Drug Discovery Today: Technologies

Vol. 3, No. 4 2006

Editors-in-Chief
Kelvin Lam – Pfizer, Inc., USA
Henk Timmerman – Vrije Universiteit, The Netherlands

In silico techniques

Target, chemical and bioactivity databases – integration is key

Tudor I. Oprea[1,*], Alexander Tropsha[2]

[1]Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, 1 University of New Mexico, Albuquerque, NM 87131-0001, USA
[2]Laboratory for Molecular Modeling, CB #7360 Beard Hall, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Various biological, chemical and bibliographic databases have grown tremendously in recent years with respect to both their diversity and size. In many cases, these databases have been established to address specific interests of their developers in biological or chemical systems with relatively little attention paid to the integration of different types of biological, chemical and literature data. Pharmaceutical industry increasingly relies on such databases for generating new ideas for identifying potential drug candidates and their targets. In this paper, we present a brief overview of many available public and commercial databases that contain information on biological targets or small molecule ligands as well as several integrated databases. We emphasize the database integration and federation as most important current trends in informatics driven pharmaceutical discovery.

Section Editors:
Tudor Oprea – University of New Mexico School of Medicine, Albuquerque, USA
Alex Tropsha – University of North Carolina, Chapel Hill, USA

Target, chemical and bioactivity information: branches of the same tree

Target (macromolecule) and lead (small molecule) identification are keys to successful drug discovery. Today's 'prior art' intellectual property climate requires that medicinal chemists familiarize themselves not only with the proprietary data but also with public domain information related to both target and lead structures active on the intended, or related targets. Access to such information is facilitated by chemical databases such as SciFinder (see Weblinks) and Beilstein (see Weblinks), by patent databases such as the MDL Drug Data Report (MDDR) (see Weblinks), and the World Drug Index (WDI) (see Weblinks) and by an increasing number of public-domain and private databases focused on structure-activity data. An increasing demand is currently placed on indexing not only the information related to chemical structures and their measured (biological) properties, that is, cheminformatics, but also details related to the biological assays and targets associated with these measurements, that is, bioinformatics. Although the average end-user continues to emphasize datamining focused on prior art chemistry, it is expected that related targets are also evaluated for reasons related to selectivity, or perhaps related to the absence of any suitable leads on the target of interest. Thus, database systems that seamlessly mine chemical, biological and target-related data in an integrated manner and provide information-rich content are anticipated to replace those systems that do not offer the appropriate contextual background. Such database systems are currently under development. In this note we focus on bioactivity databases with bioinformatics and cheminformatics content, and discuss some aspects of their integration.

Mining the chemistry/biology interface requires adequate tools that capture, store and retrieve the appropriate information in a contextual manner, as discussed elsewhere [1].

[*]*Corresponding author:* T.I. Oprea (toprea@salud.unm.edu)

1740-6749/$ © 2006 Elsevier Ltd. All rights reserved. DOI: 10.1016/j.ddtec.2006.12.003

A variety of chemical, for example, SciFinder, or medicinal chemistry, for example, MDDR, or directly drug-related databases, for example, the Physician Desk Reference (PDR) (see Weblinks), capture such information, but only partially. For example, one can search a chemical (sub)structure in SciFinder, and retrieve a several literature references indexed for 'biological' or perhaps 'toxicological' content; or retrieve a certain activity class in MDDR. However, no searchable field to query the quantitative aspects of biological activity are available, and one has to mine compounds through peer-reviewed papers and patents individually, to retrieve quantitative biology data. Designed for clinical practitioners, PDR provides drug monographs, where chemical structures and biological properties are provided per query, but without searchable fields. Furthermore, these database systems do not capture, in a context-dependent manner, information related to the drug targets that have been addressed by these small molecules or marketed drugs.

There are several essential questions that could be answered in the context of pre-clinical drug discovery by mining available information in scientific literature or available databases:

(1) Has this, or a related target been validated in the clinic? The end-user may find it useful to learn that the target of interest can be manipulated therapeutically; or that targets that are related by sequence (e.g. subtypes of the same receptor), or by function (e.g. sodium-dependent re-uptake transporters), or by pathway (e.g. enzymes from the estrogen biosynthesis pathway) have been manipulated with small molecule therapeutics, or perhaps with antibodies.

(2) Has this, or a related chemical scaffold, been successfully used in the clinic? It is relevant to identify scaffolds that are often present in launched drugs for multiple (e.g. benzodiazepines) or for the same (e.g. quinolones) therapeutic indication; it is also relevant to identify 'unwanted' scaffolds related to toxicology (e.g. thioureas) or to an over-crowded patent landscape (e.g. steroids).

(3) Is this scaffold selective? Or does it have activity at (un)related targets? The ability to search quantitative aspects of biology allows the identification of suitable scaffolds, as perhaps selectivity (or lack thereof) is intended. One should keep in mind that not all scaffolds indexed in databases are as active as claimed, and that *in vitro* bioactivity is not necessarily related to relevant activity in humans.

