*Send Orders of Reprints at reprints@benthamscience.net*

*Current Topics in Medicinal Chemistry*, 2013, *13*, 000-000 1

# Open Source Software and Web Services for Designing Therapeutic Molecules

Deepak Singla[1,2], Sandeep Kumar Dhanda[1], Jagat Singh Chauhan[1], Anshu Bhardwaj[3], Samir K. Brahmachari[3,4], Open Source Drug Discovery Consortium[3] and Gajendra P.S. Raghava[1,*]

[1]*Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India;* [2]*Centre for Microbial Biotechnology, Panjab University, Chandigarh, India;* [3]*CSIR-Open Source Drug Discovery Unit, New Delhi, India;* [4]*CSIR-Institute of Genomics and Integrative Biology, New Delhi, India*

**Abstract:** Despite the tremendous progress in the field of drug designing, discovering a new drug molecule is still a challenging task. Drug discovery and development is a costly, time consuming and complex process that requires millions of dollars and 10-15 years to bring new drug molecules in the market. This huge investment and long-term process are attributed to high failure rate, complexity of the problem and strict regulatory rules, in addition to other factors. Given the availability of 'big' data with ever improving computing power, it is now possible to model systems which is expected to provide time and cost effectiveness to drug discovery process. Computer Aided Drug Designing (CADD) has emerged as a fast alternative method to bring down the cost involved in discovering a new drug. In past, numerous computer programs have been developed across the globe to assist the researchers working in the field of drug discovery. Broadly, these programs can be classified in three categories, freeware, shareware and commercial software. In this review, we have described freeware or open-source software that are commonly used for designing therapeutic molecules. Major emphasis will be on software and web services in the field of chemo- or pharmaco-informatics that includes *in silico* tools used for computing molecular descriptors, inhibitors designing against drug targets, building QSAR models, and ADMET properties.

**Keywords:** Open source drug discovery, QSAR models, software, machine learning techniques, chemoinformatics, pharmacoinformatics.

## INTRODUCTION

The development of drug molecules for the treatment and prevention of diseases has played a critical role in the field of medicine. Due to advancement in the field of medicine, quality of life and average life expectancy have increased significantly at the last century. In the nineteenth and earlier centuries, drugs were derived mainly from medicinal plants in addition to natural extracts derived from animal species. In order to improve qualities of drugs based on natural extracts, process of isolation of pure biologically active molecules began in later part of nineteenth century. History of modern drug discovery research is not older than a century [1]. Generally, drug discovery process can be divided into three phases; first period belongs to nineteenth century where the drugs were discovered from serendipit [2, 3]. Second period starts in the early 20$^{th}$ century with the discovery of antibiotics [4]. The modern day rational drug discovery and development approach can broadly be classified into three phases 1) Research phase 2) Clinical Phase 3) Applied phase. Research Phase is marked by target identification and validation, hit and lead identification and optimization (Fig. **1**). Clinical phase is characterized by evaluating the efficacy and safety of the new chemical entity first in animal models and then human population followed by approval to market it. The applied phase composed of post marketing surveillance.

Recent study has shown that approximately US$800 million is the estimated cost for bringing a new drug into the market through passing these phases [5]. Therefore, pharmaceutical companies are seeking new ways to reduce this cost and increasing profit margins. Cheminformatics is providing an alternative for reducing huge money investment to nearly half in drug discover [6]. The recently approved drug Indinavir and earlier Haloperidol have shown up a way to *in silico* approaches in drug designing [7, 8].

The computational methods for *in silico* drug discovery have been broadly categories into two fields bioinformatics and cheminformatics. In case of bioinformatics, major emphasis is on identification and validation of drug targets, mainly based on functional/structural annotation of genomes. The field of bioinformatics is dominated by freewares where thousands of databases like UniProt [9], PDB [10], NCBI (http://www.ncbi.nlm.nih.gov/), and software/webservers like BLAST [11], GPCRpred [12], FASTA [13], ATPINT [14], Clustal-W [15], ESLPRED2 [16], Psipred [17] are available free for public use. These bioinformatics resources are heavily used by community for identification and designing of drug targets in addition to functional annotation of genomes. In case of cheminformatics or pharmacoinformat-

*Address correspondence to this author at the Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India; Tel: +91-172-2690557; Fax: +91-172-2690632; E-mail: raghava@imtech.res.in

1568-0266/13 $58.00+.00 

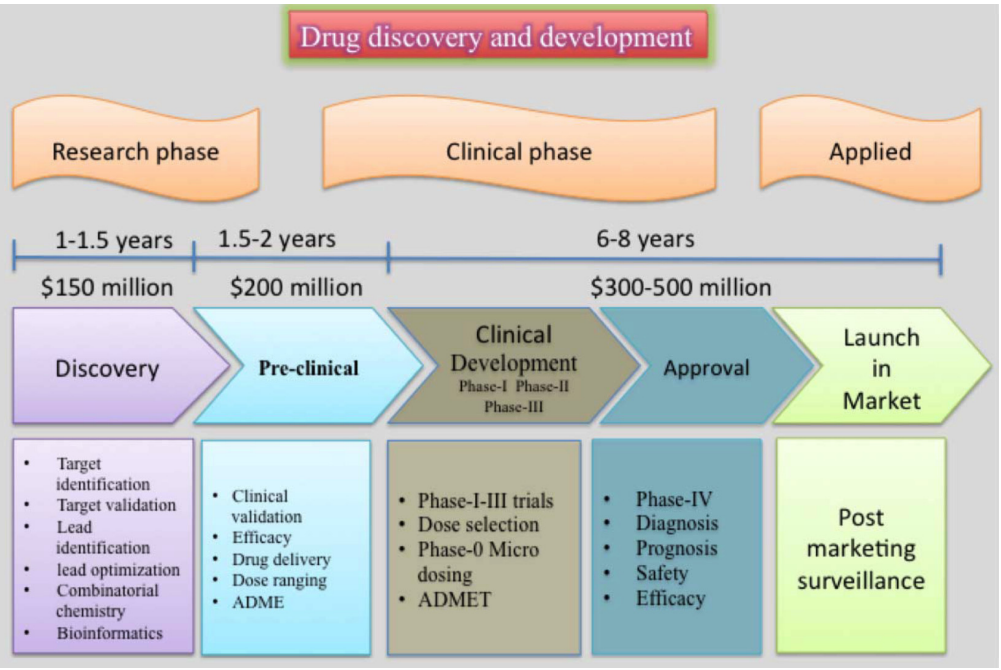


**Fig. (1).** Flowchart of drug discovery – The amount of fund required depends on the success rate at the clinical trial stage.

ics, major emphasis is on designing of drug molecules or ligands and their interaction with drug targets. In contrast to bioinformatics, cheminformatics is dominated by proprietary or commercial software/webservers where software are costly with stringent license conditions. Due to heavy cost of chemoinformatics resources, computer-aided drug discovery is still a costly affair.

Despite the existence of a large number of computational methods which are available to evaluate and priortize inhibitors, very few are ready for public use. Some of those which are published are not easy to use due to lack of a webserver or standalone package. These issues can only be resolved efficiently by enhancing collaborations in an open and free knowledge sharing environment. Recent research also suggests that open collaborative drug discovery will be the future paradigm of biomedical research [18-20]. In order to overcome the limitations of existing approaches, open source/ freely available software have been developed by different organizations like OpenTox [21], OSDD (Open Source Drug Discovery), CDD [22], Blue Obelisk [23] etc. In past, number of reviews had been published in this area. In this review, our major focus is on software that are freely accessible and could be used at different stages of the drug discovery process (Fig. **2**). Our focus will be on chemoinformatics or pharmacoinformatics related software that are used for designing drug molecules/ligands/inhibitors. In principle the success of designing a drug which has high selectivity and specificity to a given functional target is the key challenge in order to have a potential drug with least possible side effects. Also, the present tools have limited predictive capacity for estimating pharmacokinetic parameters which are imperative to have an ideal drug. We have covered the following topics in this review.

- **Source of Molecules**: Resources in the field of drug design like databases on chemicals assays/properties, drug molecules. These resources are used for developing various models for predicting inhibitors.
- **Molecular Editors**: This topic will cover software and web services for drawing new molecules and for editing of existing molecules. This topic will also include tools used for visualization of molecules.
- **Analog Generators**: In this section, we will describe software used to generate analogs of molecules. It will also include software used to generate the virtual library of chemicals.
- **Structure Optimization**: Software used for generating 2D/3D structure, and for optimization of energy/geometry of molecules will be covered under this topic.
- **Molecular Descriptors**: Calculation of molecular descriptors is fundamental requirement for developing QSAR models. We will describe the software available for the same.
- **Similarity Search**: This topic will describe software or web services which are frequently used to perform chemical similarity search.
- **QSAR/QSPR Models**: Software used for developing models like QSAR, QSPR, QSTR will be described.
- **Chemical Clustering**: Classification and clustering of small molecules is important to understand property of a scaffold, major free software will be described in this section.
- **Molecular Docking**: This topic will describe commonly used software packages for docking small molecules in macromolecules.
- **Pharmacophore Tools**: In this section, we will cover resources important for pharmacophore search.

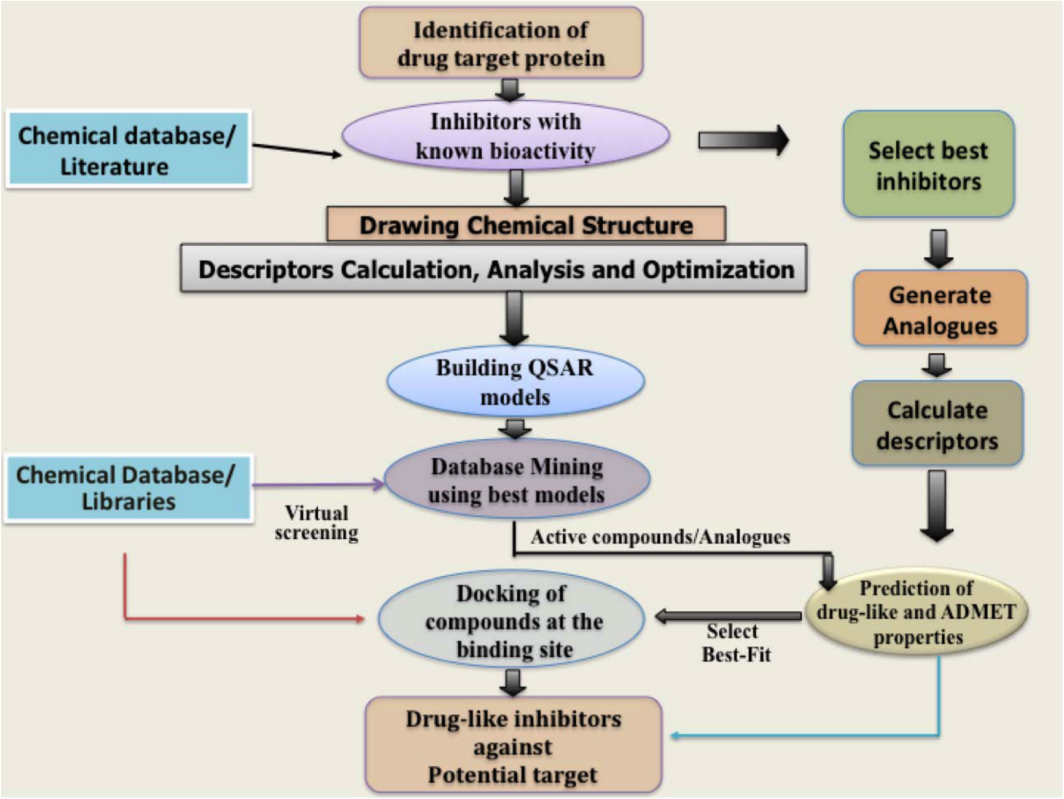


**Fig. (2).** An overview of the workflow of *in silico* drug designing process.

- **ADMET Techniques**: Software/webservers used for computing drugability of molecules, and ADMET properties will be described.
- **Drug Target Prediction**: This section describes the softwares and webservice important in prediction of drug targets.
- **Designing of Inhibitors**: This topic describes different tools that allow users to predict inhibitors against a target. These tools generally used diverese techniques (like QSAR models, docking, screening) for designing inhibitors.
- **Other resources**: Miscellaneous major resources over the internet that are serving community will be covered under this topic.
- **Major Initiatives**: Numerous organizations and groups working towards affordable drugs will be covered in this section.
- **Future Prospects**: This section will describe forthcoming prospects of open source in drug discovery including limitations of existing resources.

