BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 5. 579–591
Advance Access publication May 11, 2009
doi:10.1093/bib/bbp023

Recent advances in computer-aided drug design

Chun Meng Song, Shen Jean Lim and Joo Chuan Tong

Submitted: 2nd March 2009; Received (in revised form): 8th April 2009

Abstract

Modern drug discovery is characterized by the production of vast quantities of compounds and the need to examine these huge libraries in short periods of time. The need to store, manage and analyze these rapidly increasing resources has given rise to the field known as computer-aided drug design (CADD). CADD represents computational methods and resources that are used to facilitate the design and discovery of new therapeutic solutions. Digital repositories, containing detailed information on drugs and other useful compounds, are goldmines for the study of chemical reactions capabilities. Design libraries, with the potential to generate molecular variants in their entirety, allow the selection and sampling of chemical compounds with diverse characteristics. Fold recognition, for studying sequence-structure homology between protein sequences and structures, are helpful for inferring binding sites and molecular functions. Virtual screening, the *in silico* analog of high-throughput screening, offers great promise for systematic evaluation of huge chemical libraries to identify potential lead candidates that can be synthesized and tested. In this article, we present an overview of the most important data sources and computational methods for the discovery of new molecular entities. The workflow of the entire virtual screening campaign is discussed, from data collection through to post-screening analysis.

Keywords: *computer-aided drug design; virtual screening; computational modeling*

INTRODUCTION

Introduction of new therapeutic solutions is an expensive and time-consuming process. It is estimated that a typical drug discovery cycle, from lead identification through to clinical trials, can take 14 years [1] with cost of 800 million US dollars [2]. In the early 1990s, rapid developments in the fields of combinatorial chemistry and high-throughput screening technologies have created an environment for expediting the discovery process by enabling huge libraries of compounds to be synthesized and screened in short periods of time. However, these concerted efforts not only failed to increase the number of successfully launched new molecular entities, but seemingly aggravated the situation [3, 4]. Hit rates are often low and many of these identified hits fail to be further optimized into actual leads and preclinical [5–7]. Among the late-stage failures, 40–60% was reportedly due to absorption, distribution, metabolism, excretion and toxicity (ADME/Tox) deficiencies [8–10]. Collectively, these issues underscore the need to develop alternative strategies that can help remove unsuitable compounds before the exhaustion of significant amount of resources [7].

In time, a new paradigm in drug discovery came underway, calling for early assessment of potency (activity) and selectivity of lead candidates, as well as their potential ADME/Tox liabilities. This helps reduce costly late-stage failures and accelerates successful development of new molecular entities. At the core of this, paradigm shift is the application of

Corresponding author. Joo Chuan Tong, Data Mining Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01, Connexis South Tower, Singapore 138632. Tel: +65-64082156; Fax: +65-67761378; E-mail: victor@bic.nus.edu.sg

Chun Meng Song is a Research Officer at the Institute for Infocomm Research, Singapore, working on computational structural biology and virtual library design.

Shen Jean Lim is a Teaching Assistant at the Yong Loo Lin School of Medicine, National University of Singapore, working on bioinformatics methods and applications.

Joo Chuan Tong is a Principal Investigator at the Institute for Infocomm Research, Singapore and Adjunct Assistant Professor at the Yong Loo Lin School of Medicine, National University of Singapore. His research focuses on computational immunology, virtual library design, computational structural biology, and classification methods.

computational techniques to facilitate the discovery of new molecular entities. Computer-aided drug design (CADD) is a widely-used term that represents computational tools and resources for the storage, management, analysis and modeling of compounds. It includes development of digital repositories for the study of chemical interaction relationships, computer programs for designing compounds with interesting physicochemical characteristics, as well as tools for systematic assessment of potential lead candidates before they are synthesized and tested. The more recent foundations of CADD were established in the early 1970s with the use of structural biology to modify the biological activity of insulin [11] and to guide the synthesis of human haemoglobin ligands [12]. At that time, X-ray crystallography was expensive and time-consuming, rendering it infeasible for large-scale screening in industrial laboratories [13]. Over the years, new technologies such as comparative modeling based on natural structural homologues have emerged and began to be exploited in lead design [14]. These, together with advances in combinatorial chemistry, high-throughput screening technologies and computational infrastructures, have rapidly bridged the gap between theoretical modeling and medicinal chemistry. Numerous successes of designed drugs were reported, including Dorzolamide for the treatment of cystoid macular edema [15], Zanamivir for therapeutic or prophylactic treatment of influenza infection [16], Sildenafil for the treatment of male erectile dysfunction [17], and Amprenavir for the treatment of HIV infection [18].

CADD now plays a critical role in the search for new molecular entities [7, 13, 19]. Current focus includes improved design and management of data sources, creation of computer programs to generate huge libraries of pharmacologically interesting compounds, development of new algorithms to assess the potency and selectivity of lead candidates, and design of predictive tools to identify potential ADME/Tox liabilities. Here, we review major tools and resources that have been developed for expediting the search for novel drug candidates. The pipeline anatomy of a typical virtual screening campaign from data preparation to post-screening analysis is discussed.

DATA SOURCES

Data accessibility is critical for the success of a drug discovery and development campaign. Huge amounts of organic molecules, biological sequences and related information have been accumulated in scientific literature and case reports. These data are collected and stored in a structured way in a number of databases. Every year, hundreds of biological databases are described in [20]. At the same time, computational algorithms are actively developed to facilitate the design of combinatorial libraries. The most important data sources are reviewed in this section.