Research related to drug discovery produces large amounts of data in seemingly unrelated fields, such as molecular and cellular biology, chemical biology, combinatorial and medicinal chemistry, genetics and toxicology. This information needs to be organized, queried and structured to guide the scientific process, and to transform data into information and knowledge. Three major components of this process have been identified and discussed elsewhere [1]:

- Chemical and bioactivity information: combines chemical structures with experimental or calculated chemical and physical properties;
- Target and protocol information: biological target and experimental protocol data;
- Reference information: bibliographic information for all units in the database.

Chemical and bioactivity information

Chemical and bioactivity information relates to the storage of chemical structures and associated molecular data in machine-readable format. Key to storing chemical structures is the atomic connectivity, expressed in connection tables that store two- and/or three-dimensional atomic coordinates. Essential to the proper handling of chemical information is the annotation with external identifiers, which include generic (trivial, commercial) and IUPAC names [2], CAS registry numbers [3] as well as unique structural representations. Far from redundant, this type of information can prove very useful where chemical structure errors, tautomers and protomers, or different salt formulations are concerned [4]. Key to the proper annotation of such database systems, bioactivity relates biological activity data – primarily activity type and value – with unique indexes identifying the chemical compound, the biological target or cell (rarely whole organisms, with the experimental protocol and bibliographic references. Frequently used bioactivity types are: inhibitory concentration at 50% (IC_{50}), the molar concentration of an antagonist/inhibitor that reduces the response/reaction velocity to an agonist/substrate by 50%; A_2 – the molar concentration of an antagonist that requires double concentration of the agonist to elicit the same submaximal response, obtained in the absence of antagonist; effective concentration 50% (EC_{50}), the molar concentration of an agonist/substrate that produces 50% of the maximal possible effect (or reaction velocity) of that agonist (substrate); inhibition and direct binding experiment equilibrium dissociation constants (K_i and K_d, respectively) [5]. Additional bioactivity fields include experimental observations and errors, sometimes images (e.g. Schild plots), as well as keywords such as 'partial', 'inverse', 'competitive', 'agonist', 'antagonist' and 'inhibitor'. The PubChem (see Weblinks) keywords 'active' and 'inactive' can also help to rapidly discriminate between the different bioactivity categories. Relevant to drug discovery are also pharmacokinetic (PK) data, derived from clinical observations or from animals. Properties such as fraction of the compound/drug absorbed (%F), oral bioavailability, plasma protein binding, volume of distribution at steady-state, half-life, total blood clearance, etc. are also considered bioactivity end-points and need to be queried, where available [6].

Target and protocol information

Target and protocol information relates to the storage of target and gene information, as well as associated bioassay data in machine-readable format. Many bioinformatics databases are freely available on the Internet. Proper unique identifiers (the equivalent of chemical names), such as those from Swiss-Prot (see Weblinks), enable the end-user to navigate across these databases using uniform resource locators (URL) (see Weblinks) hyperlinks. Therefore, extended target names and functions, as well as information related to their classification and species, need to be stored. For example, using functional criteria, a target might be an enzyme, or receptor, or ion-channel, or transporter or 'other' (unspecified) protein, as well as nucleic acid (DNA or RNA). Furthermore, an enzyme falls in one of six major biochemical classes: oxidoreductases, transferases, hydrolyses, lyases, isomerases and ligases (see Weblinks), whereas receptors are further classified as nuclear hormone receptors (NHRs) [7], G-protein coupled receptors (GPCRs) (see Weblinks) [8], etc. The use of a controlled vocabulary enables the query of protocol information via pre-defined keywords, which stores information related to specific/non-specific (radio)ligands, substrates, temperature, pH, buffer, incubation time, etc. Both target and protocol information is subject to change, as newly discovered targets are re-classified and the occasional errata are published.

References

References contain bibliographic information, such as authors or inventors, title, source (e.g. journal name or patent), as well as other pertinent information (volume, page numbers, patent number etc.). Using unique identifiers, for example, PubMed (see Weblinks) or digital object identifiers (DOIs) (see Weblinks) entries can be hyperlinked to the appropriate abstract or full-text publication via MEDLINE or other databases. Publisher-provided or Medline subject headings (MeSHs) keywords can provide further content to the target and protocol fields. In-house reports, as well as Internet references should also be indexed, as they provide valuable content.

Target, chemical and bioactivity database integration

Drug discovery is an eminently multi-disciplinary effort. Therefore, the use of isolated, focused databases is no longer expected to meet user needs. Open access databases such as the PubChem system, as well as commercial products such as DiscoveryGate are available using the Internet. The information is often stored in heterogeneous manner, using multiple data types and formats. This often leads to the query-within-query scenario, where the results of an earlier query, on a particular database, are used to query other databases. However, such manual multi-database queries are often time consuming. Automated data integration has become a crucial need.