## SOURCE OF MOLECULES

In ancient times, natural products obtained from various sources like plants were used to test whether they have biological activity. These natural products were major sources for discovering therapeutics. Subsequently, synthetic chemists have synthesised large number of chemical compounds and generated library of synthetic molecules in the last century. Presently, there are number of databases and repositories that are managing comprehensive information about millions of chemicals. This section describes major source of molecules that are freely available for public use. Chemical databases/resources are the backbone of computer-aided drug discovery, whether it is chemoinformatics or pharmacoinformatics or bioinformatics. These databases provide information that is used to build knowledge-based models for discovering and designing drug molecules. Here, we have covered major databases that are available free for public use, (Table **1**) provides brief description of each database. A number of commercial/in-house databases such as WOMBAT [24], World Drug Index (http://www.daylight.com/products/wdi.html) etc. are into existence for a long time. However, more recently the availability of molecule databases such as PubChem [25, 26], Zinc [27], ChEMBL [28] has dramatically changed the landscape of publicly available cheminformatics resources. Some of the databases like PubChem BioAssay, ChEMBL also encapsulates the information regarding the target protein, organism on which the chemical is effective and sometimes along with their activity score. In the foreground, we have described major databases in brief, these database are the backbone for developing models for drug discovery (Table **1**).

PubChem project is an open public repository and maintain three types of information, namely, substance, compound and BioAssays [25, 26]. PubChem Substance contains original chemical structure submitted by different vendors, publishers or other government agencies. PubChem Compound maintains the index of unique chemical structures present in PubChem Substance. PubChem BioAssay currently contains information about 500,000 assays, covering

5000 protein targets, 30,000 gene targets and providing over 130 million bioactivity outcomes [25]. ChEMBL is a manually curated database that provides comprehensive information about 1 million bioactive (small drug-like molecules) compounds with 8200 drug targets [28]. This database also contains the different dataset for neglected diseases like malaria from both commercials as well as academics sources. ZINC database maintains information about all the commercially available compounds. This database contains 21 million compounds available for virtual screening. The most important feature of this database is that users can filter data using various features like molecular weight, logP etc., which leads to a smaller dataset with most relevant properties.

**Table 1. Databases and Resources Managing and Hosting Chemical Compounds**

| Database | Brief Description with URL |
|---|---|
| PubChem | A comprehensive database of bioassays, compounds and substances (http://pubchem.ncbi.nlm.nih.gov/) |
| ChEMBL | Database of drug like molecules (https://www.ebi.ac.uk/chembldb) |
| Zinc | Maintain commercially-available compounds for virtual screening (http://zinc.docking.org/) |
| ChemDB | Collection of small-molecules (http://cdb.ics.uci.edu/) |
| ChemSpider | A chemical database (http://www.chemspider.com/) |
| MMsINC | Commercial compounds (http://mms.dsfarm.unipd.it/MMsINC/) |
| KEGG | Maintain comprehensive information (http://www.genome.jp/kegg/) |
| SMPDB | Small molecule Pathway database (http://www.smpdb.ca) |
| HMDB | Human Metabolites (http://www.hmdb.ca/) |
| PDBeChem | Dictionary of chemical components refered in PDB entries (http://www.ebi.ac.uk/pdbe-srv/pdbechem/) |
| PDB-Bind | Binding affinity information for PDB Ligands (http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp) |
| BindingDB | Binding affinity of PDB Ligands (http://www.bindingdb.org/) |
| NCI | Small molecules related to cancer (http://cactus.nci.nih.gov/ncidb2.1/) |
| CDD | Collaborative drug discovery (https://www.collaborativedrug.com/) |
| DrugBank | All kind of drugs (http://www.drugbank.ca/) |
| HMRbase | Hormones and their Receptors (http://crdd.osdd.net/raghava/hmrbase/) |
| BIAdb | Benzyl-isoquinaloid alkaloids (http://crdd.osdd.net/raghava/biadb/) |
| NPACT | Plant derived natural compounds (http://crdd.osdd.net/raghava/npact/) |
| SuperNatural | A searchable database of available natural compounds |
| HIT | Herb ingredients targets (http://lifecenter.sgst.cn/hit/) |
| Drugs@FDA | FDA approved drug products (http://www.fda.gov/Drugs/) |

ChemDB is a database of commercially available small molecules. it contains around five million chemicals [29]. This database provides different types of information of chemicals that includes predicted or experimentally determined physicochemical properties, such as 3D structure, melting temperature and solubility. ChemSpider contains more than 28 million unique chemical entities aggregated from more than 400 diverse data sources [30]. Each structure entry in ChemSpider is associated with a list of predicted molecular properties as well as possibly available experimental data, spectra etc. It has also been integrated with the SureChem (http://surechem.com/) patent database collection of structures to facilitate structure-based linking to patents between the two data collections. NCI database has more than 275,000 small molecule structures, a very useful resource for researchers working in the filled of cancer/AIDS [31]. In addition to big databases, there are some databases that maintain specialized information. These databases maintain chemical compounds information about their role in the biological system like KEGG contains association of chemicals in pathway and diseases [32]. Similarly, number of databases maintain interaction of target-ligand interaction that is essential for target based drug discovery [33, 34].

## MOLECULAR EDITORS

Molecular editors are commonly used tools in the field of cheminformatics, to draw and manipulate chemical structures. These tools provide a number of facilities like geometry optimization, structure visualization, energy minimization. There are several software packages available, which allow users to sketch a molecular diagram on a computer. Comprehensive list of molecular editors is given in (Table **2**). BKchem (http://bkchem.zirael.org/) is a free software written in Python. It works on major operating systems like Linux, WinXP and MacOS X. It allows users to draw, edit, visualize the molecules and provide various options to the users. ChemSketch (http://www.acdlabs.com/resources/freeware/chemsketch/) is a free comprehensive chemical drawing package that allows users to draw chemical structures including organics, organometallics, polymers, and Markush structures. Free version of ChemSketch has limited facilities, and it is only available for Windows platform. JChemPaint (http://jchempaint.github.com/) is a java-based open-source software developed for drawing, editing and viewing 2D chemical structure. This software developed using the Chemistry Development Kit (CDK). XDrawChem (http://xdrawchem.sourceforge.net/) is a two-dimensional molecule drawing program, it is an open-source software. This software can read molecules in various formats and can create images in popular formats like PNG, EPS. JME Molecular Editor (http://www.molinspiration.com/jme/) is a software developed to draw, edit, and view molecules and reactions. It is a java based software available free for non-commercial user, available in a stand-alone mode or as an applet for integrating in the web page.

**Table 2. List of Major Molecular Editors, Frequently Used for Drawing and Editing Molecules**

| Editors | Brief Description |
|---|---|
| BKchem | Python based free 2D molecule editor (http://bkchem.zirael.org/) |
| Chem-Sketch | ACD/ChemSketch Freeware is a free software for drawing chemicals (http://www.acdlabs.com/resources/freeware/chemsketch/) |
| JChemPaint | Editor for 2D chemical structures (http://jchempaint.github.com/) |
| Accelrys Draw | Draw and edit complex molecules, no fee for academic community (http://accelrys.com/products/informatics/cheminformatics/draw/index.html) |
| XDraw-Chem | Molecule drawing program (http://xdrawchem.sourceforge.net/) |
| MedChem Designer | Drawing molecules and integration with ADMET property. (http://simplus-downloads.com/) |
| JME | JME Molecular Editor (http://www.molinspiration.com/jme/) |
| PubChem Sketcher [117] | A web-based tool for sketching, integrated in PubChem (http://pubchem.ncbi.nlm.nih.gov/edit2/index.html) |

**Table 3. Analogs Generation Softwares**

| Software | Brief Description |
|---|---|
| Library synthe-sizer | Virtual chemical enumeration (http://tripod.nih.gov/?p=370) |
| SmiLib [118] | Enumerates combinatorial libraries with very high rate (http://gecco.org.chemie.uni-frankfurt.de/smilib/) |
| GLARE [119] | Generate combinatorial library (http://glare.sourceforge.net/) |
| CLEVER [120] | Chemical Library Editing, Visualization and Enumerating Resource (http://datam.i2r.a-star.edu.sg/clever/) |
| Newlead [121] | Generate of combinatorial library from bioactive conformations (http://www.ccl.net/cca/software/MAC/index.shtml) |

## ANALOG GENERATORS

Virtual library generation approaches have a major impact on drug designing process where small therapeutic molecules are generated from basic scaffolds with attachments sites and lists of R-groups. In (Table **3**), we briefly summarized software packages available for the combinatorial enumeration of virtual chemical compound libraries. Some of these packages are available in open source while others are commericialy available software packages. All these software packages are based on similar methodology to generate virtual chemical libraries. Basically, all these packages required three basic substructures: core scaffolds, linkers and building blocks (R-groups).

The most commonly used combinatorial chemistry and analogs designing tool is SmiLib. It is a freely available Linux based chemoinformatics tool which can be used as command line and graphical user inferface for generating combinatorial library. It requires three substructures: Scaffolds, Building blocks and Linkers to generate a combinatorial library. GLARE (Global Library Assessment of REagents) is an Open Source package to generate the combinatorial library. CLEVER (chemical library editing, visualizing and enumerating resource) is a free chemoinformatics tool that enumerates chemical libraries using customized fragments; it also computes the physicochemical properties of the generated compounds. Another tool is Library synthesizer, an open-source java based tool for chemical library enumeration and profiling. NEWLEAD is also used for the automatic generation of combinatorial library from bioactive conformations of reference molecules. The input for this software is a set of fragments in the 3D orientation corresponding to a given pharmacophore model. ORganic VIrtual Library (ORVIL) is a perl program to generate the combinatorial library organic substituents without using scaffold hopping. It is designed to explore the organic chemical space in the given query structure without affecting the entire backbone of the molecule enabling minimum molecular complexities.

## STRUCTURE OPTIMIZATION

The pharmacological properties of small therapeutics molecules depend on their specific conformation. In chemoinformatics, various techniques such as structure-based screening, ligand-based screening, molecular modeling and molecular docking requires suitable multi-conformer stuctures (2D/3D) of compounds to facilitate the drug discovery process. The compound conformers are the basic need in medicinal chemistry to explore more complex structural motifs and different topologies, because there is a relationship between different conformers and function [35-37]. There are several chemoinformatics tools as described in (Table **4**), are available to generate 2D/3D structure/conformers like Openbabel [38], Frog [39], Balloon [40].

Openbabel is a free software, used in the inter-conversion of chemical structure/conformers (2D/3D) and different chemical file formats, substructure search, force field calculation, extraction of stereochemical information and fingerprint calculation. It is available for different platforms like Window, Linux and Mac. FROG (FRee Online druG) is a free online drug conformation generation tool for small molecules starting from their 1D or 2D descriptions. FROG also identifies the different unambiguous isomers corresponding to each molecule. Smi23d (3D Cordinate Generation) program converts one or more SMILIES strings into 3D. It uses two-stages, which it first generates the rough co-

ordinates then it optimizes and refines the final coordinates by *mengine* program. Cyndi is fast and powerful structure conformation generation package based on the multi-objective evolution algorithm. It is capable of generating geometrically diverse conformers at the large scale. It has an option to remove the redundant geometrical conformers with the RMSD filter and finally optimize remaining conformers with energy minimization.

**Table 4. List of Software and Web Servers Used for Structure Optimization of Molecules**

| Software | Brief Description |
|---|---|
| Balloon | Conformer Ensembles (http://web.abo.fi/~mivainio/balloon/index.php) |
| MOPAC | Semiempirical quantum chemistry program (http://openmopac.net/) |
| Openbabel [38] | The Open Source Chemistry Toolbox (http://openbabel.org/) |
| Frog [39] | Generation of free online drug conformation (http://bioserv.rpbs.jussieu.fr/cgi-bin/Frog) |
| DG-AMMOS [41] | Generate 3D conformation using distance geometry (http://www.mti.univ-paris-diderot.fr/fr/downloads.html) |
| SMI23D [122] | Generation of 3D (http://www.chembiogrid.org/cheminfo/smi23d/) |
| Cyndi [123] | Generate geometrically extended or compact conformations (http://www.biomedcentral.com/1471-2105/10/101/additional/) |
| TINKER [124] | Software Tools for Molecular Design (http://dasher.wustl.edu/tinker/) |

Balloon generates 3D atomic coordinates using molecular connectivity via distance geometry. It uses multi-objective genetic algorithm for generating of 3D conformers. DG-AMMOS is an open-source program, which allows the generation of the 3D conformation of small molecules using distance geometry and their energy minimization. AMMP force field sp4 is used in implementing DG-AMMOS [41]. TINKER is a widely used molecular modeling software having several features such as molecular dynamics, minimization and conformational sampling. It generates structure conformers by unconstrained molecular dynamics and each conformer is simulated and energy minimized. Molecular Orbital PACkage (MOPAC) is an open source semi-empirical quantum chemistry program that is used to study molecular structures and reactions. molecular orbital package. It is one of the old and famous program used in the field of quantum chemistry.