Small molecule databases

Small molecule databases represent a major resource for the study of biochemical interactions and play an increasing role in modern discovery with the accumulation of data. A variety of repositories of biologically interesting small molecules and their physicochemical properties have been compiled [21]. These include databases of known chemical compounds, drugs, carbohydrates, enzymes, reactants, natural products and natural-product-derived compounds (Table 1) [22, 23]. PubChem (http://pubchem.ncbi.nlm.nih.gov/), under the umbrella of National Institute of Health (NIH) Molecular Library Roadmap Initiative (http://nihroadmap.nih.gov/), provides information on the biological activities of more than 40 million small molecules and 19 million unique structures. The Available Chemicals Directory (ACD) from the Molecular Design Limited (http://www.mdli.com) serves as a central resource for docking studies. As of January 2009, the database details information of >571 000 purchasable compounds, while its screening compound counterpart Screening Compounds Directory stores over 4.5 million unique structures. ZINC [24], a free database of purchasable compounds, contains 20 089 615 3D structures of molecules annotated with biologically relevant properties (molecular weight, calculated Log *P* and number of rotatable bonds). LIGAND [22] provides records on 15 395 chemical compounds, 8031 drugs, 10 966 carbohydrates, 5043 enzymes, 7826 chemical reactions and 11 113 reactants (February 2009). DrugBank [25] stores detailed information on nearly 4800 drugs, including >1350 FDA-approved small molecule drugs, 123 FDA-approved biotech drugs, 71 nutraceuticals and >3243 experimental drugs. ChemDB [21] includes nearly 5 million commercially available compounds. Other small molecule databases exist and have been reviewed elsewhere [26].

Biological databases

Sequencing of the human and other model organism genomes have produced increasingly huge amounts of data relevant to the study of human disease. Some of these data sources are described in Table 2. The international collaborative GenBank [27], DNA Data Bank of Japan (DDBJ) [28] and European Molecular Biology Laboratory (EMBL) [29] serve as worldwide repositories for nucleotide sequences of diverse origins. The three databases synchronize their records on a daily basis. Swiss-Prot [30] and Protein Information Resource (PIR) [31] provide comprehensive and expertly annotated protein sequence and functional information. A total of 410 518 protein sequences are currently (February 2009) indexed by Swiss-Prot. Translated EMBL (TrEMBL), a computer-annotated protein sequence database supplement of Swiss-Prot, includes all translation of EMBL nucleotide sequences that are not available in the database [30]. Protein Data Bank (PDB) [32] is the single worldwide archive of structural data of biological macromolecules. As of February 2009, a total of 56 066 biological macromolecular structures have been deposited in PDB.

Apart from the wealth of information from general-purpose biological databases, a variety of specialist databases have also been developed. Collectively, these sources represent current accumulated knowledge on human biology and disease. Gene expression profiles provide hints of potential targets that may be signatures of diseases. For this, databases such as ArrayExpress [33], Gene Expression Omnibus (GEO) [34] and CIBEX [35] are popular repositories. In the field of proteomics, data from 2D gel electrophoresis have been deposited into resources

Table 1: Some small molecule databases reviewed in this article

Name	URL
PubChem	http://pubchem.ncbi.nlm.nih.gov/
ACD	http://www.mdli.com
ZINC	http://zinc.docking.org/
LIGAND	http://www.genome.jp/ligand/
DrugBank	http://www.drugbank.ca/
ChemDB	http://cdb.ics.uci.edu/

Table 2: Some biological databases reviewed in this article

Type	Name	URL
DNA sequences	GenBank	http://www.ncbi.nlm.nih.gov/Genbank/
	DDBJ	http://www.ddbj.nig.ac.jp/
	EMBL	http://www.embl-heidelberg.de/
Protein sequences	Swiss-Prot	http://www.expasy.ch/sprot/
	PIR	http://pir.georgetown.edu/
Protein structures	PDB	http://www.rcsb.org/pdb
Gene expression	ArrayExpress	http://www.ebi.ac.uk/microarray-as/ae/
	GEO	http://www.ncbi.nlm.nih.gov/geo/
	CIBEX	http://cibex.nig.ac.jp/index.jsp
2D gel electrophoresis	SWISS-2DPAGE	http://www.expasy.ch/ch2d/
	GELBANK	http://gelbank.anl.gov/
Mass spectrometry	OPD	http://bioinformatics.icmb.utexas.edu/OPD/
	GPMDB	http://www.thegpm.org/GPMDB/index.html
Metabolomics	HMDB	http://www.metabolomics.ca
	MDL Metabolite database	http://www.mdl.com/products/predictive/metabolite/index.jsp
	METLIN	http://metlin.scripps.edu/
Protein–protein interactions	BIND	www.bind.ca/
	HPRD	http://www.hprd.org/
	IntAct	www.ebi.ac.uk/intact/
Transcriptional regulation	TRANSFAC	http://www.biobase-international.com/pages/index.php?id=transfac
	TRED	rulai.cshl.edu/TRED/
Post-translational modifications	dbPTM	dbptm.mbc.nctu.edu.tw/
	RESID	www.ebi.ac.uk/RESID/
Biological pathways	KEGG	www.genome.jp/kegg/
	BioCarta	http://www.biocarta.com/

such as SWISS-2DPAGE [36] and GELBANK [37], while mass spectrometry data is available in databases such as Open Proteomics Database (OPD) [38] and Global Proteome Machine Database (GPMDB) [39]. Metabolomic databases, which detail information of biological pathways and their workings, are available through resources such as the Human Metabolite Database (http://www.metabolomics.ca; HMDB), MDL Metabolite database (http://www.mdl.com/products/predictive/metabolite/index.jsp) and METLIN [40]. The Biomolecular Interaction Network Database (BIND) [41], Human Protein Reference Database (HPRD) [42] and IntAct [43] provide data on protein–protein interactions while transcriptional relationships are available from resources like TRANSFAC [44] and TRED [45]. Post-translational modifications to proteins are also characterized and gathered in dbPTM [46], RESID [47], among others. Collectively, some of these pair-wise relationships have been abstracted into biologically related pathways and networks and made available through resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [48] and BioCarta pathways (http://www.biocarta.com/). These resources allow in-depth analysis of selected biomolecules and their roles in molecular pathways of disease. More detailed reviews are available elsewhere [20].