Bringing together different information from different sources into a single data model can often be difficult, even for data of the same kind, because of the disparity of purposes between various information sources. A distributed system that acts as a front end to multiple local databases, or is perhaps structured as a global layer on top of local databases [9], can be used to address differences in data representation and function between local sources. Federated databases are considered a better approach to integrating multiple databases: continued operation of existing applications is allowed, controlled integration of existing databases is supported, and incorporation of new applications and new databases is facilitated [10]. Technologies such as CORBA (see Weblinks) [11], Java (see Weblinks), XML (see Weblinks) and HTML (see Weblinks) [12] provide a powerful and flexible method of integrating data from different databases. For chem-bio-informatics data integration, external identifiers that correspond to, for example, target information from other databases (e.g. Swiss-Prot) can be embedded in a particular bioactivity database. This becomes valuable for web-enabled links, because the URL [8] acts as entry point into the other databases, often associated with a numeric or alphanumeric identifier for a specific resource.

As the scientific disciplines covered by chem-bio-informatics database systems continue to evolve, and so does the captured data. Furthermore, the sheer volume of information continues to expand at an exponential pace. This situation has become too difficult for any one group to capture and index appropriately. Often, scientists focus on a class (e.g. serine proteases) or family (e.g. integrins) of macromolecules, or perhaps on a therapeutic area (e.g. cardio-vascular), which explains the need for class-specific, specialized databases.

In the field of bioinformatics, these databases collect and organize data around a single class of macromolecules (e.g. GPCRs), or around a particular topic of interest (e.g. cancer), whereas in the field of cheminformatics, they often collect target-focused small molecules (e.g. kinase inhibitors). When curated well, these resources allow scientists to query a single source to obtain a significant fraction of the most relevant data. What follows is a succinct overview of several bioinformatics and cheminformatics databases with the common theme of bioactivity. Since (as mentioned above) the size of each database is a rapidly moving target, we did not include this information on purpose; we suggest that readers visit the respective web sites for the most up-to-date information on each database. Table 1 provides a summary of all databases discussed in this review.

Table 1. Databases of relevance to pharmaceutical drug discovery

Database type	Database name	Public (P) or commercial (C)
Target	UniProtKb/Swiss-Prot	P
	EC-PDB	P
	PDBRTF	P
	Entrez Protein	P
Chemical/bioactivity	PubChem	P
	PubChem BioAssay	P
	KEGG	P
	GPCRDB [8]	P
	NucleaRDB [7]	P
	NURSA	P
	IUPHAR	P
	DrugBank [16]	P
	BIDD	P
	SMID	P
	BIND [24]	P
	ChemBank	P
	*Ki*Bank	P
	ZINC	P
	AurSCOPE	C
	MediChem	C
	Merck Index	C
	DiscoveryGate	C
	PathArt	C
	WOMBAT	C

Target databases

Many bioinformatics databases cover the 'Target' space, often in an integrated manner. Such an example is the array of databases from the European Bioinformatics Institute (EBI) (see Weblinks). Maintained by the Swiss Institute for Bioinformatics (SIB) and EBI, the UniProtKB/Swiss-Prot protein knowledgebase (Swiss-Prot) is a curated protein sequence database that provides a uniquely non-redundant high level of annotation. Together with UniProtKB/TrEMBL, it forms the Universal Protein Resource (UniProt) knowledgebase (see Weblinks). The primary object of this database is proteins, for which sequence data, references and the taxonomic data is provided. It further captures information related to protein function, domains and sites, secondary and quaternary structure, similarities to other proteins, etc. Since Swiss-Prot is cross-referenced with more than 60 different databases, this makes it one of the most integrated target databases in the public domain. Key to this integration are the pointers to Swiss-Prot entries. A related EBI database, Enzyme Structures Database (EC-PDB) (see Weblinks), contains the known enzyme structures that have been deposited in the Protein Data Bank (PDB) (see Weblinks) [13]. A related database, applied to enzymes and nuclear receptors, Representativity of Target Families in the Protein Data Bank (PDBRTF) is available from IMIM (see Weblinks).

Transporters are indexed using the IUBMB approved classification system for membrane transport proteins, known as the transporter classification (TC) system (see Weblinks), in the TCDB (see Weblinks) database. Analogous to the enzyme commission (EC) system for enzyme classification, the TC system incorporates phylogenetic information and provides a hierarchical index based on five criteria: transporter class, subclass, family or superfamily, subfamily and substrate or range of substrates transported.