## MOLECULAR DESCRIPTORS

Molecular descriptors repersent the characterstics or features of a molecule in numerical values [42]. Descriptor can be defined as an outcome of logical procedure where chemical information is represented in the form of some values or numbers for a property in considerations [43]. A key steps in classical quantitative structure-activity/property relationship (QSAR/QSPR) modeling is the encoding of a chemical compound into a vector of numerical descriptors. These molecular descriptors may be result of some experiment, for example logP and are highly correlated with that property of chemicals. Based on these descriptors QSAR/QSPR model are developed, which are helpful in designing new chemical entity (NCE) having the property similar to used dataset [44]. Today, huge numbers of software are available in public domain to calculate molecular descriptor, some of which are listed in (Table **5**).

**Table 5. Important Software and Webserver for Computing Molecular Descriptors**

| Software | Brief Description |
|---|---|
| Joelib | Descriptor calculation software (http://sourceforge.net/projects/joelib) |
| Afgen | Fragment-based descriptors (http://glaros.dtc.umn.edu/gkhome/afgen/overview) |
| ISIDA-fragmentor | Calculate of Substructural Molecular Fragments and ISIDA Fragments (http://infochim.u-strasbg.fr/spip.php?rubrique49) |
| ODDesri-potrs | Simple java-based command level tool for descriptor calculation (http://www.softpedia.com/get/Science-CAD/ODDescriptors.shtml) |
| MOLD2 [45] | Calculating descriptors from a two-dimensional chemical structure (http://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2/default.htm) |
| PowerMV [125] | Window based calculation of descriptors (http://nisla05.niss.org/PowerMV/index.html) |
| PaDEL [126] | Fingerprints calculation (http://padel.nus.edu.sg/software/padeldescriptor) |
| CDK [127] | Chemistry Development Kit (http://cdk.sourceforge.net) |
| Filter-it | Filter-it is used for calculating descriptors and filtering drug-like molecules. (http://silicos-it.com/software/software.html) |
| MODEL [128] | A webserver for molecular descriptor based upon 3D structure http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi |

PaDEL has been recently published software to calculate molecular descriptors and fingerprints. This current version of PaDEL calculates 863 molecular descriptors (including 1D, 2D and 3D). Additionally, this software also computes different types of binary fingerprints using Chemistry Development Kit (CDK) and count of chemical substructures identified by Klekota and Roth. Some additional parameters like an atom type electrotopological state descriptors, ex-

tended topochemical atom (ETA) descriptors, McGowan volume, molecular linear free energy relation descriptors etc. were calculated as identified by Laggner. PowerMV is a Window based tool for descriptor generation, and similarity search. This software is capable of calculating binary descriptors as well as the descriptors used to derive drug-likeness based upon the Lipinski's rule of five (Ro5). JOELib/JOELib2 is another java based cheminformatics library, which is being widely used for descriptor calculation, SMARTS substructure search, conversion of file formats.

Mold2 calculates 779 1D and 2D molecular descriptors from diverse information like pyhico-chemial properties, topology, atom counts, Eigen values. It has been shown that Mold2 descriptors perform better than the number of commercial software [45]. AFGen is a program that can calculate the graph based properties of chemicals. These graph properties include paths (PF), acyclic subgraphs (AF), and arbitrary topology subgraphs (GF). The ISIDA Fragmentor2011 calculates substructural molecular fragment and ISIDA property labeled fragments from a Structure-Data File (SDF).

ODDescriptors/BlueDesc is a freely available Java-based user friendly tool that calculates cheminformatics descriptor to be used to develop the model for QSAR/QSPR. This software is based upon the CDK and JOElib2 for descriptor calculations and can generate libsvm or arff file as output. CDK is a java based descriptor calculation tool developed in 2003. The tool is capable of calculating topological, geometrical, charge based and constitutional descriptors. A number of software/libraries have been developed for computing molecular descriptors using CDK. Our group made web-interface for CDK see http://crdd.osdd.net:8081/webcdk/. MODEL is molecular Descriptor Lab, for computing a comprehensive set of 3,778 molecular descriptors from following six categories: constitutional descriptors, electronic descriptors, physical chemistry properties, topological indices, geometrical molecular descriptors, and quantum chemistry.

## CHEMICAL SIMILARITY SEARCH

Searching similar molecules in chemoinformatics is an important tool for chemicals classification, database searching or the relationship between molecules and their activity. In the past, a number of tools have been developed to calculate the similarity matrix (Table **6**). These algorithms vary from simple molecular properties based, graph based, shape based and volume based and so on. Each module has its own pros and cons in searching of database. The similarity is measured in terms of Tanimoto cofficient (varies from 0.0 to 1.0) or euclidian distance. There are some algorithms which consider both shape matching and feature matching. The simplest way to calculate the similarity is provided by Open babel on MCCS166 key based similarity search [38]. The PubChem database also provides PubChem881 key based 2D similarity search. In addition, PubChem also provides a facility to search based on shape and chemical feature mapping. The JC search tool from ChemAxon has the capability to search similar molecules at a given cut-off value. Sometimes users are interested in finding the substrucure/ supestructure of similar to active molecules. This could also be done using the JC search tool.

**Table 6. Similarity Search Algorithms and Their Web Links**

| Software | Description |
|---|---|
| JCsearch | JC search used for searching similar structure, substructureas well as super structure from a given database. (http://www.chemaxon.com/jchem/doc/user/Jcsearch.html) |
| PubChem | PubChem provide the facility to search similar chemical in PubChem database using PubChem based binary fingerprints. (http://pubchem.ncbi.nlm.nih.gov/search/) |
| SIMCOMP [129] | Chemical structure similarity search against KEGG COMPOUND, KEGG DRUG, and other databases. SIMCOMP is based on 2D graph representation. (http://www.genome.jp/tools/simcomp/) |
| SUBCOMP [129] | SUBCOMP is based on bit-string representation of chemical structures. (http://www.genome.jp/tools/subcomp/) |
| SMSD [130] | SMSD is a Java based software library for calculating Maximum Common Subgraph (MCS) between small molecules. This will help us to find similarity/distance between two molecules. (http://www.ebi.ac.uk/thornton-srv/software/SMSD/) |

## QSAR/QSPR MODELS

Quantitative structure-activity relationship (QSAR) is a mathematical relationship linking chemical structure and biological/pharmacological activity in a quantitative manner for a series of chemical compounds. Related term quantitative structure–property relationship (QSPR) is used to represent the relationship between structure and physico-chemical properties [46, 47]. There are two types of QSAR models: 2D-QSAR, the models constructed using 2D descriptors, and it is established in predicting physicochemical properties as well as in providing quantitative estimates of various biological effects [48]. Another type is 3D-QSAR, when QSAR model is generated by descriptors of 3D structure of molecules [49]. The application of any QSAR models is to predict the biological activities of new compounds based on structural properties of chemicals against a particular target or whole cell [50, 51].

There are number of techniques that are frequently used to build QSAR models, this section describes major techniques used for building prediction of activity of molecules (Table **7**). Support Vector Machine (SVM) is a machine learning algorithm program which is based on statistical and optimization theory and having capability to handle structural feature data [52]. The SVM$^{light}$ software is an implementation of SVM is widely used in chemoinformatics. The WEKA package contains a wide range of tools and algorithms for data analysis and predictive modeling [53]. The system is written in JAVA, a platform independent object-oriented programming language. WEKA is a complete data mining software that could be used for pre-processing of data, clustering, model building, visualization, and feature selection. The most common file format recognize by WEKA is ARFF (attribute-relation file format) and csv for-

mat. Artificial Neural Network (ANN) is powerful machine learning technique, commonly used for solving the classification problems. SNNS (Stuttgart Neural Network Simulator) is a software simulator for neural networks on Unix (http://www.ra.cs.uni-tuebingen.de/SNNS/). The multi-layer feed-forward network, back propagation multi-layer perceptron (MLP) are the most popular application of ANN used in generating QSAR model. Memory-Based Learning is a direct descendant of the classical k-Nearest Neighbor (k-NN) approach, which is a powerful pattern classification algorithm for numeric data. K-Nearest Neighbor may be implemented using the free software TiMBL (Tilburg Memory-Based Learner) (http://ilk.uvt.nl/timbl).

**Table 7. Machine Learning and Feature Selection Techniques in Cheminformatics**

| Software | Brief Description |
|---|---|
| **Software used for developing QSAR model** | |
| SVM | SVM is a supervised learning technique, used for classification and regression analysis. The QSAR models can be optimized using different SVM parameters and kernels. (http://www.cs.cornell.edu/People/tj/svm_light/) |
| ANN | ANN is based on supervised learning, unsupervised learning and reinforcement learning. SNNS (Stuttgart Neural Network Simulator) is a free software simulator for neural networks. (http://www.ra.cs.uni-tuebingen.de/SNNS/) |
| kNN | The k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples. TiMBL is an open source software package implementing k-nearest neighbor classification. (http://ilk.uvt.nl/timbl/) |
| Weka [53] | Weka is a collection of visualization tools and algorithms for data analysis and predictive modeling, It contains libSVM, SMO, NaiveBayes, LMT, Random Forest etc learning algorithms. (http://www.cs.waikato.ac.nz/~ml/weka/) |
| **Feature Selection techniques** | |
| Weka | Weka is a popular java based tool used in feature selection. There are various feature selection methods and evaluators are available in Weka package. (http://www.cs.waikato.ac.nz/~ml/weka/) |
| Rapidminer | RapidMiner is a freely available software. It contains Brute force, Forward selection and Backward elimination algorithms for feature selection. (http://rapid-i.com/content/view/181/190/) |
| Orange | orngFSS (Orange.feature.selection) module provides feature selection facilities. It contains attMeasure, bestNAtts, selectBestNAtts, filterRelieff etc function for feature selection. (http://orange.biolab.si/) |
| RRF [54] | Regularized Random Forest (RRF) is an R package for feature selection. In RRF variables are selected based on a subsample of data at each node. (http://cran.r-project.org/web/packages/RRF/index.html) |

Feature selection is a key step to eliminate correlation, multi-collinearity and remove useless attributes from all descriptors. In (Table **7**), we describe commonly used feature selection softwares like Weka [53], Rapidminer, Orange and RRF [54]. Weka (Waikato Environment for Knowledge Analysis) is a popular java based tool used in feature selection. There are various feature selection approaches like Genetic algorithms (GA), Greedy stepwise forward selection, wrapper selection method and F-stepping remove-one are implemented in Weka. Rapidminer is a open-source software widely used for machine learning, data mining and feature selection. Brute force, Forward selection, Backward elimination etc. are important feature selection algorithms in Rapidminer (http://rapid-i.com/content/view/181/190/). Orange is a data mining and machine learning tool used in feature selection and data analysis. Orange.feature.selection module provides feature selection facilities. Regularized Random Forest (RRF) is an R package based feature selection techniques. In RRF, a set of non-redundant features can be selected without loss of predictive information [54].

## CHEMICAL CLUSTERING

Clustering of chemical is playing a very crucial role in computational chemistry [55]. The chemical clustering is used to identify the outliers in a given dataset, to understand the behaviour of a particular functional group, and also in identification of a common scaffold etc. Numbers of approach have been used for clustering the chemical compounds like the binary fingerprints based, graph properties based, maximum common substructure based [56]. Based on these, lots of softwares (commercial as well as open-source) has been developed in the past (Table **8**).

**Table 8. List of Chemical Clustering Tools and Their Web Addresses**

| Software | Brief Description |
|---|---|
| ChemMine | Chemical clustering and analysis (http://chemmine.ucr.edu/) |
| ChemMineR | It is R based open source chemical clustering tool. (http://manuals.bioinformatics.ucr.edu/home/chemminer) |
| Jcluster | ChemAxon Cluster (http://www.chemaxon.com/products/jklustor/) |
| ChemBioserver [57] | Chemical clustering webserever (http://bioserver-3.bioacademy.gr/Bioserver/ChemBioServer/) |

ChemBioServer [57] is a free web-based application that performs the clustering by two methods, the hierarchical as well as the modern Affinity Propagation (AP) clustering algorithm. While clustering, the web-server also displays the cluster in an attractive graphical manner along with the representative scaffold for a particular cluster. The compound

screening and analysis can be performed using the server based upon vdW energy, geometry, physicochemical properties, and undesired/toxic moieties.

ChemMine Tools is aimed at searching, comparing and clustering of chemicals [58]. The tool can cluster chemical by three clustering algorithms: hierarchical, binning and multidimensional scaling and the clustering of numerical data is also provided. Additionally, the property calculation module is also inbuilt in the webserver. The webserver is limiting in comparison for more than two chemicals at a time and to fish out the representative chemical of a cluster.

ChemMineR inbuilt in R environment, an open-source tool that also provides various functions for clustering entire compound libraries and visualizing clustering results and compound structures [59]. The tool supports SDF file for import molecules. ChemAxon's JKlustor Suite can be used to search similarity, calculate diversity and structural comparison and chemical clustering based on the molecular descriptors. The suite is capable of showing the representative structures for a cluster as well as the number of chemical in that particular cluster.