Virtual combinatorial libraries

Combinatorial chemistry is now a critical component of modern drug discovery. Often, such libraries are far too large to be synthesized or screened in their entirety. It is common that these resources may contain a large number of highly similar compounds in terms of their physicochemical characteristics. The potential for improved design that allows optimizing a library's diversity or similarity to a target can help minimize redundancy or maximize the number of discovered true leads. Concepts such as diversity, coverage and representativeness are commonly adopted to ensure a good sampling of a library using the minimum number of molecules [49]. Virtual library design usually begins with the explicit enumeration of all molecular variants within appropriate chemical spaces, followed by subsetting to allow good sampling of all products in the library [50]. Two approaches are normally used for the enumeration of molecular variants: (1) Markush techniques which attach a list of alternative functional groups to variable sites on a common scaffold [51], and (2) chemical transforms which specifies part of the reacting molecules that undergo chemical transformations and the nature of these transformations [52, 53]. These libraries may be optimized for molecular diversity or similarity using descriptors such as chemical composition, chemical topology, 3D structures and functionality [54], or drug-likeness using heuristic rules to detect ADME/Tox deficiencies [49].

COMPUTATIONAL MODELING, ANALYSIS, OPTIMIZATION AND PIPELINING

Fold recognition and comparative modeling

Fold recognition plays an integral role in modern drug discovery process, fueled by the rapid production of data from initiatives such as the Human Genome Project [55]. A potential drug target that is structurally similar to that of a well characterized protein with known biochemical function may help identify binding sites and molecular function [13]. Existing methods include sequence comparison and protein threading [56]. Sequence comparison typically involves searching a query sequence against a database of known protein sequences with experimentally defined 3D structures and evaluating the alignment using substitution matrices, gap penalties, or propensity scales [57–59]. In contrast, protein threading or side-chain conformation search involves substituting the backbone coordinates of a source structure with the probe sequence and assessing the plausibility of the model by means of a set of empirical potentials [60–62]. Such methods offer the potential to identify structurally conserved proteins with no evolutionary links, and are useful for modeling highly conserved receptor-ligand complexes. The search for T-cell epitopes and the subsequent design of peptide-based modalities exemplifies the applicability of this approach [63].

Once the structure of a homologous protein has been identified, a 3D model of the target structure may be constructed by comparative modeling, which provides a foundation for drug design by structures [64]. Such methods are based on the fact that the structures of evolutionary related proteins are more conserved than their primary sequences [65]. Hence, models of a protein with unknown structure may be constructed based on proteins with similar sequences

[66]. Successful model construction requires at least one experimentally solved 3D structure with significant amino acid sequence similarity to the target sequence [67]. It has been estimated that templates with 50% sequence identity can reliably generate 3D models with approximately 1 Å agreements between matched Cα atoms, while templates with 25% identity produces models with Cα root-mean-square-deviation (r.m.s.d.) of more than 2 Å [68]. A variety of techniques have been developed for model construction, such as fragment assembly, as implemented in SWISS-MODEL [67], COMPOSER [69] and 3D-JIGSAW [70], segment matching, as implemented in SegMod/ENCAD [71], spatials constraints, as implemented in Modeller [64], or ab initio prediction, as implemented in Rosetta [72]. Several inhibitors including human lipoxygenase inhibitors [73], kinase inhibitors [74] and cannabinoid CB2 receptor agonists [75] have been discovered using virtual screening with homology models.

Ligand selectivity

The discovery of new molecular entities for drug intervention is a highly combinatorial science due to the diversity of protein targets, as well as huge variability of possible lead candidates. The theoretical number of natural proteins is approximately 250 000 [76], while the number of real organic compounds with molecular weight <2000 Da is more than 10^{60} [77, 78]. Due to the astronomically huge chemical space, the cost required for systematic studies can be extraordinarily high. Computational tools are increasingly used as a cost-effective way for the selection, modeling, analysis and optimization of potential lead candidates. This section surveys the computational methods that have been developed for the prediction of ligand selectivity.

Receptor-based techniques

The availability of a protein target structure is usually helpful in identifying potential ligand interactors. Such approaches usually involve explicit molecular docking of ligands into the receptor binding site, producing a predicted binding mode for each candidate compound [79]. Predicting the preferred binding poses of ligands within a protein active site is difficult. First, the location and geometry of the binding site must be known, which may not always be addressed by X-ray crystallography or NMR studies [80]. Second, the method must find the correct positioning of a compound in the active site of the protein [81]. Incremental construction algorithms are potentially useful in guiding the search for optimal binding poses. Fragments are placed in the binding site of proteins and then 'grown' to fill the space available. An example of such approach has been reported by Rarey and colleagues [82], in which the conformational space of the ligand is sampled on the basis of a discrete model and a tree-search technique is used for extending the ligand within the active site. Boehm and coworkers [83] applied the use of needle screening to identify compounds that bind to the bacterial enzyme DNA gyrase ATP binding site. There are also an increasing number of reports on the use of Monte Carlo procedures for protein modeling and design. An early use of such procedure was described by Abagyan and Totrov [84], which randomly selects a conformational subspace and makes a step to a new position independent of the previous position, but according to the predefined continuous probability distribution. The use of conformational ensembles [85] and genetic algorithms [86] to predict the bound conformations of flexible ligands to macromolecular targets was also explored. Other docking algorithms exist and have been described elsewhere [87]. A comparative evaluation of eight docking programs (DOCK, FlexX, FRED, GLIDE, GOLD, SLIDE, SURFLEX and QXP) for their capacity to recover the X-ray pose of 100 small-molecular-weight ligands was reported [88]. It was found that at a 1 Å r.m.s.d. threshold, docking was successful for up to 63% of cases, while at an r.m.s.d. threshold of 2 Å, the maximum success rate was 90%.