Target-bioactivity databases

Hosted by the National Library of Medicine the *Entrez* life sciences search engine (see Weblinks) is the largest public-domain system of databases that captures information relevant to targets [via e.g. the *Entrez* Protein system (see Weblinks)], chemicals (via PubChem) and bioactivity [via PubChem Bioassay (see Weblinks)]. This small molecule portal, part of the National Institutes of Health Roadmap (see Weblinks) [14], is organized in three databases: (1) PubChem Substance, which contains descriptions of chemical samples from a variety of sources, and links to PubMed citations, protein 3D structures and biological screening results that are available in PubChem BioAssay; (2) PubChem Compound, which contains chemical compound information related to substances; structures stored within PubChem Compounds are pre-clustered and cross-referenced by identity and similarity; additionally, calculated properties and descriptors are available for searching and filtering of chemical structures and (3) PubChem BioAssay, which captures bioactivity screens of chemical substances described in PubChem Substance; it provides searchable descriptions of each bioassay, including descriptions of the conditions and readouts specific to that screening procedure, as well as an 'active/inactive' label. Finally, PubChem Chemical Structure search (see Weblinks) allows the user to query these databases online using chemical substructures, descriptive terms, chemical properties such as Lipinski's rule of five [15] and structural similarity. All the above databases rely on PubMed [10] for appropriate literature references.

Another broad purpose integrated publicly available database has been developed by Kanehisa and colleagues in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo. Called the Kyoto Encyclopedia of Genes and Genomes (KEGG), this project aims at the development of 'complete computer representation of the cell, the organism and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information' (see Weblinks). The KEGG database includes four constituent databases: the PATHWAY database summarizes literature data on pathways maps and molecular interaction and reaction networks for various biological processes; the BRITE database that supplements the

PATHWAY database by going beyond molecular interactions to deduce and infer functional relationships between various biochemical systems and sub-systems; the GENES database that catalogs information on all complete and partially complete genomes in eukaryotes, bacteria and archaea; and finally, the LIGAND database that contains data on biologically active compounds, drugs, glycans, reactions and reactant pairs of compounds.

Targets with pharmacological relevance are indexed in GPCRDB [8], which collects GPCR data, and in NucleaRDB [7], which collects NHR data, respectively. Target information (sequence data, gene information, mutants, experimental 3D structures) is linked to small molecule information (ligand binding data). Computationally derived data, for example, multiple sequence alignments, phylogenetic trees and 3D models, complement these databases. An important element to the organization of these databases is the pharmacological classification of receptors; access to the data is provided using hierarchical lists of receptor families. Entries for GPCRs and NHRs from other databases (e.g. Swiss-Prot) point to these cross-reference tables. Another database dedicated to nuclear receptors, associated ligands and co-regulators is the Nuclear Receptor Signaling Atlas (NURSA) (see Weblinks). NURSA focuses primarily on orphan nuclear receptor biology, bringing together bioinformatic aspects (target information) with cheminformatic aspects (ligand information), supported by experimental data based on, for example, microarray and Q-PCR technologies.

The IUPHAR receptor database (see Weblinks) is a target/bioactivity database that currently captures GPCR information related to receptor sequence and structure, receptor classification, as curated by the IUPHAR nomenclature committee and published in Pharmacological Reviews, and the associate ligand information, which includes bioactivities in numerical form, the type of activity (e.g. agonism, allosteric modulation, etc.), together with the corresponding references. The ion channel compendium (currently available only in publication format) will be included in this database system within the next coming years.

DrugBank is a targets/drugs database [16] from the University of Alberta that combines drug target information (i.e. target sequence and structure, pathway, splice variants, etc.) with drug information (i.e. chemical structure, pharmacokinetics data, pharmacological mode of action and pharmaceutical details). The database contains over 1000 FDA-approved small molecule drugs, and over 3000 candidate drugs (i.e. under development).

The PDSP K_i database [17] is a public domain resource for psychoactive drugs and their binding properties to a large number of targets. This data warehouse captures both published and in-house K_i values or affinity, covering a large number of drugs and drug candidates binding to GPCRs, ion channels, transporters and enzymes. The user interface provides tools for customized data mining (K_i graphs, receptor and ligand selectivity mining), and is cross-linked with PubChem and PubMed [10] (if available). Other searchable fields include: receptor name, species name, tissue source, radiolabeled and tested ligand, bibliographic references as well as K_i value range.

The National University of Singapore Bioinformatics and Drug Design group (see Weblinks) develops methods, software and databases for drug discovery. Their BIDD Databases provide information about drugs, natural products, protein targets, ADME/Tox, drug-protein binding and other relevant information. The following databases are searchable online: Therapeutic Target Database [18], Drug Adverse Reaction Target Database [19], Drug ADME Associated Protein Database (321 protein entries) [20], Therapeutically Relevant Multiple Pathways Database [21], Computed Ligand Binding Energy Database [22] and Kinetic Data of Bio-molecular Interaction [23]. These databases contain cross-references to other relevant databases, associated references and ligand structures.