## MOLECULAR DOCKING

Molecular docking technique is most preferably used to predicts the preferred orientation of molecule with in the active site of target molecule where it binds to to form a stable complex. So, it is widely used in hit identification and lead optimization [60, 61]. Mostly docking algorithm generate the large number of possible structures and finally selects the most favorable structure geometry by scoring function. Depending on the interacting partner of the proteins, docking can be divided into two classes: Protein-protein docking, where two different protein molecules interact with each other and this is mainly rigid body docking and protein-ligand docking, where protein binds with small molecules.

AutoDock developed and maintained by Scripps Research Institute, is an open source molecular modeling software mainly for protein ligand docking is Autodock (Table **9**). This software include two important programs: AutoGrid pre-calculates grid maps of interaction energies for different atom types and AutoDock is used for docking of the ligand with in the predefind grid based on genetic algorithm. Dock is another anchor-and-grow based docking program. It is applied both for rigid body and flexible ligand docking. Latest, Dock version can predict binding poses by adding new features like force field scoring enhanced by solvation and receptor flexibility (Table **9**). It is developed by UCSF [62, 63]. Autodock Vina is a new open-source program for protein ligand docking and virtual screening. It is improved version of AutoDock 4, which is fast and improving the accuracy of the binding mode predictions [64]. Hex is an academically free program for protein and DNA docking. It can also use protein-ligand docking [65]. High Ambiguity Driven biomolecular DOCKing (HADDOCK) is a docking software that use the biophysical interaction data, mutagenesis data or bioinformatic predictions. It is developed for protein-protein docking; it can also be applied to protein-ligand docking. FTDOCK is a software package based on Fourier correlation algorithm used for rigid-body docking. It performs translational and rotational search in possible direction between two molecules [66].

**Table 9. Listing for Molecular Docking Tools with Brief Description**

| Name | Brief Description |
|---|---|
| Dock [62, 63] | Anchor-and-Grow based docking program, for flexible ligand and flexible protein. (http://dock.compbio.ucsf.edu/). |
| Autodock [64] | For Flexible ligand, Flexible protein side chains. Compatible for Linux, Window and Mac OS. (http://autodock.scripps.edu/). |
| Hex [65] | Mainly for protein-protein and protein-DNA docking. (http://hex.loria.fr/) |
| FTDock [66] | For rigid-body docking, based on based on Fourier correlation algorithm. (http://www.sbg.bio.ic.ac.uk/docking/ftdock.html) |
| AutoDock Vina [131] | Improved version of AutoDock4, fast, betters binding energy. (http://vina.scripps.edu/) |
| HADDock [132] | It is use for protein-protein/protein-ligand docking. (http://www.nmr.chem.uu.nl/haddock/) |

## PHARMACOPHORE TOOLS

Pharmacophore search is a key component of drug discovery programs that could be used as alternative method to molecular docking for fast and efficient screening of compound library. It represents the spatial arrangement of chemical features that is essential for a molecule to interact with a specific target receptor. Pharmacophore search is an established and effective mechanism of virtual screening [67, 68]. A brief list of freely available Pharmacophore generation software is given in (Table **10**).

The Pharmapper is a freely available webserver for identification of potential target candidates of a small-molecule [69]. This server maintains a database repository of ~7000 targets based pharmacophore models. Based on triangle hashing based method, it finds the best matching poses of input ligand against all known pharmacophore based models. This is highly useful for fast searching as it took around 1hr to screen the whole database. PharmaGist is a freely available webserver for searching pharmacophore from a set of ligand molecules [70]. This server only requires the set of ligands known to interact with a particular target without any prior knowledge of target structure. This software initially align input molecules, detect the subsets of molecules having similar features, with the possibility that a particular subset may bind to different binding sites or with different binding modes. This software also address cases where the input ligands have different affinities by defining weighted pharmacophore based on the number of ligands that share them, and automatically select the most appropriate pharmacophore for virtual screening. Therefore, it is an important tool for virtual screening of large database. Pharmer is an open-source, fast, and an efficient pharmacophore tool for

virtual screening [71]. The search time depends upon the complexity of a query molecule rather than size of database. This software takes only one pdb file at a time and use KDB-tree and Bloom fingerprint for pharmacophore searching. This software also supported the different kind of pharmacophore format like pml (ligand scout), ph4 (MOE) etc. ZincPharma is an extension of this software that could be used for screening of zinc database.

**Table 10. Different Types of Pharmocophore Searching Softwares**

| Softwares | Brief Description |
|---|---|
| Boomer | Pharmacokinetic drug monitering (http://www.boomer.org/) |
| Cyber Patient | A software for pharmacokinetic simulations (http://www.labsoft.com/www/software.html) |
| PKfit | A tool for pharmacokinetic modeling (http://cran.csie.ntu.edu.tw/web/packages/PKfit/index.html) |
| JPKD | Therapeutic drug monitoring (http://pkpd.kmu.edu.tw/jpkd/) |
| Tdm | Therapeutic drug monitoring (http://pkpd.kmu.edu.tw/tdm/) |
| mobilePK | (http://pkpd.kmu.edu.tw/mobilepk/) |
| Pharmapper [69] | Ligand based Pharmacophore search (http://59.78.96.61/pharmmapper/) |
| PharmaGist [70] | Ligand based Pharmacophore search (http://bioinfo3d.cs.tau.ac.il/PharmaGist/) |
| Pharmer [71] | Both PDB and ligand based pharmacophore search (http://smoothdock.ccbb.pitt.edu/pharmer/) |
| ZincPharma [133] | Both PDB and ligand based pharmacophore search (http://zincpharmer.csb.pitt.edu/pharmer.html) |

## ADMET TECHNIQUES

In recent years, awareness for developing computational model for predicting the absorption, distribution, metabolism, excretion and toxicity (ADMET) properties of small chemical compounds are increasing. In past, numbers of *in silico* models have been developed for fast screening and evaluating ADMET properties of small molecules based on a set of simple empirical rules (Table **11**). Although, these rules cannot evaluate the full complexity of the *in vivo* system but can provide valuable information and help decision-making.

Toxtree is an open-source user-friendly software, which can estimate the toxic potential of a chemical based on decision tree approach. This software has been designed in such a way that the implementation of new plugin is very easy and highly flexible. This software application is suitable for running on any platform, supported by Java 1.5 or higher. Presently, it could predict the cancer-causing potential in different organisms or cell lines, calculation of some important physiochemical properties, effects on human health like skin irritation, eye irritation etc. Prediction of Activity Spectra for Substances (PASS) could predict simultaneously 3678 kinds of activity with mean accuracy of prediction about 95% (leave-one-out cross validation) based on the compound's structural formula. The online webserver is able to predict ~1244 types of biological activity, including pharmacological effects, mechanisms of action, toxic and adverse effects, interaction with metabolic enzymes and transporters, influence on gene expression, etc. Therefore, it's very easy to screen large compound database with in a short period of time. So, this software is very useful for the prediction of the biological activity spectrum for existing compounds and compounds, which are virtually predicted.

**Table 11. Various ADMET Properties Prediction Tools**

| Softwares | Brief Description |
|---|---|
| OncoLogic™ | Predicting cancer causing potential of a chemical (http://www.epa.gov/oppt/cahp/pubs/can.htm) |
| OSIRIS | ADMET (http://www.organic-chemistry.org/prog/peo/) |
| Metabolizer | Drug metabolism (http://www.chemaxon.com/products/online-tryouts/metabolizer/) |
| DrugMint | A webserver for predicting druglikeness of chemical compounds. (http://crdd.osdd.net/oscadd/drugmint) |
| QED | A webserver for quantitative estimating the drug-likeness of a molecule. (http://crdd.osdd.net/oscadd/qed) |
| ToxTree [134] | Toxicity estimation (http://toxtree.sourceforge.net/download.html#Toxtree_2.5.0) |
| PASS [135] | Prediction of Activity Spectra for Substances (http://pharmaexpert.ru/Passonline/index.php) |
| admetSAR [136] | ADMET prediction (http://www.admetexp.org/predict/) |
| FAF-Drugs2 [137] | ADMET (http://www.mti.univ-paris-diderot.fr/fr/downloads.html) |
| DrugLogit [138] | Classifies compound as drug or non-drug (http://hermes.chem.ut.ee/~alfx/druglogit.html) |
| MetaSite [139] | Compound metabolism prediction (http://www.moldiscovery.com/soft_metasite.php) |
| 2D SMARTCyp [140] | Metabolism site prediction (http://www.farma.ku.dk/smartcyp/download.php) |
| MetaPred [141] | Compound metabolism prediction (http://crdd.osdd.net/raghava/metapred/) |

admetSAR is the manually curated, most comprehensive database of diverse chemicals associated with known ADMET profiles. In addition to database search, admetSAR could predict around 50 ADMET endpoints by chemoinformatics-based toolbox, entitled ADMET-Simulator, which integrates high quality and predictive QSAR models. This

webserver is helpful for *in-silico* screening ADMET profiles of drug candidates and environmental chemicals. FAF-Drugs2 is command line free software developed in python. This software has the capability to identifies key functional groups, and also some toxic and unstable molecules/functional groups. As it stands, FAF-Drugs2 implements different filtering rules such as 23 physicochemical rules and 204 substructure searching rules that can be easily customized. This software is also implemented with Gnuplot software to plot several distribution diagrams of major physicochemical properties of the screened compound libraries. OncoLogic™ is a desktop application that can analyze a chemical structure to determine the likelihood that it may cause cancer. This is based on some rules derived by applying structure activity relationship (SAR) analysis and incorporating knowledge of how chemicals cause cancer in animals and humans, and mimicking the decision logic of human experts.

OSIRIS Property Explorer is a freely available webserver that can predict physico-chemical and toxicological molecular properties, need to be optimized at the time of designing pharmaceutically active compounds. This tool was originally developed by T. Sander and later released in the public domain in 1999 on Actelion's web site to demonstrate the applicability of Java applets for real-time cheminformatics applications. This tool calculates drug-likeness, drug score etc. along with some important physic-chemical properties. DrugLogit is available in the form of freely available webserver to predict the probability of a compound to act as drug or non-drug. This tool is based on simple, readily available molecular properties of a compound. A selection of the equations also allows classifying the disease category of a compound. Approximately, 23 equations have been used in this software for prediction. They are rationalized based on the different mechanism of action, administration mode, and target organs of different disease categories.

## TOOLS FOR DRUG METABOLISM

Drug metabolism is an important aspect in the drug discovery process. A number of cytochrome proteins classified into different protein families are known to be responsible for drug metabolism. The high or low metabolism of a drug is related with its dose requirement. Therefore, it is very important to predict the fate of a compound to be metabolized or not, site of action etc. Towards this, a number of tools have been developed such as MetaSite, MetaPred etc. (Table **11**).

MetaSite is freely available software that could predict metabolic transformations related to cytochrome-mediated reactions. This software also provides the structure of the metabolites formed, highlights the molecular moieties that help to direct the molecule in the cytochrome cavity. This software also claims that primary site of metabolism has been accurately predicted in more than 85% of the cases. SMARTCyp 2.0 is another freely available JAVA based downloadable software for metabolic site prediction for all five major drug-metabolizing. Metabolizer computes all the possible metabolites of a given molecule, predicts the major metabolites, and estimates metabolic stability. MetaPred is a freely available webserver for predicting whether the compound will be metabolized or not. It also predicts the family of cytochrome responsible for its metabolism.

## DRUG TARGET IDENTIFICATION

During the last two decades or so, tremendous progress have been made in medicinal chemistry for discovering new potential drug targets. Identifying drug targets by experiments alone would be a very time-consuming and costly affair. In addition, it is also important to use systems based approaches to identify drug targets for a polypharmacology approach to ensure targeting more than one protein at time for better efficacy. These approaches need quality annotation of proteins both at structural and functional level which are then used to construct interaction graphs to identify central proteins. These central hub proteins may then be selected in combination depending on their structural features, expression profiles and localization to identify best pairs for polypharmacology. Therefore, a lot of computational algorithms such as those for identifying drug-target interaction networks [72, 73], inhibitor design [74-76], multiple drug-target prediction [77-79], classifying body fluids [80], identifying recombination spots with pseudo dinucleotide compositioniden [81], classifying hepalar cirrhosis and carcinoma [82], classifying anatomical therapeutic chemical (ATC) classification of drugs [83] etc. have been developed for target discovery. These software provides very useful insights for both medicinal chemistry research and drug development [84].

Understanding the location of a protein molecule is of prime importance in prioritizing the drug targets. This would be very helpful in understanding the biological phenomenon like protein-protein interactions, protein-ligand interaction, pathways analysis. Over the years, a number of software have been developed in predicting the protein localization in the differnet compartment of various cell types as summarized in (Table **12**), such as eukarotic, human, plant, bacterial subcellular localization [16, 85-90]. Similarly, algorithms for predicting antimicrobial peptides [91], identifying HIV cleavage sites in proteins [92, 93], predicting proteases and their types [94], identifying virulence factors [95] have been developed for predicting potential drug targets.