Third, the system must evaluate the relative goodness-of-fit or how well the compound can bind to the receptor in comparison with other compounds [13]. An early venture was described by Platzer and colleagues [89], on calculating the relative standard free energy of binding of substrates to α-chymotrypsin. At that time, computational limitations did not allow the inclusion of solvation or entropic effects in the simulations. Since then, new methods have been devised which allow the basic handling of such configurations [90]. Physical-based potentials uses atomic force fields to model free energies of binding, and may be coupled with methods such as free energy perturbation (FEP) [91] and thermodynamic integration (TI) [92] for higher accuracy. Tools that implement physical-based

scoring methods include Assisted Model Building and Energy Refinement (AMBER) [93], Chemistry at HARvard Molecular Mechanics (CHARMM) [94] and DOCK [95]. Empirical-based potentials are fast and hence widely used in most docking algorithms. Such an approach requires the availability of receptor–ligand complexes with known binding affinity, and uses additive approximations of several energy terms such as van der Waals potential, electrostatic potential, hydrophobicity potential, among others, for binding free energy estimations [80]. Examples of tools that deploy empirical-based scoring functions include FlexX [82], SCORE [96], Internal Coordinate Mechanics (ICM) [84] and VALIDATE [97]. Knowledge-based methods, which are implemented in Potentials of Mean Force (PMF) [98] and DrugScore [99], compute binding free energies based on the frequencies of inter-atomic contacts. Such methods are also fast to compute and do not require availability of binding affinity data [79]. The Poisson–Boltzmann equation [100], which describes inter-molecular electrostatic interactions, has also been reportedly used for assessing the quality of a virtual screen.

Ligand-based techniques

Central to screening procedures based on ligands is the Similarity Property Principle [101], which asserts that molecules with similar structures are likely to share similar properties. This forms the basis for many ligand-based screening efforts where molecular structure and property descriptors of interacting molecules are extrapolated to search for other molecules with similar characteristics [54, 102]. For this, various machine learning techniques have been described, including the use of decision trees [103], recursive partitioning [104], artificial neural networks (ANN) [105] and support vector machines (SVM) [106, 107]. More recently, several groups have also demonstrated the use of mapping methods that transforms molecular features into various representations. For instance, Godden and coworkers [108] introduced the concept of Dynamic Mapping of Consensus positions (DMC) for mapping consensus positions of specific compound sets to binary-transformed chemical descriptor spaces, as well as the idea of Distance in Activity-Centered Chemical Space (DACCS) for accurately detecting molecular similarity relationships in 'raw' chemical spaces of high dimensionality [109]. Eckert *et al.* [110] introduced an extension of DMC, DynaMAD, which maps compounds to 'activity-class-dependent' descriptor values using unmodified descriptor value distributions. Molecular fingerprints based on 2D or 3D descriptors are also applied for virtual screening applications, as seen in methods such as MOLPRINT 2D [111], Property Descriptor value Range-derived FingerPrint (PDR-FP) [112], Rapid Overlay of Chemical Structures (ROCS) [113], shape fingerprints [114], and 3D pharmacophore fingerprints [115]. Dynamic activities over the past few years have also seen the development of hybrid techniques that integrate the strength of both structure-based and ligand-based techniques. For example, Cherkasov and coworkers [116] reported a combined approach integrating docking and structure-activity modeling using ANN to predict non-steroidal compounds that bind to human sex hormone binding globulin. Although useful in practice, ligand-based procedures are usually non-generalizable and highly dependent on the quantity and quality of available experimental data. Where there is limited data or biasness in the training dataset, these models suffer from poor accuracy. Reported successes of ligand-based virtual screening include the discovery of novel cyclooxygenase 2 (COX-2) inhibitors [106] and anti-malaria compounds [117].

Assessment of ADME/Tox deficiencies

The disposition of a pharmaceutical compound may be described by its pharmacokinetic or ADME properties [118]. In order to exert a pharmacological effect in tissues, a compound has to penetrate various physiological barriers, such as the gastrointestinal barrier, the blood–brain barrier and the microcirculatory barrier, to reach the blood circulation. It is subsequently transported to its effector site for distribution into tissues and organs, degraded by specialized enzymes, and finally removed from the body via excretion. In addition, genetic variation in drug metabolizing enzymes implies that some compounds may undergo metabolic activation and cause adverse reactions or Tox in humans [119]. Accordingly, the ADME/Tox properties of a compound directly impact its usefulness and safety.

The membrane permeability of a compound is determined by a combination of factors including compound size, aqueous solubility, ionizability (pKa) and lipophilicity (log *P*). It has been reported that the polar surface area (PSA) inversely correlates

with the lipid penetration ability [120]. Compounds that are completely absorbed by humans tend to have PSA values of $\leq 60\,\text{Å}^2$, while compounds with PSA $>140\,\text{Å}^2$ are less than 10% absorbed. Lipinski [121] carefully studied the physico-chemical properties of 2245 drugs from the World Drug Index (WDI) and found that poor absorption and permeation are more likely to occur when molecular weight <500 g/mol, Clog $P < 5$, hydrogen bond donors <5 and hydrogen bond acceptors <10. A 'rule of five' was subsequently proposed with respect to drug-likeness. These rules were extended by other researchers, including Ghose *et al*. [122] and Oprea [123]. A more stringent 'rule of five' was proposed by Wenlock *et al*. [124] after analyzing 594 compounds from the Physicians Desk Reference 1999, wherein molecular weight <473 g/mol, Clog $P < 5$, hydrogen bond donors <4 and hydrogen bond acceptors <7. Congreve and coworkers [125] performed an analysis on a range of targets derived by NMR and X-ray crystallography and found that the fragments obeyed, on average, a 'rule of three' for lead-likeness, in which molecular weight is <300 g/mol, hydrogen bond donors ≤ 3 and Clog $P \leq 3$. However, these rules could only serve as the minimal criteria for evaluating drug-likeness. It has been estimated that 68.7% of compounds in the Available Chemical Directory (ACD) Screening Database (2.4 million compounds) and 55% of compounds in ACD (240 000 compounds) do not violate the 'rule of five' [126]. A collection of 1203 compounds which represents 2973 pharmacokinetic measurements is now available in the PK/DB database [127]. The general rules for assessing ADME/Tox properties have been extended to more complex computational and mathematical models. Procedures based on genetic algorithms (GAs) [128, 129], ANNs [129, 130], SVMs [131] and statistical models [132, 133] have been widely used for predicting aqueous drug solubility and human intestinal absorption. Likewise, machine-learning algorithms and mathematical models for predicting Caco-2 permeability [133, 134] and blood–brain barrier penetration [135, 136] have been described. A comprehensive description of these methods can be found in a recent review [126]. Collectively, these systems allow detailed modeling of pharmacokinetics and drug delivery. The result will be realistic models that should match the complexities of external drug administration and greatly assist our understanding of the fate of compounds ingested or otherwise delivered externally to a human. An example is the successful screening for novel inhibitors of human carbonic anhydrase II using a series of hierarchical filters to reduce the initial data sample based on functional group requirements and pharmacophore matching [137].