The Blueprint Initiative (see weblinks) is an open resource for biomolecular data focused on public databases and other software. Small Molecule Interaction Database (SMID) is a relational online database for small molecule/domain interactions, determined from the Molecular Modeling Database [MMDB (see Weblinks), hosted and maintained by NCBI]. SMID links small molecules to their protein partners and their families, giving a comprehensive picture of small molecule binding (see Weblinks). SMID further offers physico-chemical and biochemical details about the participating small and macro-molecules. The Biomolecular Interaction Network Database (BIND) is a collection of records documenting molecular interactions (see Weblinks) [24]. BIND includes high-throughput data submissions and hand-curated information from literature. SeqHound integrates biological sequence, taxonomy, annotation and 3D structures (see Weblinks). Its annotated links include Genbank (see Weblinks) [25], MMDB, MedLine [10] and BIND [24]. SeqHound is a resource for programmers, hence it features a simple web interface with limited functionality.

ChemBank (see Weblinks), implemented by the Broad Institute of Harvard and MIT, and supported by the National Cancer Institute (see Weblinks), is an open database focused on studying the effects of small molecules on biology. ChemBank features a fully developed chemical database that can be queried by compound name, sub-structure or similarity. Each query result is further expandable to information that includes chemical structure data, identifiers [e.g. CAS numbers (see Weblinks)], chemical vendors and other practical information. A list of characterized activities and observed effects are available, together with cross-references from PubMed [10].

*Ki*Bank is an open online database (see Weblinks) for *in silico* drug design from the University of Tokyo [26,27]. *Ki*Bank stores Chemical information with associated Bioactivity for given Targets, which can be queried in two search modes: by protein name/function, and by chemical name, respectively. Target-oriented queries retrieve a list with compounds that bind to the queried target. 2D and 3D structure visualizations are accessible for most compounds. In addition to chemical information, *Ki*Bank stores K_i values, species, limited experimental protocol information and PubMed [10] references. Chemistry-oriented queries retrieve a list of all targets for which bioactivity data are available on that compound. These two lists are cross-referenced, so the end-user can easily switch from chemical queries to target queries and vice versa.

ZINC is an open resource (see Weblinks) of commercially available compounds dedicated to virtual screening [28]. Compounds from the ZINC database may also be purchased directly from chemical vendors. The query can search molecular (sub)structures and several properties such as Lipinski's rule-of-five [15] criteria. Structures can be entered using the Java Molecular Editor (see Weblinks), which generates SMILES, or directly as SMILES or SMARTS. The result lists all the entries matching the query in 2D format, next to vendor identifiers and the (pre)calculated properties; 3D structures can be displayed by request. Subsets can be downloaded into the desired format, and are ready for processing with virtual screening technologies [29].

AurSCOPE is a commercial collection of annotated structure databases that capture biological and chemical information related to a given therapeutic or biopharmaceutical topic from literature, mostly patents and journals (see Weblinks). These databases capture *in vitro* and *in vivo* biological data, together with chemical information and structure-activity relationships. Complete descriptions of the biological test methods are provided (see Weblinks). AurSCOPE GPCR contains biological and chemical data relating to GPCR chemistry, pharmacology and physiology; AurSCOPE ADME/Drug–Drug Interactions contains biological and chemical information related to metabolic properties of drugs, which enables the identification of potential drug–drug interactions. AurSCOPE Ion Channel is focused on drugs described as ion channel blockers, openers, or activators, that captures all ion channels (calcium, chloride, potassium, sodium) and transmitter-gated ion channels; AurSCOPE hERG Channel database contains chemical and biological information relating to the human ether-a-go-go related gene (hERG) potassium channel.

The MediChem database from CambridgeSoft (see Weblinks) and GVK Biosciences (see Weblinks) is a commercial medicinal chemistry resource from the top medicinal chemistry journals (see Weblinks). The database captures chemical information, references (including PubMed [10] pointers) and bioactivity data (bioassay, target, activity). Assays include ADME/Tox, binding information for a target and its mutants, functional assays (e.g. cell based or *in vivo*), toxicity, etc. Records can be queried by target platform, for example, kinase, GPCR, NHR, etc.

The Merck Index is a structure searchable encyclopedia of chemicals, drugs and biologic active compounds. More than 10,000 monographs on single substances and related groups of compounds cover chemical, generic and brand names. Searchable fields include structures and stereochemistry, registry numbers, physical properties, toxicity information, therapeutic uses and literature.

MDL DiscoveryGate (see Weblinks) is web-enabled discovery environment, which integrates, indexes and links different sources of scientific information to give immediate access to compounds and related property data, reactions, original journal articles and patents, and authoritative reference works on synthetic methodologies. The strong advantage of this commercial database is that it is highly integrated making the querying and navigation between disparate types of data (e.g. chemical structures and scientific literature) easy. The system affords getting comprehensive data with a single query across multiple integrated databases such as searching an online structure and data index of chemical compounds, identifying and organizing similar chemical structures and related data, finding bioactivity data for a particular compound, finding chemical suppliers for compounds of interest, reviewing methods for synthesizing compounds, viewing reported toxic effects and metabolism pathways, and linking from literature references directly to the original publications or patent documents. The related MDL Patent Chemistry Database indexes chemical reactions, substances and substance related information from organic chemistry and life sciences patent publications.