In addition to that, there are some well known family of proteins like GPCR, nuclear receptors, kinases etc. that have great contribution in drug discovery [12, 96-102]. GPCR proteins is being targeted by nearly 50% of marketed drugs and nuclear receptors targeted by nearly 13% FDA approved drugs [103]. Therefore, characterizing these class of protein family will provide in-depth knowledge in undertanding the ligand-protein interactions and drugs side-effect. In (Table **13**), we describes a series of web-server/tools important in drug target discovery.

## DESIGNING OF INHIBITORS

Inhibitors are required to block a target or stop a signaling cascade. These inhibitors could be exploited in three approaches: structure based inhibitor design (SBID), ligand based inhibitor design (LBID) and de novo inhibitor design (DNID) [104, 105]. The structure of target is the perquisite for the SBID, which could be determined by X-ray crystallography or NMR [106-108]. Alternatively, a huge list of

**Table 12. Webserver for Proteome Annotation with Their Brief Description**

| Webserver | Description |
|---|---|
| ATPint [14] | A web based tool for prediction of ATP binding residue in protein sequence http://www.imtech.res.in/raghava/atpint |
| ESLpred2 [16] | This can predict four major protein localizations (cytoplasmic, mitochondrial, nuclear and extracellular) for eukaryotes. http://www.imtech.res.in/raghava/eslpred2 |
| iRSpot-PseDNC [81] | Identify recombination spots with pseudo dinucleotide composition. http://lin.uestc.edu.cn/server/iRSpot-PseDNC |
| iLoc-Virus [85] | A classifier for identifying the subcellular localization of virus proteins with both single and multiple sites http://icpr.jci.edu.cn/bioinfo/iLoc-Virus |
| iLoc-Hum [86] | A software for predicting subcellular locations of human proteins with both single and multiple sites http://icpr.jci.edu.cn/bioinfo/iLoc-Hum or http://www.jci-bioinfo.cn/iLoc-Hum |
| iLoc-Plant [87] | A webserver for predicting the subcellular localization of single and multiple site plant proteins http://icpr.jci.edu.cn/bioinfo/iLoc-Plant or http://www.jci-bioinfo.cn/iLoc-Plant |
| iLoc-Gpos [88] | Predicting the subcellular localization of singleplex and multiplex in Gram-positive bacterial proteins http://icpr.jci.edu.cn/bioinfo/iLoc- Gpos or http://www.jci-bioinfo.cn/iLoc-Gpos |
| iLoc-Euk [89] | A webserver for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins http://icpr.jci.edu.cn/bioinfo/iLoc-Euk |
| HIVcleave [92] | A web-server for predicting HIV protease cleavage sites in proteins http://chou.med.harvard.edu/bioinf/HIV/ |
| HSLpred [142] | Knowledge-based tool for predicting subcellular localization of human proteins http://www.imtech.res.in/raghava/hslpred |
| MitPred [143] | A web-server specifically trained to predict the proteins, which are destined to localize in mitochondria in yeast and animals particularly using domain profiles http://www.imtech.res.in/raghava/mitpred |
| iLoc-Gneg [144] | Predicting subcellular localization of both single and multiple sites in gram-negative bacterial proteins http://icpr.jci.edu.cn/bioinfo/iLoc-Gneg |
| PSLpred [145] | PSLpred for predicting subcellular localization of gram-negative bacterial proteins http://www.imtech.res.in/raghava/pslpred |
| iDNA-Prot [146] | Identification of DNA binding proteins using random forest with grey model http://icpr.jci.edu.cn/bioinfo/iDNA-Prot or http://www.jci-bioinfo.cn/iDNA-Prot |
| iSMP-Grey [147] | Prediction of Secretory proteins in Malaria Parasite http://www.jci-bioinfo.cn/iSMP-Grey |
| Signal-CF [148] | A webserver for predicting signal peptides http://chou.med.harvard.edu/bioinf/Signal-CF/ or http://202.120.37.186/bioinf/Signal-CF/ |
| GSTpred [149] | GSTpred is a computational tool for identification of the Glutathione S-transferase protein http://www.imtech.res.in/raghava/gstpred |
| ATPsite [150] | A machine learning-based predictor to identifies ATP-binding residues from protein sequences http://biomine.ece.ualberta.ca/ATPsite |
| GTPbinder [151] | Web based tool for prediction of GTP binding residue in protein sequence http://www.imtech.res.in/raghava/gtpbinder |
| FADpred [152] | FADPred is a web-server specially trained for the FAD interacting residues http://www.imtech.res.in/raghava/fadpred |
| NADbinder [153] | To predict NAD binding proteins and their interacting residues http://www.imtech.res.in/raghava/nadbinder |
| GlycoPP [154] | GlycoPP is a webserver for predicting potential N-and O-glycosites in prokaryotic protein sequence http://www.imtech.res.in/raghava/glycopp |

**(Table 12) contd….**

| Webserver | Description |
|---|---|
| NetOGlyc [155] | The server incorporates neural network to identify mucin type GalNAc O-glycosylation sites in mammalian proteins http://www.cbs.dtu.dk/services/NetOGlyc |
| Sulfinator [156] | A algorithm that can be used to predict tyrosine sulfation sites in proteins http://web.expasy.org/sulfinator |
| Sulfosite [157] | Software to analyze protein sulfotyrosine sites using machine learning methods http://sulfosite.mbc.nctu.edu.tw |
| SUMOsp [158] | In silico method for predicting sumoylation sites in proteins http://sumosp.biocuckoo.org |

**Table 13. Software for Drug Target Prediction with Their Brief Description**

| Software | Description |
|---|---|
| GPCRpred [12] | This is a server for predicting G-protein-coupled receptors and for classifying them in families and sub-families http://www.imtech.res.in/raghava/gpcrpred |
| GPCR-2L [96] | Prediction of G protein-coupled receptors and their types http://icpr.jci.edu.cn/ |
| GPCR-CA [97] | A cellular automaton image approach for predicting G-protein-coupled receptor functional classes http://icpr.jci.edu.cn/ |
| GPCR-GIA [98] | A web-server for identifying G-protein coupled receptors and their families with grey incidence analysis http://218.65.61.89:8080/bioinfo/GPCR-GIA |
| GPCRSclass [99] | This is a dipeptide composition based method for predicting Amine Type of G-protein-coupled receptors http://www.imtech.res.in/raghava/gpcrsclass |
| KinasePhos2.0 [159] | This tool could be used for identifying phosphorylation sites using protein sequence profile and protein coupling pattern features http://KinasePhos2.mbc.nctu.edu.tw |
| NetPhosK (*160*) | The server utilizes neural network to predict kinase specific eukaryotic protein phosphoylation sites http://www.cbs.dtu.dk/services/NetPhosK |
| VGIchan [161] | VGIchan server is to predict voltage gated ion channels and classify them into sodium, potassium, calcium and chloride ion channels. http://www.imtech.res.in/raghava/vgichan |
| VICMpred [162] | To classify the proteins of bacteria into virulence factors, information molecule, cellular process and metabolism molecule http://www.imtech.res.in/raghava/vicmpred |
| NRpred [163] | A SVM based tool for the classification of nuclear receptors on the basis of amino acid composition or dipeptide composition http://www.imtech.res.in/raghava/nrpred |
| NR-2L [164] | A webserver for two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features http://icpr.jci.edu.cn/bioinfo/NR2L or http://www.jci-bioinfo.cn/NR2L |
| iNR-PhyChem [165] | A physical-chemical property matrix based method for identifying nuclear receptors and their subfamilies http://www.jci-bioinfo.cn/iNR-PhysChem or http://icpr.jci.edu.cn/bioinfo/iNR-PhysChem |
| iNuc-PhysChem [166] | Software for sequence-based predictor for identifying nucleosomes via physicochemical properties http://lin.uestc.edu.cn/server/iNuc-PhysChem |
| SRTpred [167] | A tool to predict secretory proteins irrespective of presence or absence of N-terminal signal peptides using machine-learning techniques. http://www.imtech.res.in/raghava/srtpred |
| GPS-Polo [168] | In-silico tool to predict Plk-specific phospho-binding and phosphorylation sites in proteins http://polo.biocuckoo.org |

**(Table 13) contd….**

| Software | Description |
|---|---|
| ProteDNA [169] | Computer program that could be explored to identify the binding residues in a transcription factor http://serv.csbb.ntu.edu.tw/ProteDNA |
| 3DTF [170] | A knowledge-based tool for identification of transcription factors http://cogangs.biobase.de/3dtf/ |
| TESS [171] | This tool could be explored to predict transcription factor binding sites in DNA sequence http://www.cbil.upenn.edu/cgi-bin/tess/tess |
| PREDetector [172] | In-siico prediction algorithm for regulatory elements of DNA-binding proteins in bacterial genomes http://www.montefiore.ulg.ac.be/~hiard/PreDetector |
| PROMO [173] | Web-based approach for identification of transcription factor binding sites in DNA sequences http://alggen.lsi.upc.es/cgibin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3 |
| CancerPred [174] | CancerPred is a web-server specially trained for predicting the Cancerlectins http://www.imtech.res.in/raghava/cancer_pred |

software is also available for target structure prediction; whose description is beyond the scope of this review. Technically, homology modeling, threading and docking strategy are the drivers for SBID. Currently, there are several methods that have been developed using the above-mentioned approach. Here we are describing some of these methods; GDoQ predicts the inhibitors against GlmU enzyme from *Mycobacterium tuberculosis* [109], KiDoQ, is another webserver for designing inhibitors against Dihyrodipcolinate synthase (DHDPS), a potential drug target enzyme of a unique bacterial DAP/Lysine pathway [110] (Table **14**). In the absence of target structure, LBID could be applied where new or improved inhibitors could be designed computationally from a dataset of already known inhibitors [111, 112]. Toxipred is a user-friendly web server for the prediction of aqueous toxicity of small chemical molecules in *T. pyriformis*. DNID approach is applied for designing where no inhibitors were reported previously [113, 114]. In DNID, the ligand is built based upon the complementarity of the active site in a target with the ligand [115, 116]. A webserver named "Drugster" has implemented LigBuilder for building ligands. e-LEA3D is webserver dedicated to drug designing with focus on generating de novo libraries and virtual screening (Table **14**).

## OTHER RESOURCES

Although this review covers a lot about open source applications in drug discovery, a large number of freely available resources, for computational drug discovery are not covered. We are trying to summarize them in this section.

CRDD (Computational resources for drug discovery) is an open source *in silico* repository for tools being developed under the aegis of OSDD. This repository provides free access to webserver, databases and software related to drug discovery (http://crdd.osdd.net/). DrugPedia is another project from OSDD, where information of drugs is maintained in the form of pages (http://crdd.osdd.net/drugpedia). MayaChemTools is a collection of Perl script for handling and manipulating the structure files for general purpose in drug discovery (www.mayachemtools.org/). CADDSuite offers modular tools developed by Oliver Kohlbacher for data storage, docking, QSAR and analyzing result in the field of computer-aided drug design (http://www.ballview.org/caddsuite). The integration of this tool in galaxy platform helps in creating the workflow for complexes process. ChemBench (http://chembench.mml.unc.edu/) is developed by Carolina Exploratory Center for Cheminformatics Research (CECCR) to provide a platform for virtual libraries of available chemicals with predicted biological and drug-like properties, model building, property and activity predictors, and special tools for chemical library design.