Stereochemical quality assessment

Errors in protein structures may be identified by evaluating the stereochemical quality of generated models. A commonly used indicator of protein quality is the Ramachandran plot, which displays the φ and ψ backbone conformational angles for each residue in a protein [138]. Such an approach evaluates the correctness of structural coordinates based on standard deviations in φ and ψ angle pairs for residues in a protein. More complex forms of such metric exist, which incorporates addition parameters such as bond lengths for protein structure verification [139]. An alternative method for assessing protein stereochemical quality is to compare the model to its own amino-acid sequence using a 3D profile, computed from the atomic coordinates of the structure 3D profiles of correct protein structures, with its amino acid sequences [140]. A wrongly folded segment in a structure may be identified by examining the profile score in a moving-window scan. Appropriate use of these resources will enhance the quality of developed models and prediction accuracies.

Data pipelining

Data pipelining is increasingly applied to streamline and automate the process of virtual screening campaigns. In such systems, data automatically flow from one task to another allowing complete data analysis to be performed. This is achieved by constructing and executing workflows using components that perform specific data integration, calculation or analysis tasks. An example of workflow technology is Pipeline Pilot [141] developed at SciTegic, Inc. The system allows for the analysis of discovery data such as chemical series and high throughput screening results using an array of cheminformatics tools that include molecular fingerprints, similarity calculations, clustering, maximal common subgraph search and Bayesian model learning.

A ROADMAP FOR A STRUCTURE-BASED VIRTUAL SCREENING CAMPAIGN

In order to make sense of the wide assortment of tools available for virtual screening, we present a simplified solution as a roadmap, with a small set of selected options for each step (Figure 1). Specific tools and resources have been selected on the basis of their availability and performance. These tools have been shown to perform consistently either autonomously or as part of different analysis pipelines. As such, we wish to recommend them as a general procedure for small or large scale virtual screening initiatives.

A campaign usually begins with the selection of biological targets whose role in the disease pathway is established. The BLAST suite of programs [58] is usually helpful in inferring functional and evolutionary relationships between sequences, and helps in identifying members of gene families. Where available, the structures of target proteins may be downloaded from PDB [32] and the corresponding amino acid sequences from Swiss-Prot [30] or PIR [31]. PSI-BLAST [58], 3D-PSSM [142] and SAM-T2K [143] are useful for inferring the binding sites of target proteins. Next, it is necessary to prepare the library of compounds to be screened. Potential sources of publicly available chemical libraries include ZINC [24] and DrugBank [25]. Tools like Open Babel [144] and JOELib (http://sourceforge.net/projects/joelib/) are useful for inter-conversion of different chemical file formats such as PDB [32], Chemical Markup Language (CML) [145], MDL Molfile [146], simplified molecular input line entry specification (SMILES) [147] and SYBYL Line Notation (SLN) [148]. In addition, computational software exists for combinatorial library enumeration and 3D structure generation. SmiLib [149] allows the rapid combinatorial generation of chemical compounds by attaching different functional groups on a common chemical scaffold. The commercial software CORINA [150] or Converter (http://www.accelrys.com) are useful for generating the 3D structure of a small molecule. Tools such as CLEVER [151] support chemical library creation and manipulation, combinatorial chemical library enumeration using user-specified chemical components, chemical format conversion and visualization, as well as chemical compounds analysis and filtration with respect to drug-likeness, lead-likeness and fragment-likeness based on the physicochemical properties computed from the derived molecules.

For a target protein of unknown structure, a 3D model may be constructed with the help of SWISS-MODEL [67] or Modeller [64]. The de novo structure prediction algorithm, Rosetta, may be applied to predict the conformations of structurally divergent regions in comparative models [72]. Tools such as Relibase+ [152] are helpful for selecting conserved water molecules to be included into docking screens. WHAT_CHECK [139], PROCHECK [153] or Ramachandran Plot 2.0 [154] may be applied, before and after the screening process, to check stereochemical quality and identify errors in protein structures. Protonation of the target protein, energy minimization, and molecular docking may

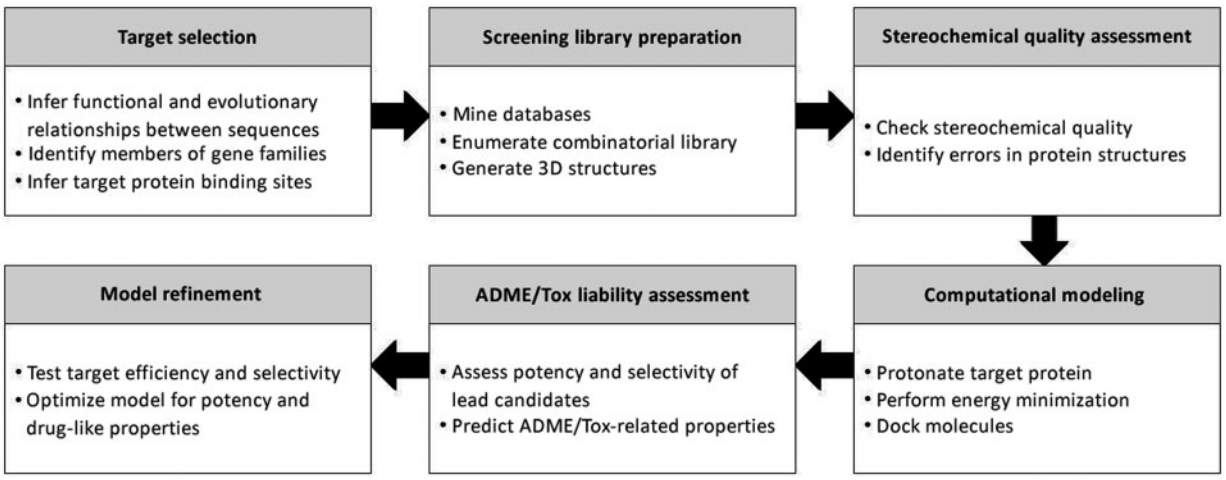

Figure 1: A roadmap for structure-based screening campaign, comprising of (i) target selection (ii) library preparation and (iii) stereochemical quality assessment, ADME/Tox assessment and computational optimization.