PathArt is a pathway database from Jubilant Biosys that dynamically builds molecular interaction networks from curated databases. This product has comprehensive information on over 900 regulatory as well as signaling pathways that allows users to upload and map microarray expression data onto the pathways. GPCR Annotator captures a wide range of therapeutically relevant areas related to GPCRs; the annotator module allows users to classify the GPCR family hierarchy from sequence input (see Weblinks). Kinase ChemBioBase, co-distributed by Accelrys is a comprehensive database, currently containing compounds active on more than 400 kinases. Drug Database captures approved drugs. GPCR Ligand Database captures small molecule GPCR agonists/antagonists from journal articles and patents, covering GPCRs. Protease Inhibitors Database includes small molecule protease inhibitors active on various proteases, captured from journal articles and patents.

Sunset Molecular Discovery LLC (see Weblinks) integrates knowledge from target-driven medicinal chemistry with clinical PK data in the World of Molecular Bioactivity

(WOMBAT)-PK database (see Weblinks), and provides up-to-date coverage of the medicinal chemistry literature in the WOMBAT database, as it appears in peer-reviewed journals. WOMBAT (current release 2006.2) contains biological activities on more than 1400 unique targets (GPCRs, ion channels, enzymes and proteins). The information is curated from more than thousands of papers published in medicinal chemistry journals. Additional experimental properties and calculated descriptors are available, as well as a comprehensive set of keywords related to biology and experimental protocols. Approximately 87% of the targets have SwissProt primary accession numbers; all indexed papers contain the DOI identifiers and/or have a direct cross-reference to the URL providing the original paper in PDF format. WOMBAT-PK 2006.1, the WOMBAT Database for Clinical pharmacokinetics (PK), captures clinical PK measurements for hundreds of drugs. Clinical data and physico-chemical properties for launched drugs are captured from multiple literature sources. Both databases are available in the RDF format.

Are these technologies evolving?

Integrated resources, where bioinformatics and cheminformatics data are seamlessly converging into a comprehensive picture are becoming reality. The integration process itself requires hierarchical classification schemes, such that knowledge related to target-focused chemical libraries and biological target families can be mined simultaneously [30]. Annotated and integrated bioactivity databases are becoming the *de facto* second-generation chemical databases [31]. As hierarchical classification schemes for biological and chemical entities mature, extracting the comprehensive knowledge from such databases becomes easier [2]. The major task for discovery scientists remains data analysis and interpretation, resulting in knowledge creation. This requires familiarity with fundamental principles in both chemistry and biology, and certainly skills with complex queries. Learning how to proper query such disjoint sources to answer complex knowledge discovery type questions might be a challenge, but the major hurdles of the past, that is, data collection and, recently, integration, are fading. The age of informatics-driven pharmaceutical discovery has arrived.