## MAJOR INITIATIVES

Along with the availability of numerous open source tools and softwares, there are also some organisations working in a collaborative manner with public or industry partner in developing affordable medicines particularly for neglected diseases like TB, malaria etc. (Table **15**). The Drugs for Neglected Diseases Initiative (DNDi) is an open source, non-profit, collaborative project for developing new treatments against Neglected Diseases with major emphasis on malaria, slepping sickness. Till date, it has contributed two compounds for malaria, one compound for sleeping sickness. The Infectious Disease Research Institute (IDRI) is another non-profit rganization with major focus on infectious diseases like tuberculsois, malaria etc. This is involved in prevention, diagnosis and treatment of infectious diseases. In collaboration with GSK, their TB vaccine are in phase-1 clinical trial. OpenTox was initiated as a collaborative project to develop *in-silico* toxicology models that could be used for the creation of predictive toxicology applications. This involving the collaboration between different university, enterprise, and government research groups to design and build the initial OpenTox framework. Blue Obelisk is the name used by diverse internet group promoting reusable chemistry via open-source software development. The three major areas of this movement are 1) open source 2) open standard 3) open data. Open source for drug discovery (OSDD) is a translational platform for drug discovery, which connects informaticians, experimental biologist, clinician,

**Table 14. Tools for Designing Inhibitors with Their Brief Description**

| Software | Description |
|---|---|
| ToxiPred | A web server for the prediction of aqueous toxicity against *T. pyriformis* (http://crdd.osdd.net/raghava/toxipred) |
| eBooster | A webserver for prediction of EC50 value of compounds against EthR of *M.Tuberculosis* (http://crdd.osdd.net/oscadd/ebooster/) |
| ntEGFR | Prediction of EGFR inhibitors (http://crdd.osdd.net/oscadd/ntegfr/) |
| TLR4hi | Prediction of inhibitors against Human TLR4 (http://crdd.osdd.net/oscadd/tlr4hi/) |
| Hivfin | A webserver for prediction of Fusion Inhibitors against HIV (http://crdd.osdd.net/oscadd/hivfin/) |
| KetoDrug | A web server for binding affinity prediction of ketoxazole derivatives against Fatty Acid Amide Hydrolase (FAAH). (http://crdd.osdd.net/oscadd/ketodrug/) |
| DMKPred | Drug molecules for kinase protein (http://www.imtech.res.in/raghava/dmkpred/) |
| CARBOTOPE | Prediction of carbohydrate based epitope (http://crdd.osdd.net/raghava/carbotope/) |
| ABMPred | Prediction of AntiBacterial Compounds against MurA Enzyme. (http://crdd.osdd.net/oscadd/abmpred/) |
| Gdoq [109] | Server to predict the Mycobacterial GlmU enzyme inhibitor (http://crdd.osdd.net/raghava/gdoq) |
| Kidoq [110] | Dihyrodipcolinate synthase inhibitors desinging software (http://crdd.osdd.net/raghava/kidoq) |
| PIRSpred [175] | HIV-1 genotype resistance/susceptibility prediction server (http://protinfo.compbio.washington.edu/pirspred/) |
| LigBuilder [176] | A computer program for strucrture based denovo drug designing (http://ligbuilder.org/intro.html) |
| Drugster [177] | A fully interactive lead and structure designing server (http://www.bioacademy.gr/bioinformatics/drugster/) |
| e-LEA3D [178] | Scaffold designing, virtual screening and generate combinatorial library (http://bioinfo.ipmc.cnrs.fr/lea.html.) |
| MDRIpred [179] | A webserver for prediction of Inhibitor against Drug Resistant *M.Tuberculosis.* (http://crdd.osdd.net/oscadd/mdri/) |

**Table 15. Some Open Source Initiatives for Drug Discovery with Their Research Area**

| Project | Research Areas |
|---|---|
| Drugs for Neglected Diseases Initiative | Sleeping sickness, visceral leishmaniasis, Chagas disease (http://www.dndi.org/) |
| Infectious Disease Research Institute | Tuberculosis, leishmaniasis, Chagas disease, malaria, leprosy and Buruli ulcer (http://www.idri.org/) |
| Blue Obelisk | Provides open source cheminformatics tools. (http://sourceforge.net/apps/mediawiki/blueobelisk/index.php?title=Main_Page) |
| OSDD | Promoting open source for neglected disease. (http://www.osdd.net/) |
| OpenTox | Toxicology (http://www.opentox.org/) |
| Global Alliance for TB | Tuberculosis (http://www.tballiance.org/) |

research organizations to provide affordable medicine against tuberculosis, malaria. This project was launched in 2008 with the motto "affordable health care for all".

## FUTURE PROSPECTS

The current review has focussed on the existing algorithms and tools in cheminformatics with specific emphasis on those that exists in public domain. It is evident that a lot of efforts are being made to enhance the preditive potential of these tools. However, there are limitations and areas where such efforts are lacking. Some of the limitations include the fact that computational methods does not create real life situation during docking experiments where a ligand is to find its target in presence of many proteins with high

precision. Also, there is lack of applications that allow inhibition of multiple proteins in a critical pathway or consider host genome polymorphism for drug metabolism and transport. Another challenge particularly in drug discovery projects is also to model combination therapy and prediciting the right dosage based on pharmacokinetic parameters. These are some of the challenges that need experimental and theoretical researchers to work in collaboration for developing better platforms for drug discovery programs.

As most of the drug discovery research has been a part of pharmaceutical industries, the field of computer-aided drug designing is dominated by commercial tools. Academia and institutes are making efforts to design better predictive models and provide them as open source tools which can be worked on by others thus ensuring continuous improvements. Given that most efforts in academia and institutes are not directly linked to drug discovery and development, the prediction accuracy and fine tuning of these models is limited and needs to be benchmarked more comprehensively. Unlike pharmaceutical set up where this is done as a routine exercise, validating predictions in a research environment is mostly restricted to smaller datasets and workable experiments. On the contrary, it is also worth mentioning that most of the advancements in understanding of chemical space and their drug like properties is studied and published by academic researchers which ultimately feeds into predictive tools, both open source and proprietary. Thus, it is evident that novel methods and intelligent designs published by researchers in public funded organizations is utilized by industry and translated into pipelines for drug research. These platforms provide an integrated environment for researchers to study, evaluate and prioritize compounds for their drug-like properties in a user and time friendly manner. Creating such platforms in open source environment is need of the hour which demands concerted effort in ensuring consistent data formats and ontologies for curating data and sharing the results of data analyses. Open source communities working towards developing these platforms need to ensure that these standards are followed and the community at large should be made aware of using such standards for better data organization and analyses. It is also imperative to systematically benchmark these models and their predictive capability which feeds back and provide more clues for improvement towards faster drug development with reduced failure rates. These collaborative platforms allow researchers to work together and solve challenging problems by sharing ideas and discover alternate, more efficient mechanisms of resolving them which are beyond the expertise of any individual group or laboratory.

Researchers, mostly, from public funded organizations share their scientific discoveries and resources via peer-reviewed publications. With the advent of open access, that provides a open platform to researchers for sharing their research outcomes, there has been a constant pressure for the researcher to arrange for publishing charges. It is mandatory for researchers to provide access to the details of the tools or algorithms they have developed through peer-reviewed publications. Hence, in order to overcome this barrier, we think that the article should be published without any publication charges or the country scientific department should provide the grants for the publication fee. Open access ensures larger readership which is a good primer for better citations of the publications and works as strong incentive and motivation for young researchers.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

## ACKNOWLEDGEMENTS

The authors are thankful to the Council of Scientific and Industrial Research, India for funding (Grant No. HCP0001).

## ABBREVIATIONS

| | | |
|---|---|---|
| CDD | = | Collaborative Drug Discovery |
| SMILES | = | Simplified Molecular-Input Line Entry System |
| OSDD | = | Open Source Drug Discovery |
| WOMBAT | = | WOrld of Molecular BioAcTivity |
| QSAR | = | Quantitative Structure Activity Relationship |
| QSPR | = | Quantitative Structure Property Relationship |
| QSTR | = | Quantitative Structure Toxicity Relationship |
| CADD | = | Computer Aided Drug Design |
| ADMET | = | Absorption, Distribution, Metabolism, Toxicity |

## REFERENCES

[1] Drews, J. Drug discovery: a historical perspective. *Science*, **2000**, *287*(5460), 1960–1964.

[2] Ban, T. A. The role of serendipity in drug discovery. *Dialogues Clin. Neurosci.*, **2006**, *8*(3), 335–344.

[3] Jones, A. W. Early drug discovery and the rise of pharmaceutical chemistry. *Drug Test. Anal.*, **2011**, *3*(6), 337–344.

[4] Jagusztyn-Krynicka, E. K.; Wyszyńska, A. The decline of antibiotic era--new approaches for antibacterial drug discovery. *Pol. J. Microbiol.*, **2008**, *57*(2), 91–98.

[5] O'Connor, K. A.; Roth, B. L. Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat. Rev. Drug Discov.*, **2005**, *4*(12), 1005–1014.

[6] Fda Challenge and Opportunity on the Critical Path to New Medical Products. *Review Literature And Arts Of The Americas*, **2004**, 1–31.

[7] Wlodawer, A. Rational approach to AIDS drug design through structural biology. *Annu. Rev. Med.*, **2002**, *53*(6), 595–614.

[8] DesJarlais, R. L.; Seibel, G. L.; Kuntz, I. D.; Furth, P. S.; Alvarez, J. C.; Ortiz De Montellano, P. R.; DeCamp, D. L.; Babé, L. M.; Craik, C. S. Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease. *Proc. Natl. Acad. Sci. U. S. A.*, **1990**, *87*(17), 6644–6648.

[9] Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **2012,** *41*(Database issue), D43-D47.

[10] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.*, **1978**, *185*(2), 584–591.

[11] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.*, **1990**, *215*(3), 403–410.

[12] Bhasin, M.; Raghava, G. P. S. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.*, **2004**, *32*(Web Server issue), W383–389.

[13] Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.,* **1988**, *85*, (8), 2444–2448.

[14] Chauhan, J. S.; Mishra, N. K.; Raghava, G. P. S. Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics*, **2009**, *10*, 434.

[15] Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics*, **2007**, *23*(21), 2947–2948.

[16] Garg, A.; Raghava, G. P. S. ESLpred2: improved method for predicting subcellular localization of eukaryotic proteins. *BMC Bioinformatics*, **2008**, *9*, 503.

[17] McGuffin, L. J.; Bryson, K.; Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics*, **2000**, *16*(4), 404–405.

[18] Bhardwaj, A.; Scaria, V.; Raghava, G. P. S.; Lynn, A. M.; Chandra, N.; Banerjee, S.; Raghunandanan, M. V; Pandey, V.; Taneja, B.; Yadav, J.; Dash, D.; Bhattacharya, J.; Misra, A.; Kumar, A.; Ramachandran, S.; Thomas, Z.; Brahmachari, S. K. Open source drug discovery--a new paradigm of collaborative research in tuberculosis drug development. *Tuberculosis (Edinb)*, **2011**, *91*(5), 479–486.

[19] Ardal, C.; Røttingen, J. A. Open source drug discovery in practice: a case study. *PLoS Negl. Trop. Dis.*, **2012**, *6*(9), E1827.

[20] Ardal, C.; Alstadsæter, A.; Røttingen, J. A. Common characteristics of open source software development and applicability for drug discovery: a systematic review. *Health Res. Policy Syst.*, **2011**, *9*, 36.

[21] Tcheremenskaia, O.; Benigni, R.; Nikolova, I.; Jeliazkova, N.; Escher, S. E.; Batke, M.; Baier, T.; Poroikov, V.; Lagunin, A.; Rautenberg, M.; Hardy, B. OpenTox predictive toxicology framework: toxicological ontology and semantic media wiki-based OpenToxipedia. *J. Biomed. Semantics*, **2012**, *3 SUPPL 1*, S7.

[22] Ekins, S.; Bradford, J.; Dole, K.; Spektor, A.; Gregory, K.; Blondeau, D.; Hohman, M.; Bunin, B. A. A collaborative database and computational models for tuberculosis drug discovery. *Mol. Biosyst.*, **2010**, *6*(5), 840–851.

[23] O'Boyle, N. M.; Guha, R.; Willighagen, E. L.; Adams, S. E.; Alvarsson, J.; Bradley, J.C.; Filippov, I. V; Hanson, R. M.; Hanwell, M. D.; Hutchison, G. R.; James, C. A.; Jeliazkova, N.; Lang, A. S.; Langner, K. M.; Lonie, D. C.; Lowe, D. M.; Pansanel, J.; Pavlov, D.; Spjuth, O.; Steinbeck, C.; Tenderholt, A. L.; Theisen, K. J.; Murray-Rust, P. Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on. *J. Cheminform.*, **2011**, *3*(1), 37.

[24] Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mractc, M.; Oprea, T. I. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In *Chemical Biology*; Wiley-VCH Verlag GmbH, 2008; Vol. 1-3, pp. 760–786.

[25] Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay Database. *Nucleic Acids Res.*, **2012**, *40*(Database issue), D400–412.

[26] Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **2009**, *37*(Web Server issue), W623–633.

[27] Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.*, **2012**, *52*(7), 1757-1768.

[28] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **2012**, *40*(Database issue), D1100–1107.

[29] Chen, J. H.; Linstead, E.; Swamidass, S. J.; Wang, D.; Baldi, P. ChemDB update--full-text search and virtual chemical space. *Bioinformatics*, **2007**, *23*(17), 2348–2351.

[30] Little, J. L.; Williams, A. J.; Pshenichnov, A.; Tkachenko, V. Identification of "known unknowns" utilizing accurate mass data and ChemSpider. *J. Am. Soc. Mass Spectrom.*, **2012**, *23*(1), 179–185.

[31] Milne, G. W.; Nicklaus, M. C.; Driscoll, J. S.; Wang, S.; Zaharevitz, D. National Cancer Institute Drug Information System 3D database. *J. Chem. Inf. Comput. Sci.,* **1994**, *34*(5), 1219–1224.

[32] Goto, S.; Okuno, Y.; Hattori, M.; Nishioka, T.; Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **2002**, *30*(1), 402–404.