be performed using DOCK [85], AutoDock [86] or the commercial software ICM [84]. The commercial software ACD/Log*D* Suite (http://www.acdlabs.com/products/phys_chem_lab/logd/suite.html) by Advanced Chemistry Development can be used to predict ADME-related properties including hydrophobicity, lipophilicity and pKa, while Pharma Algorithms provide a suite of products via ADME Boxes (http://www.ap-algorithms.com/adme_boxes.htm) that addresses issues such as solubility, oral bioavailability, absorption and distribution. To remove toxic hits from the chemical libraries, counter pharmacophore screening may also be performed, using compounds whose inhibition leads to toxic effects [155]. This entire process may be iterated, with the inclusion of modified analogues or additional compound sets, for further optimization of potency and drug-like properties.

Key Points

- Numerous bioinformatics tools and resources have been developed to expedite drug discovery process. This article provides an overview of the most important data sources and computational methods for the discovery of new molecular entities. We have also provided guidelines on the workflow of the entire virtual screening campaign, from data collection through to post-screening analysis.
- Data accessibility is critical for the success of a drug discovery and development campaign. Small molecule databases represent a major resource for the study of biochemical interactions. Biological databases represent current accumulated knowledge on human biology and disease. Combinatorial libraries allow for optimization of a library's diversity or similarity to a target and can help minimize redundancy or maximize the number of discovered true leads.
- A campaign usually begins with the selection of biological targets whose role in the disease pathway is established. Next, it is necessary to prepare the library of compounds to be screened. For a target protein of unknown structure, a 3D model may be constructed using homology modeling techniques. This is usually followed by protonation of the target protein, energy minimization, molecular docking and stereochemical quality assessments.

CONCLUSION

Since the first reported success of discovery by design over three decades ago, there has been an explosion in the number, variety and sophistication of resources and analysis tools. CADD is now widely recognized as a viable alternative and complement to high-throughput screening. The search for new molecular entities has led to the construction of high quality datasets and design libraries that may be optimized for molecular diversity or similarity. On the other hand, advances in molecular docking algorithms, combined with improvements in computational infrastructure, are enabling rapid improvement in screening throughput. Propelled by increasingly powerful technology, distributed computing is gaining popularity for large-scale screening initiatives. Recent examples include the European Union funded WISDOM (World-wide In Silico Docking on Malaria) project which analyzed over 41 million malaria-relevant compounds in ~1 month using 1700 computers from 15 countries [155], and the Chinese funded Drug Discovery Grid (DDGrid) for anti-SARS and anti-diabetes research with a calculation capacity of >1 Tflops per second [156]. Combined with concerted efforts towards the design of more detailed physical models such as solubility and protein solvation, these advancements will, for the first time, allow the realization of the full potential of lead discovery by design.

References

1. Myers S, Baker A. Drug discovery—an operating model for a new era. *Nat Biotechnol* 2001;**19**:727–30.
2. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ* 2003;**22**:151–85.
3. Lahana R. How many leads from HTS? *Drug Discov Today* 1999;**4**:447–8.
4. Lobanov V. Using artificial neural networks to drive virtual screening of combinatorial libraries. *Drug Discov Today: BIOSILICO* 2004;**2**:149–56.
5. Hann MM, Leach AR, Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J Chem Inf Comput Sci* 2001;**41**:856–64.
6. Oprea TI, Davis AM, Teague SJ, *et al*. Is there a difference between leads and drugs? A historical perspective. *J Chem Inf Comput Sci* 2001;**41**:1308–15.
7. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* 2006;**11**:580–94.
8. Kennedy T. Managing the drug discovery/development interface. *Drug Discov Today* 1997;**2**:436–44.
9. Venkatesh S, Lipper RA. Role of the development scientist in compound lead selection and optimization. *J Pharm Sci* 2000;**89**:145–54.
10. Hou T, Xu X. Recent development and application of virtual screening in drug discovery: an overview. *Curr Pharm Des* 2004;**10**:1011–33.
11. Blundell TL, Dodson GG, Mercola D, *et al*. The structure, chemistry and biological activity of insulin. *Adv Protein Chem* 1972;**26**:279–402.
12. Beddell CR, Goodford PJ, Norrington FE, *et al*. Compounds designed to fit a site of known structure in human haemoglobin. *Br J Pharmacol* 1976;**57**:201–9.