Weblinks

- AurSCOPE, Aureus Pharma: http://www.aureus-pharma.com/Pages/Products/Aurscope.php
- Aureus Pharma, Paris, France: http://www.aureus-pharma.com/
- Bioinformatics & Drug Design group, Computational Science Department, National University of Singapore: http://bidd.nus.edu.sg/
- The Biomolecular Interaction Network Database and related tools 2005 update (Alfarano *et al.*, 2005) *Nucleic Acids Res.* 33 (Database issue), D418-D424: http://bind.ca/
- The Blueprint Initiative, Samuel Lunenfeld Research Institute, Toronto: http://www.blueprint.org/
- CambridgeSoft Corporation, Cambridge, USA: http://www.cambridgesoft.com/
- CambridgeSoft Corporation, Chemical Database: http://www.cambridgesoft.com/databases/
- CAS online/SciFinder, American Chemical Society: http://www.cas.org/SCIFINDER/
- ChemBank project, Broad Institute, Cambridge: http://chembank.broad.harvard.edu/
- Chemical Abstracts Service, CAS Registry, American Chemical Society http://www.cas.org/EO/regsys.html
- Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems (Horn *et al.*, 2001) *Nucleic Acids Res.* 29, 346–349: http://www.receptors.org/NR/
- CORBA – Common Object Request Broker Architecture, Object Management Group, Inc.: http://www.corba.org/; http://www.omg.org/
- CrossFire Beilstein database, Elsevier MDL: http://www.beilstein.com/
- The Digital Object Identifier System, The International DOI Foundation: http://www.doi.org/
- DiscoveryGate, Elsevier MDL: https://www.discoverygate.com/
- EBI–European Bioinformatics Institute: http://www.ebi.ac.uk/; an exhaustive list of available databases can be found at: http://www.ebi.ac.uk/Databases/
- *Entrez*, The Life Sciences Search Engine: http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
- *Entrez* Protein: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
- *Entrez* PubMed, National Library of Medicine: http://www.ncbi.nlm.nih.gov/entrez/
- Enzyme Structures Database, European Bioinformatics Institute: http://www.ebi.ac.uk/thornton-srv/databases/enzymes/
- ExPASy Proteomics Server/Swiss-Prot Protein knowledgebase, Swiss Institute of Bioinformatics: http://www.expasy.org/sprot/
- Extensible Markup Language (XML), World Wide Web Consortium (W3C): http://www.w3.org/XML/
- GenBank: update (Benson *et al.*, 2004) *Nucleic Acids Res.* 32 (Database issue), D23-D26: http://www.ncbi.nih.gov/Genbank/
- GPCRDB: an information system for G protein-coupled receptors (Horn *et al.*, 1998) *Nucleic Acids Res.* 26, 275–279: http://www.gpcr.org/7tm/
- GVK Biosciences Private Limited, Hyderabad, India: http://www.gvkbio.com/
- HyperText Markup Language (HTML), World Wide Web Consortium (W3C): http://www.w3.org/MarkUp/
- International Union of Biochemistry and Molecular Biology (IUBMB), Enzyme Classification: http://www.chem.qmul.ac.uk/iubmb/
- Java Database Connectivity (JDBC), Java Technologies, Java SE: http://java.sun.com/javase/technologies/database/index.jsp
- Java Molecular Editor, Molinspiration Cheminformatics: http://www.molinspiration.com/jme/
- Jubilant Biosys Ltd., Products: http://www.jubilantbiosys.com/products.htm
- *Ki*Bank, Quantum Molecular Interaction Analysis Group, Institute of Industrial Science, University of Tokyo: http://kibank.iis.u-tokyo.ac.jp/
- Kyoto Encyclopedia of Genes and Genomes (KEGG): http://www.genome.ad.jp/kegg/
- MDDR – MDL Drug Data Report, Elsevier MDL: http://www.mdli.com/products/knowledge/drug_data_report/
- Membrane Transport Proteins, Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB): http://www.chem.qmul.ac.uk/iubmb/mtp/

- MMDB - The Molecular Modeling Database, National Center for Biotechnology Information, The NCBI Structure Group: http://www.ncbi.nlm.nih.gov/Structure/
- National Cancer Institute, National Institutes of Health: http://www.cancer.gov/
- NIH Molecular Libraries Initiative (Austin *et al.*, 2004) *NIH Molecular Libraries Initiative. Science* 306, 1138–1139: http://nihroadmap.nih.gov/
- Nomenclature Committee of the International Union of Pharmacology (NC-IUPHAR), Receptor Compendium: http://www.iuphar-db.org/
- Nuclear Receptor Signaling Atlas: http://www.nursa.org/index.cfm
- PDBRTF Database, Chemogenomics Lab Research Unit, Institute Municipal D'Investigacio Medica, Barcelona: http://cgl.imim.es/pdbrtf/
- The Physician Desk Reference, 2003, Thomson Healthcare: http://www.pdr.net/
- The Protein Data Bank (Berman *et al.*, 2000) *Nucleic Acids Res.* 28, 235–242: http://www.rcsb.org/pdb/
- Prous Science, Barcelona, Spain: http://www.prous.com/
- PubChem: http://pubchem.ncbi.nlm.nih.gov/
- PubChem Bioassay: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pcassay
- PubChem Chemical Structure search: http://pubchem.ncbi.nlm.nih.gov/search/
- SeqHound database, The Blueprint Initiative: http://www.blueprint.org/seqhound/index.html
- SMID - Small Molecule Interaction Database, The Blueprint Initiative: http://smid.blueprint.org/
- Sunset Molecular Discovery LLC, Products: http://www.sunsetmolecular.com/products/
- TCDB – Transport Classification Database: http://www.tcdb.org/
- Uniform Resource Identifiers (URI): Generic Syntax – Draft Standard RFC 2396, 1998: http://www.ietf.org/rfc/rfc2396.txt
- UniProtKB/Swiss-Prot Protein Knowledgebase Database, Swiss Institute for Bioinformatics and the European Bioinformatics Institute: http://www.ebi.ac.uk/swissprot/
- WDI - World Drug Index, Derwent Publications Ltd.: http://thomsonderwent.com/products/lr/wdi/
- ZINC, University of California San Francisco: http://blaster.docking.org/zinc/

Acknowledgments

This work was supported in parts by National Institutes of Health grant U54 MH074425-01 (National Institutes of Health Molecular Libraries Screening Center Network) and by the New Mexico Tobacco Settlement Fund (T.I.O.) as well as by the National Institutes of Health planning grant P20-RR20751 (AT). The authors express their gratitude to Dr Marius Olah (Sunset Molecular Discovery LLC) and to Dr Dan C. Fara (University of New Mexico) for their help with editing the manuscript and their valuable suggestions.

