*, 4568.

[68] Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. *Proc. Natl. Acad. Sci. USA,* **2005**, *102*, 17272.

[69] Lameijer, E.-W.; Kok, J. N.; Back, T.; IJzerman, A. P. *J. Chem. Inf. Model.,* **2006**, *46*, 553.

[70] Do, Q. T.; Bernard P. *IDrugs*, **2004**, *7,* 1017.

[71] Do, Q. T.; Rénimel, I; André, P.; Lugnier, C.; Muller, C.; Bernard, P. *Curr. Drug Disc. Techn.,* **2005**, *2*, 161.

[72] Bernard, P.; Scior, T.; Didier, B.; Hibert, M.; Berthon, J. Y. *Phytochemistry,* **2001**, *58*, 865.

[73] Tollman, P.; Guy, P.; Altshuler, J.; Flanagan, A.; Steiner, M. *A Revolution in R & D. How Genomics and Genetics are Transforming the Biopharmaceutical Industry*. The Boston Consulting Group. **2001**.
[74] Harvey, A. L. *Trends Pharmacol. Sci.,* **1999**, *20*, 196.
[75] Harvey, A. *Drug Discov. Today,* **2000**, *5*, 294.
[76] Demain, A. L. *Nat. Biotechnol*., **2002**, 20, 331.
[77] Baurin, N.; Arnoult, E.; Scior, T.; Do, Q.T.; Bernard, P. *J. Ethnopharmacol.,* **2002**, 82, 155.
[78] Harrigan, G. G.; Brackett, D. J.; Boros, L. G. *Mini Rev. Med. Chem.,* **2005**, *5*, 13.

Received: 20 October, 2006 Revised: 30 November, 2006 Accepted: 01 December, 2006