13. Congreve M, Murray CW, Blundell TL. Structural biology and drug discovery. *Drug Discov Today* 2005;**10**:895–907.
14. Blundell TL. Structure-based drug design. *Nature* 1996;**384**: 23–6.
15. Grover S, Apushkin MA, Fishman GA. Topical dorzolamide for the treatment of cystoid macular edema in patients with retinitis pigmentosa. *Am J Ophthalmol* 2006;**141**:850–8.
16. Von Itzstein M, Wu WY, Kok GB, *et al*. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* 1993;**363**:418–23.
17. Terrett NK, Bell AS, Brown D, *et al*. Sildenafil (Viagra^{TM}), a potent and selective inhibitor of Type 5 CGMP phosphodiesterase with utility for the treatment of male erectile dysfunction. *Bioorg Med Chem Lett* 1996;**6**:1819–24.
18. Goodgame JC, Pottage JC Jr, Jablonowski H, *et al*. Amprenavir in combination with lamivudine and zidovudine versus lamivudine and zidovudine alone in HIV-1-infected antiretroviral-naive adults. Amprenavir PROAB3001 International Study Team. *Antivir Ther* 2000;**5**:215–25.
19. Muegge I, Oloff S. Advances in virtual screening. *Drug Discov Today Tech* 2006;**3**:405–11.
20. Galperin MY. The molecular biology database collection: 2007 update. *Nucleic Acids Res* 2007;**35**:D3–4.
21. Chen J, Swamidass SJ, Dou Y, *et al*. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* 2005;**21**:4133–9.
22. Goto S, Okuno Y, Hattori M, *et al*. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res* 2002;**30**:402–4.
23. Ortholand JY, Ganesan A. Natural products and combinatorial chemistry: back to the future. *Curr Opin Chem Biol* 2004;**8**:271–80.
24. Irwin JJ, Shoichet BK. ZINC— a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005;**45**:177–82.
25. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**:D668–72.
26. Jónsdóttir SO, Jørgensen FS, Brunak S. Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics* 2005;**21**:2145–60.
27. Benson DA, Karsch-Mizrarchi I, Lipman DJ, *et al*. GenBank. *Nucleic Acids Res* 2009;**34**:D16–20.
28. Tateno Y, Imanishi T, Miyazaki S, *et al*. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res* 2002;**30**:27–30.
29. Kanz C, Aldebert P, Althorpe N, *et al*. The EMBL nucleotide sequence database. *Nucleic Acids Res* 2005;**33**:D29–33.
30. O'Donovan C, Martin MJ, Gattiker A, *et al*. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform* 2002;**3**:275–84.
31. Wu CH, Huang H, Arminski L, *et al*. The protein information resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res* 2002;**30**:35–7.
32. Westbrook J, Feng Z, Jain S, *et al*. The protein data bank: unifying the archive. *Nucleic Acids Res* 2002;**30**:245–8.
33. Parkinson H, Kapushesky M, Shojatalab M, *et al*. ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 2005;**33**: D553–5.
34. Boyle J. Gene-Expression Omnibus integration and clustering tools in SeqExpress. *Bioinformatics* 2005;**21**: 2550–1.
35. Ikeo K, Ishi-i J, Tamura T, *et al*. CIBEX: center for information biology gene expression database. *C R Biol* 2003; **326**:1079–82.
36. Hoogland C, Mostaguir K, Sanchez JC, *et al*. SWISS-2DPAGE, ten years later. *Proteomics* 2004;**4**:2352–6.
37. Babnigg G, Giometti CS. GELBANK: a database of annotated two-dimensional gel electrophoresis patterns of biological systems with complete genomes. *Nucleic Acids Res* 2004;**32**:D582–5.
38. Prince JT, Carlson MW, Wang R, *et al*. The need for a public proteomics repository. *Nat Biotechnol* 2004;**22**: 471–2.
39. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 2004;**3**:1234–42.
40. Smith CA, O'Maille G, Want EJ, *et al*. METLIN: a metabolite mass spectral database. *Ther Drug Monit* 2005;**27**: 747–51.
41. Alfarano C, Andrade CE, Anthony K, *et al*. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* 2005;**33**:D418–24.
42. Peri S, Navarro JD, Kristiansen TZ, *et al*. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* 2004;**32**:D497–501.
43. Kerrien S, Alam-Faruque Y, Aranda B, *et al*. IntAct: open source resource for molecular interaction data. *Nucleic Acids Res* 2007;**35**:D561–5.
44. Wingender E. TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. *In Silico Biol* 2004;**4**:55–61.
45. Jiang C, Xuan Z, Zhao F, *et al*. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* 2007;**35**:D137–40.
46. Lee TY, Huang HD, Hung JH, *et al*. dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res* 2006;**34**:D622–7.
47. Farriol-Mathis N, Garavelli JS, Boeckmann B, *et al*. Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* 2004;**4**:1537–50.
48. Kanehisa M, Goto S, Kawashima S, *et al*. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004;**32**:D277–80.
49. Brown RD, Hassan M, Waldman M. Combinatorial library design for diversity, cost efficiency, and drug-like character. *J Mol Graph Model* 2000;**18**:427–37.
50. Agrafiotis DK, Lobanov VS, Salemme FR. Combinatorial informatics in the post-genomics era. *Nat Rev Drug Discov* 2002;**1**:337–46.
51. Leland BA, Christie BD, Nourse JG, *et al*. Managing the combinatorial expansion. *J Chem Inf Comput Sci* 1997;**37**: 62–70.
52. Leach AR, Bradshaw J, Green DVS, *et al*. Implementation of a system for reagent selection and library enumeration, profiling and design. *J Chem Inf Comput Sci* 1999;**39**: 1161–72.
53. Lobanov VS, Agrafiotis DK. Scalable methods for the construction and analysis of virtual combinatorial libraries. *Combin Chem High-Throughput Screen* 2002;**5**:167–78.