[33] Schreyer, A.; Blundell, T. CREDO: a protein-ligand interaction database for drug discovery. *Chem. Biol. Drug Des.*, **2009**, *73*(2), 157–167.

[34] Hendlich, M.; Bergner, A.; Günther, J.; Klebe, G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.*, **2003**, *326*(2), 607–620.

[35] Chen, I. J.; Foloppe, N. Conformational sampling of druglike molecules with MOE and catalyst: implications for pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.*, **2008**, *48*(9), 1773–1791.

[36] Foloppe, N.; Chen, I. J. Conformational sampling and energetics of drug-like molecules. *Curr. Med. Chem.*, **2009**, *16*(26), 3381–3413.

[37] Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. Conformational analysis of macrocycles: finding what common search methods miss. *J. Chem. Inf. Model.*, **2009**, *49*(10), 2242–2259.

[38] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.*, **2011**, *3*, 33.

[39] Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tufféry, P. Frog: a FRee Online druG 3D conformation generator. *Nucleic Acids Res.*, **2007**, *35*(Web Server issue), W568–572.

[40] Puranen, J. S.; Vainio, M. J.; Johnson, M. S. Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. *J. Comput. Chem.*, **2010**, *31*(8), 1722–1732.

[41] Lagorce, D.; Pencheva, T.; Villoutreix, B. O.; Miteva, M. A. DG-AMMOS: a new tool to generate 3d conformation of small molecules using distance geometry and automated molecular mechanics optimization for in silico screening. *BMC Chem. Biol.*, **2009**, *9*, 6.

[42] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley & Sons: New York, 2000.

[43] Katritzky, A. R.; Gordeeva, E. V Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.*, *33*(6), 835–857.

[44] Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.*, *44*(5), 1630–1638.

[45] Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Model.*, **2008**, *48*(7), 1337–1344.

[46] Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D. Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. *Chem. Rev.*, **2002**, *102*(3), 783–812.

[47] Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.*, **2007**, *152*(1), 9–20.

[48] Dudek, A. Z.; Arodz, T.; Gálvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb. Chem. High Throughput Screen.*, **2006**, *9*(3), 213–228.

[49] Tamura, H.; Ishimoto, Y.; Fujikawa, T.; Aoyama, H.; Yoshikawa, H.; Akamatsu, M. Structural basis for androgen receptor agonists and antagonists: interaction of SPEED 98-listed chemicals and related compounds with the androgen receptor based on an *in vitro* reporter gene assay and 3D-QSAR. *Bioorg. Med. Chem.*, **2006**, *14*(21), 7160–7174.

[50] Kubinyi, H. QSAR and 3D QSAR in drug design Part 1: methodology. *Drug Discov. Today*, **1997**, *2*(11), 457–467.

[51] Ebalunode, J. O.; Zheng, W.; Tropsha, A. Application of QSAR and shape pharmacophore modeling approaches for targeted chemical library design. *Methods Mol. Biol.*, **2011**, *685*, 111–133.

[52] Joachims, T. Making large-Scale SVM Learning Practical. *Advances in Kernel Methods Support Vector Learning*, **1999**, 169–184.

[53] Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics*, **2004**, *20*(15), 2479–2481.

[54] Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.,* **2003**, 3, 1157–1182.

[55] Schuffenhauer, A.; Brown, N.; Ertl, P.; Jenkins, J. L.; Selzer, P.; Hamon, J. Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *J. Chem. Inf. Model.*, **2007**, *47*(2), 325–336.

[56] Chu, C. W.; Holliday, J. D.; Willett, P. Effect of data standardization on chemical clustering and similarity searching. *J. Chem. Inf. Model.,* **2009**, *49*(2), 155–161.

[57] Athanasiadis, E.; Cournia, Z.; Spyrou, G. ChemBioServer: A web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery. *Bioinformatics*, **2012**, *28*(22), 3002–3003.

[58] Backman, T. W. H.; Cao, Y.; Girke, T. ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Res.*, **2011**, *39*(Web Server issue), W486–491.

[59] Cao, Y.; Charisi, A.; Cheng, L. C.; Jiang, T.; Girke, T. ChemmineR: a compound mining framework for R. *Bioinformatics*, **2008**, *24*(15), 1733–1734.

[60] Gschwend, D. A.; Good, A. C.; Kuntz, I. D. Molecular docking towards drug discovery. *J. Mol. Recognit.*, **1996**, *9*(2), 175–186.

[61] Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, **2004**, *3*(11), 935–949.

[62] Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA*, **2009**, *15*(6), 1219–1230.

[63] Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput. Aided Mol. Des.*, **2006**, *20*, (10–11), 601–619.

[64] Morris, G. M.; Huey, R.; Olson, A. J. Using AutoDock for ligand-receptor docking. *Curr. Protoc. Bioinformatics*, **2008**, *CHAPTER 8*, UNIT 8.14.

[65] Macindoe, G.; Mavridis, L.; Venkatraman, V.; Devignes, M. D.; Ritchie, D. W. HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res.*, **2010**, *38*(Web Server issue), W445–449.

[66] Jackson, R. M.; Gabb, H. A.; Sternberg, M. J. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.*, **1998**, *276*(1), 265–285.

[67] Horvath, D. Pharmacophore-based virtual screening. *Methods Mol. Biol.*, **2011**, *672*, 261–298.

[68] Spitzer, G. M.; Heiss, M.; Mangold, M.; Markt, P.; Kirchmair, J.; Wolber, G.; Liedl, K. R. One concept, three implementations of 3D pharmacophore-based virtual screening: distinct coverage of chemical search space. *J. Chem. Inf. Model.*, **2010**, *50*(7), 1241–1247.

[69] Liu, X.; Ouyang, S.; Yu, B.; Liu, Y.; Huang, K.; Gong, J.; Zheng, S.; Li, Z.; Li, H.; Jiang, H. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. *Nucleic Acids Res.*, **2010**, *38*(Web Server issue), W609–614.

[70] Schneidman-Duhovny, D.; Dror, O.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. PharmaGist: a webserver for ligand-based pharmacophore detection. *Nucleic Acids Res.*, **2008**, *36*(Web Server issue), W223–228.

[71] Koes, D. R.; Camacho, C. J. Pharmer: efficient and exact pharmacophore search. *J. Chem. Inf. Model.*, **2011**, *51*(6), 1307–1314.

[72] He, Z.; Zhang, J.; Shi, X. H.; Hu, L. L.; Kong, X.; Cai, Y. D.; Chou, K. C. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS One*, **2010**, *5*(3), E9603.

[73] Huang, T.; Zhang, J.; Xu, Z. P.; Hu, L. L.; Chen, L.; Shao, J. L.; Zhang, L.; Kong, X. Y.; Cai, Y. D.; Chou, K. C. Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. *Biochimie*, **2012**, *94*(4), 1017–1025.

[74] Wang, J. F.; Chou, K. C. Insights into the mutation-induced HHH syndrome from modeling human mitochondrial ornithine transporter-1. *PLoS One*, **2012**, *7*(1), E31048.

[75] Li, X. B.; Wang, S. Q.; Xu, W. R.; Wang, R. L.; Chou, K. C. Novel inhibitor design for hemagglutinin against H1N1 influenza virus by core hopping method. *PLoS One*, **2011**, *6*(11), E28111.

[76] Lian, P.; Wei, D. Q.; Wang, J. F.; Chou, K. C. An allosteric mechanism inferred from molecular dynamics simulations on phospholamban pentamer in lipid membranes. *PLoS One*, **2011**, *6*(4), E18587.

[77] Yu, H.; Chen, J.; Xu, X.; Li, Y.; Zhao, H.; Fang, Y.; Li, X.; Zhou, W.; Wang, W.; Wang, Y. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS One*, **2012**, *7*(5), E37608.

[78] Ma, Y.; Wang, S. Q.; Xu, W. R.; Wang, R. L.; Chou, K. C. Design novel dual agonists for treating type-2 diabetes by targeting peroxisome proliferator-activated receptors with core hopping approach. *PLoS One*, **2012**, *7*(6), E38546.

[79] Anand, P.; Sankaran, S.; Mukherjee, S.; Yeturu, K.; Laskowski, R.; Bhardwaj, A.; Bhagavat, R.; Brahmachari, S. K.; Chandra, N. Structural Annotation of Mycobacterium tuberculosis Proteome. *PLoS One*, **2011**, *6*(10), E27044.

[80] Hu, L. L.; Huang, T.; Cai, Y. D.; Chou, K. C. Prediction of body fluids where proteins are secreted into based on protein interaction network. *PLoS One*, **2011**, *6*(7), E22989.

[81] Li, B. Q.; Huang, T.; Liu, L.; Cai, Y. D.; Chou, K. C. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PLoS One*, **2012**, *7*(4), E33393.

[82] Huang, T.; Wang, J.; Cai, Y. D.; Yu, H.; Chou, K. C. Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. *PLoS One*, **2012**, *7*(4), E34460.

[83] Chen, L.; Zeng, W. M.; Cai, Y. D.; Feng, K. Y.; Chou, K. C. Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One*, **2012**, *7*(4), E35254.

[84] Chou, K. C. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **2004**, *11*(16), 2105–2134.

[85] Xiao, X.; Wu, Z. C.; Chou, K. C. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.*, **2011**, *284*(1), 42–51.

[86] Chou, K. C.; Wu, Z. C.; Xiao, X. iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.*, **2012**, *8*(2), 629–641.

[87] Wu, Z. C.; Xiao, X.; Chou, K. C. iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. Biosyst.*, **2011**, *7*(12), 3287–3297.

[88] Wu, Z. C.; Xiao, X.; Chou, K. C. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex Gram-positive bacterial proteins. *Protein Pept. Lett.*, **2012**, *19*(1), 4–14.

[89] Chou, K. C.; Wu, Z. C.; Xiao, X. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS One*, **2011**, *6*(3), E18258.

[90] Chou, K. C.; Shen, H. B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nature Protoc.*, **2008**, *3*(2), 153–162.

[91] Wang, P.; Hu, L.; Liu, G.; Jiang, N.; Chen, X.; Xu, J.; Zheng, W.; Li, L.; Tan, M.; Chen, Z.; Song, H.; Cai, Y. D.; Chou, K. C. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS One*, **2011**, *6*(4), E18476.

[92] Shen, H. B.; Chou, K. C. HIVcleave: a web-server for predicting human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem.*, **2008**, *375*(2), 388–390.

[93] Chou, K. C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.*, **1993**, *268*(23), 16938–16948.

[94] Chou, K. C.; Shen, H. B. ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Commun.*, **2008**, *376*(2), 321–325.

[95] Zheng, L. L.; Li, Y. X.; Ding, J.; Guo, X. K.; Feng, K. Y.; Wang, Y. J.; Hu, L. L.; Cai, Y. D.; Hao, P.; Chou, K. C. A comparison of computational methods for identifying virulence factors. *PLoS One*, **2012**, *7*(8), E42517.

[96] Xiao, X.; Wang, P.; Chou, K. C. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different

modes of pseudo amino acid compositions. *Mol. Biosyst.*, **2011**, *7*(3), 911–919.

[97] Xiao, X.; Wang, P.; Chou, K. C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J. Comput. Chem.*, **2009**, *30*(9), 1414–1423.

[98] Lin, W. Z.; Xiao, X.; Chou, K. C. GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. *Protein Eng. Des. Sel.*, **2009**, *22*(11), 699–705.

[99] Bhasin, M.; Raghava, G. P. S. GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. *Nucleic Acids Res.*, **2005**, *33*(Web Server issue), W143–147.

[100] Chou, K. C. Prediction of G-protein-coupled receptor classes. *J. Proteome Res.*, **2005**, *4*(4), 1413–1418.

[101] Elrod, D. W.; Chou, K. C. A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Eng.*, **2002**, *15*(9), 713–715.

[102] Chou, K. C.; Elrod, D. W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.*, *1*(5), 429–433.

[103] Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discov.*, **2006**, *5*(12), 993–996.

[104] Veselovsky, A. V; Ivanov, A. S. Strategy of computer-aided drug design. *Curr. Drug Targets Infect. Disord.,* **2003**, *3*(1), 33–40.

[105] Liguory, C.; Coffin, J. C. Oral choledocoscopy after endoscopic sphincterotomy. *Nouv. Presse Med.*, **1979**, *8*(2), 136.

[106] Kalyaanamoorthy, S.; Chen, Y. P. P. Structure-based drug design to augment hit discovery. *Drug Discov. Today*, **2011**, *16*, (17–18), 831–839.

[107] Klebe, G. Recent developments in structure-based drug design. *J. Mol. Med. (Berl)*, **2000**, *78*(5), 269–281.

[108] Ferenczy, G. Structure-based drug design. *Acta Pharm. Hung.*, **1998**, *68*(1), 21–31.

[109] Singla, D.; Anurag, M.; Dash, D.; Raghava, G. P. S. A web server for predicting inhibitors against bacterial target GlmU protein. *BMC Pharmacol.*, **2011**, *11*, 5.