Disclosure: Tudor Oprea is the founder and CEO of Sunset Molecular Discovery LLC, which develops and markets the WOMBAT and WOMBAT-PK databases.

References

1 Olah, M. and Oprea, T.I. (2006) Bioactivity databases. In *Comprehensive Medicinal Chemistry II*, (Vol. 3) (Taylor, J.B. and Triggle, D.J., eds) pp. 293–313, Elsevier

2 Wisniewski, J.L. (2003) Chemical nomenclature and structure representation: algorithmic generation and conversion. In *Handbook of Cheminformatics*, (Vol. 2) (Gasteiger, J., ed.), pp. 51–79, Wiley-VCH

3 Fisanick, W. and Shively, E.R. (2003) The CAS information system: applying scientific knowledge and technology for better information. In *Handbook of Cheminformatics*, (Vol. 2) (Gasteiger, J., ed.), pp. 556–607, Wiley-VCH

4 Olah, M. *et al.* (2005) WOMBAT: world of molecular bioactivity. In *Chemoinformatics in Drug Discovery*, (Vol. 23) (Oprea, T.I., ed.), pp. 223–239, Wiley-VCH

5 Neubig, R.R. *et al.* (2003) International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on terms and symbols in quantitative pharmacology. *Pharmacol. Rev.* 55, 597–606

6 Oprea, T.I. *et al.* (2005) Rapid ADME filters for lead discovery. In *Molecular Interaction Fields*, (Vol. 24) (Cruciani, G., ed.), pp. 249–272, Wiley-VCH

7 Horn, F. *et al.* (2001) Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res.* 29, 346–349 (http://www.receptors.org/NR/)

8 Horn, F. *et al.* (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* 26, 275–279 (http://www.gpcr.org/7tm/)

9 Bright, M.W. *et al.* (1992) A taxonomy and current issues in multidatabase systems. *Computer* 25, 50–60

10 Larson, J.A. (1995) *Database Directions: From Relational to Distributed, Multimedia, and Object-Oriented Database Systems.* Prentice-Hall PTR, New Jersey pp. 45–56

11 Siegel, J. (1996) *CORBA Fundamentals and Programming.* John Wiley & Sons, Inc., New York

12 Powell, T.A. (1999) *HTML: The Complete Reference* (2nd edn), Osborne/McGraw-Hill, Berkeley

13 Berman, H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.* 28, 235–242 (http://www.rcsb.org/pdb/)

14 Austin, C.P. *et al.* (2004) NIH molecular libraries initiative (Austin *et al.*, 2004). *Science* 306, 1138–1139 (http://nihroadmap.nih.gov/)

15 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

16 Wishart, D.S. *et al.* (2006) DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* 34 (Database issue), D668–D672

17 Roth, B.L. *et al.* (2000) The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* 6, 252–262

18 Chen, X. *et al.* (2002) TTD: therapeutic target database. *Nucleic Acids Res.* 30, 412–415

19 Ji, Z.L. *et al.* (2003) Drug adverse reaction target database (DART): proteins related to adverse drug reactions. *Drug Saf.* 26, 685–690

20 Sun, L.Z. *et al.* (2002) Absorption, distribution, metabolism, and excretion-associated protein database. *Clin. Pharmacol. Ther.* 71, 405–416

21 Zheng, C.J. *et al.* (2004) TRMP: a database of therapeutically relevant multiple pathways. *Bioinformatics* 20, 2236–2241

22 Chen, X. *et al.* (2002) CLiBE: a database of computed ligand binding energy for ligand-receptor complexes. *Comp. Chem.* 26, 661–666

23 Ji, Z.L. *et al.* (2003) KDBI: kinetic data of bio-molecular interactions database. *Nucleic Acids Res.* 31, 255–257

24 Alfarano, C. *et al.* (2005) The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.* 33 (Database issue), D418–D424 (http://bind.ca/)

25 Benson, D.A. *et al.* (2004) GenBank: update. *Nucleic Acids Res.* 32 (Database issue), D23–D26 (http://www.ncbi.nih.gov/Genbank/)

26 Aizawa, M. *et al.* (2004) KiBank: a database for computer-aided drug design based on protein-chemical interaction analysis. *Yakugaku Zasshi: J. Pharm. Soc. Jpn.* 124, 613–619

27 Zhang, J.-W. *et al.* (2004) Development of KiBank, a database supporting structure-based drug design. *Comput. Biol. Chem.* 28, 401–407

28 Irwin, J.J. *et al.* (2005) ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45, 177–182

29 Oprea, T.I. *et al.* (2004) Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* 8, 349–358

30 Cases, M. *et al.* (2005) Chemical and biological profiling of an annotated compound library directed to the nuclear receptor family. *Curr. Top. Med. Chem.* 5, 763–772

31 Savchuck, N.P. *et al.* (2004) Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr. Opin. Chem. Biol.* 8, 412–417