[110] Garg, A.; Tewari, R.; Raghava, G. P. S. KiDoQ: using docking based energy scores to develop ligand based model for predicting antibacterials. *BMC Bioinformatics*, **2010**, *11*, 125.

[111] Acharya, C.; Coop, A.; Polli, J. E.; Mackerell, A. D. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr. Comput. Aided Drug Des.*, **2011**, *7*(1), 10–22.

[112] Bacilieri, M.; Moro, S. Ligand-based drug design methodologies in drug discovery process: an overview. *Curr. Drug Discov. Technol.*, **2006**, *3*(3), 155–165.

[113] Dean, P. M.; Lloyd, D. G.; Todorov, N. P. De novo drug design: integration of structure-based and ligand-based methods. *Curr. Opin. Drug Discov. Devel.,* **2004**, *7*(3), 347–353.

[114] Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol.*, **2011**, *672*, 299–323.

[115] Dean, P. M. Chemical genomics: a challenge for de novo drug design. *Mol. Biotechnol.*, **2007**, *37*(3), 237–245.

[116] Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.*, **2005**, *4*(8), 649–663.

[117] Ihlenfeldt, W. D.; Bolton, E. E.; Bryant, S. H. The PubChem chemical structure sketcher. *J. Cheminform.*, **2009**, *1*(1), 20.

[118] Schüller, A.; Schneider, G.; Byvatov, E. SMILIB: Rapid Assembly of Combinatorial Libraries in SMILES Notation. *QSAR Comb. Sci.*, **2003**, *22*(7), 719–721.

[119] Truchon, J. F. GLARE: A tool for product-oriented design of combinatorial libraries. *Methods Mol. Biol.*, **2011**, *685*, 337–346.

[120] Lam, T. H.; Bernardo, P. H.; Chai, C. L. L.; Tong, J. C. CLEVER: A general design tool for combinatorial libraries. *Methods Mol. Biol.*, **2011**, *685*, 347–356.

[121] Tschinke, V.; Cohen, N. C. The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses. *J. Med. Chem.*, **1993**, *36*(24), 3863–3870.

[122] Junker, J. Statistical filtering for NMR based structure generation. *J. Cheminform.*, **2011**, *3*(1), 31.

[123] Liu, X.; Bai, F.; Ouyang, S.; Wang, X.; Li, H.; Jiang, H. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics*, **2009**, *10*, 101.

[124] Tosco, P.; Balle, T.; Shiri, F. SDF2XYZ2SDF: how to exploit TINKER power in cheminformatics projects. *J. Mol. Model.*, **2011**, *17*(11), 3021–3023.

[125] Liu, K.; Feng, J.; Young, S. S. PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. *J. Chem. Inf. Model.*, *45*(2), 515–522.

[126] He, Y.; Liew, C. Y.; Sharma, N.; Woo, S. K.; Chau, Y. T.; Yap, C. W. PaDEL-DDPredictor: Open-source software for PD-PK-T prediction. *J. Comput. Chem.*, **2012**.

[127] Truszkowski, A.; Jayaseelan, K. V.; Neumann, S.; Willighagen, E. L.; Zielesny, A.; Steinbeck, C. New developments on the cheminformatics open workflow environment CDK-Taverna. *J. Cheminform.*, **2011**, *3*, 54.

[128] Li, Z. R.; Han, L. Y.; Xue, Y.; Yap, C. W.; Li, H.; Jiang, L.; Chen, Y. Z. MODEL-molecular descriptor lab: a web-based server for computing structural and physicochemical features of compounds. *Biotechnol. Bioeng.*, **2007**, *97*(2), 389–396.

[129] Hattori, M.; Tanaka, N.; Kanehisa, M.; Goto, S. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res.*, **2010**, *38*(Web Server issue), W652–656.

[130] Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. Small Molecule Subgraph Detector (SMSD) toolkit. *J. Cheminform.*, **2009**, *1*(1), 12.

[131] Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **2010**, *31*(2), 455–461.

[132] Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, **2003**, *125*(7), 1731–1737.

[133] Koes, D. R.; Camacho, C. J. ZINCPharmer: pharmacophore search of the ZINC database. *Nucleic Acids Res.*, **2012**, *40*(Web Server issue), W409–414.

[134] Patlewicz, G.; Jeliazkova, N.; Safford, R. J.; Worth, A. P.; Aleksiev, B. An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ. Res.*, **2008**, *19*, (5–6), 495–524.

[135] Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: prediction of activity spectra for biologically active substances. *Bioinformatics*, **2000**, *16*(8), 747–748.

[136] Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P. W.; Tang, Y. admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties. *J. Chem. Inf. Model.*, **2012**.

[137] Lagorce, D.; Sperandio, O.; Galons, H.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs2: free ADME/tox filtering tool to assist drug discovery and chemical biology projects. *BMC Bioinformatics*, **2008**, *9*, 396.

[138] García-Sosa, A. T.; Oja, M.; Hetényi, C.; Maran, U. DrugLogit: logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties. *J. Chem. Inf. Model.*, **2012**, *52*(8), 2165–2180.

[139] Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.*, **2005**, *48*(22), 6970–6979.

[140] Liu, R.; Liu, J.; Tawa, G.; Wallqvist, A. 2D SMARTCyp reactivity-based site of metabolism prediction for major drug-metabolizing cytochrome P450 enzymes. *J. Chem. Inf. Model.*, **2012**, *52*(6), 1698–1712.

[141] Mishra, N. K.; Agarwal, S.; Raghava, G. P. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacol.*, **2010**, *10*, 8.

[142] Garg, A.; Bhasin, M.; Raghava, G. P. S. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.*, **2005**, *280*(15), 14427–14432.

[143] Kumar, M.; Verma, R.; Raghava, G. P. S. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *J. Biol. Chem.*, **2006**, *281*(9), 5357–5363.

[144] Xiao, X.; Wu, Z. C.; Chou, K. C. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. *PloS one*, **2011**, *6*(6), E20592.

[145] Bhasin, M.; Garg, A.; Raghava, G. P. S. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics*, **2005**, *21*(10), 2522–2524.

[146] Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. iDNA-Prot: identification of DNA binding proteins using random forest with grey model. *PLoS One*, **2011**, *6*(9), E24756.

[147] Lin, W. Z.; Fang, J. A.; Xiao, X.; Chou, K. C. Predicting Secretory Proteins of Malaria Parasite by Incorporating Sequence Evolution Information into Pseudo Amino Acid Composition via Grey System Model. *PLoS One*, **2012**, *7*(11), E49040.

[148] Chou, K. C.; Shen, H. B. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.,* **2007**, *357*(3), 633–640.

[149] Mishra, N. K.; Kumar, M.; Raghava, G. P. S. Support vector machine based prediction of glutathione S-transferase proteins. *Protein Pept. Lett.*, **2007**, *14*(6), 575–580.

[150] Chen, K.; Mizianty, M. J.; Kurgan, L. ATPsite: sequence-based prediction of ATP-binding residues. *Proteome Sci.*, **2011**, *9 SUPPL 1*, S4.

[151] Chauhan, J. S.; Mishra, N. K.; Raghava, G. P. S. Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC Bioinformatics*, **2010**, *11*, 301.

[152] Mishra, N. K.; Raghava, G. P. S. Prediction of FAD interacting residues in a protein from its primary sequence using evolutionary information. *BMC Bioinformatics*, **2010**, *11 SUPPL 1*, S48.

[153] Ansari, H. R.; Raghava, G. P. S. Identification of NAD interacting residues in proteins. *BMC Bioinformatics*, **2010**, *11*, 160.

[154] Chauhan, J. S.; Bhat, A. H.; Raghava, G. P. S.; Rao, A. GlycoPP: a webserver for prediction of N- and O-glycosites in prokaryotic protein sequences. *PLoS One*, **2012**, *7*(7), E40155.

[155] Julenius, K.; Mølgaard, A.; Gupta, R.; Brunak, S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*, **2005**, *15*(2), 153–164.

[156] Monigatti, F.; Gasteiger, E.; Bairoch, A.; Jung, E. The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics*, **2002**, *18*(5), 769–770.

[157] Chang, W. C.; Lee, T. Y.; Shien, D. M.; Hsu, J. B. K.; Horng, J. T.; Hsu, P. C.; Wang, T. Y.; Huang, H. D.; Pan, R. L. Incorporating support vector machine for identifying protein tyrosine sulfation sites. *J. Comput. Chem.*, **2009**, *30*(15), 2526–2537.

[158] Xue, Y.; Zhou, F.; Fu, C.; Xu, Y.; Yao, X. SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res.*, **2006**, *34*(Web Server issue), W254–257.

[159] Wong, Y. H.; Lee, T. Y.; Liang, H. K.; Huang, C. M.; Wang, T. Y.; Yang, Y. H.; Chu, C. H.; Huang, H. D.; Ko, M. T.; Hwang, J. K. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **2007**, *35*(Web Server issue), W588–594.

[160] Blom, N.; Sicheritz-Pontén, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **2004**, *4*(6), 1633–1649.

[161] Saha, S.; Zack, J.; Singh, B.; Raghava, G. P. S. VGIchan: prediction and classification of voltage-gated ion channels. *Genomics Proteomics Bioinformatics*, **2006**, *4*(4), 253–258.

[162] Saha, S.; Raghava, G. P. S. VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics*, **2006**, *4*(1), 42–47.

[163] Bhasin, M.; Raghava, G. P. S. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.,* **2004**, *279*(22), 23262–23266.

[164] Wang, P.; Xiao, X.; Chou, K. C. NR-2L: a two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. *PLoS One*, **2011**, *6*(8), E23505.

[165] Xiao, X.; Wang, P.; Chou, K. C. iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. *PLoS One*, **2012**, *7*(2), E30869.

[166] Chen, W.; Lin, H.; Feng, P. M.; Ding, C.; Zuo, Y. C.; Chou, K. C. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One*, **2012**, *7*(10), E47843.

[167] Garg, A.; Raghava, G. P. S. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol.*, **2008**, *8*(2), 129–140.

[168] Liu, Z.; Ren, J.; Cao, J.; He, J.; Yao, X.; Jin, C.; Xue, Y. Systematic analysis of the Plk-mediated phosphoregulation in eukaryotes. *Brief. Bioinform*, **2012**, BBS041–.

[169] Chu, W. Y.; Huang, Y. F.; Huang, C. C.; Cheng, Y. S.; Huang, C. K.; Oyang, Y. J. ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Res.*, **2009**, *37*(Web Server issue), W396–401.

[170] Gabdoulline, R.; Eckweiler, D.; Kel, A.; Stegmaier, P. 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. *Nucleic Acids Res.*, **2012**, *40*(Web Server issue), W180–185.

[171] Schug, J. Using TESS to predict transcription factor binding sites in DNA sequence. *Curr. Protoc. Bioinformatics*, **2008**, *CHAPTER 2*, UNIT 2.6.

[172] Hiard, S.; Marée, R.; Colson, S.; Hoskisson, P. A.; Titgemeyer, F.; Van Wezel, G. P.; Joris, B.; Wehenkel, L.; Rigali, S. PREDetector: a new tool to identify regulatory elements in bacterial genomes. *Biochem. Biophys. Res. Commun.*, **2007**, *357*(4), 861–864.

[173] Messeguer, X.; Escudero, R.; Farré, D.; Núñez, O.; Martínez, J.; Albà, M. M. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics*, **2002**, *18*(2), 333–334.

[174] Kumar, R.; Panwar, B.; Chauhan, J. S.; Raghava, G. P. Analysis and prediction of cancerlectins using evolutionary and domain information. *BMC Res. Notes*, **2011**, *4*, 237.

[175] Jenwitheesuk, E.; Wang, K.; Mittler, J. E.; Samudrala, R. PIR-Spred: a web server for reliable HIV-1 protein-inhibitor resistance/susceptibility prediction. *Trends Microbiol.*, **2005**, *13*(4), 150–151.

[176] Yuan, Y.; Pei, J.; Lai, L. LigBuilder 2: A Practical de Novo Drug Design Approach. *J. Chem. Inf. Model.*, **2011**.

[177] Vlachakis, D.; Tsagrasoulis, D.; Megalooikonomou, V.; Kossida, S. Introducing Drugster: a comprehensive and fully integrated drug design, lead and structure optimization toolkit. *Bioinformatics*, **2012**.

[178] Douguet, D. e-LEA3D: a computational-aided drug design web server. *Nucleic Acids Res.*, **2010**, *38*(Web Server issue), W615–621.

[179] Singla, D.; Tewari, R.; Kumar, A.; Raghava, G. P. Designing of inhibitors against drug tolerant Mycobacterium tuberculosis (H37Rv). *Chem. Cent. J.*, **2013**, *7*(1), 49.

**Received: December 12, 2012 Revised: January 20, 2013 Accepted: February 08, 2013**