54. Livingston DJ. The characterization of molecular structures using molecular properties. A survey. *J Chem Inf Comput Sci* 2000;**40**:195–209.
55. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.
56. Mizuguchi K. Fold recognition for drug discovery. *Drug Discov. Today: TARGETS* 2004;**3**:18–23.
57. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 1991;**253**:164–70.
58. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
59. Rice DW, Eisenberg D. A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol* 1997;**267**: 1026–38.
60. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. *Nature* 1992;**358**:86–9.
61. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasichemical approximation. *Macromolecules* 1985;**18**:534–52.
62. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci* 1999;**8**:361–9.
63. Tong JC, Tan TW, Ranganathan S. Methods and protocols for predicting immunogenic epitopes. *Brief Bioinform* 2007; **8**:96–108.
64. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 1993;**234**:779–815.
65. Lesk AM, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globulins. *J Mol Biol* 1980;**136**: 225–70.
66. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. *Protein Sci* 2005;**14**: 1315–27.
67. Schwede T, Kopp J, Guex N, *et al*. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res* 2003;**31**:3381–85.
68. Chung SY, Subbiah S. A structural explanation for the twilight zone of protein sequence homology. *Structure* 1996;**4**: 1123–7.
69. Sutcliffe MJ, Haneef I, Carney D, *et al*. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng* 1987;**1**:377–84.
70. Bates PA, Kelley LA, MacCallum RM, *et al*. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins* 2001;**5**(Suppl):39–46.
71. Levitt M. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol* 1992;**226**:507–33.
72. Roh CA, Strauss CE, Chivian D, *et al*. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* 2004;**55**:656–77.
73. Kenyon V, Chorny I, Carvajal WJ, *et al*. Novel human lipogenase inhibitors discovered using virtual screening with homology models. *J Med Chem* 2006;**49**:1356–63.
74. Diller DJ, Li R. Kinases, homology models, and high-throughput docking. *J Med Chem* 2003;**46**:4638–47.
75. Salo OM, Raitio KH, Savinainen JR, *et al*. Virtual screening of novel CB2 ligands using a comparative model of the human Cannabinoid CB2 receptor. *J Med Chem* 2005;**48**: 7166–71.
76. O'Donovan C, Apweiler R, Bairoch A. The human proteomics initiative (HPI). *Trends Biotechnol* 2001;**19**: 178–81.
77. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modelling perspective. *Med Res Rev* 1996;**16**:3–50.
78. Martin YC. Challenges and prospects for computational aids to molecular diversity. *Perspect Drug Discov Design* 1997;**7**(8):159–72.
79. Lyne PD. Structure-based virtual screening: an overview. *Drug Discov Today* 2002;**7**:1047–55.
80. Fernández-Recio J, Totrov M, Abagyan R. Identification of protein–protein interaction sites from docking energy landscapes. *J Mol Biol* 2004;**335**:843–65.
81. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. *J Comput Aided Mol Des* 2002;**16**:151–66.
82. Rarey M, Kramer B, Lengauer T, *et al*. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 1996;**261**:470–89.
83. Boehm HJ, Boehringer M, Bur D, *et al*. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J Med Chem* 2000;**43**:2664–74.
84. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic caculations for peptides and proteins. *J Mol Biol* 1994;**21**:983–1002.
85. Lorber DM, Shoichet BK. Flexible ligand docking using conformational ensembles. *Protein Sci* 1998;**7**:938–50.
86. Morris GM, Goodsel DS, Halliday RS, *et al*. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comp Chem* 1998;**19**:1639–62.
87. Halperin I, Ma B, Wolfson H, *et al*. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* 2002;**47**:409–43.
88. Kellenberger E, Rodrigo J, Muller P, *et al*. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* 2004;**57**:225–42.
89. Platzer KE, Momany FA, Scheraga HA. Conformational energy calculations of enzyme-substrate interactions. II. Computation of the binding energy for substrates in the active site of alpha-chymotrypsin. *Int J Peptide Protein Res* 1972;**4**:201–19.
90. Kitchen DB, Decornez H, Furr JR, *et al*. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 2004;**3**: 935–49.
91. Pearlman DA, Kollman PA. A new method for carrying out free energy perturbation calculations: dynamically modified windows. *J Chem Phys* 1989;**90**:2460–70.
92. Kollman PA. Free energy calculations: application to chemical and biochemical phenomena. *Chem Rev* 1993;**93**: 2395–417.

93. Case DA, Cheatham TE, 3rd, Darden T, *et al*. The Amber biomolecular simulation programs. *J Comput Chem* 2005;**26**: 1668–88.
94. Brooks BR, Bruccoleri RE, Olafson BD, *et al*. CHARMM: a program for macromolecular energy, minimization, and dynamics calculation. *J Comput Chem* 1983;**4**:187–217.
95. Ewing TJ, Makino S, Skillman AG, *et al*. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* 2001;**15**:411–28.
96. Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* 1994;**8**:243–56.
97. Head RD, Smythe ML, Oprea TI, *et al*. VALIDATE: a new method for the receptor-based prediction of binding affinities of novel ligands. *J Am Chem Soc* 1996;**118**:3959–69.
98. Muegge I. A knowledge-based scoring function for protein-ligand interactions: probing the reference state. *Perspect Drug Des Discov* 2000;**20**:99–114.
99. Gohkle H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* 2000;**295**:337–56.
100. Fogolari F, Brigo A, Molinari H. The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J Mol Recognit* 2002;**15**:377–92.
101. Johnson M, Maggiora GM. *Concepts and Applications of Molecular Similarity*. New York: Wiley, 1990.
102. Bajorath J. Selected concepts and investigations in compounds classification, molecular descriptor analysis, and virtual screening. *J Chem Inf Comput Sci* 2001;**41**:233–45.
103. Wagener M, van Geerestein VJ. Potential drugs and non-drugs: prediction and identification of important structural features. *J Chem Inf Comput Sci* 2000;**40**:280–92.
104. Hawkins DM, Young SS, Rusinko A 3rd. Analysis of a large structure-activity data set using recursive partitioning. *Quant Struct-Active Relat* 1997;**16**:296–302.
105. Plewczynski D, Spieser SA, Koch U. Assessing different classification methods for virtual screening. *J Chem Inf Model* 2006;**46**:1098–106.
106. Franke L, Byvatov E, Werz O, *et al*. Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J Med Chem* 2005;**48**: 6997–7004.
107. Jorissen RN, Gilson MK. Virtual screening of molecular databases using a support vector machine. *J Chem Inf Model* 2005;**45**:549–61.
108. Godden JW, Furr JR, Xue L, *et al*. Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-transformed chemical descriptor spaces with variable dimensionality. *J Chem Inf Comput Sci* 2004;**44**: 21–9.
109. Godden JW, Bajorath J. A distance function for retrieval of active molecules from complex chemical space representations. *J Chem Inf Model* 2006;**46**:1094–7.
110. Eckert H, Vogt I, Bajorath J. Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: design of DynaMAD and comparison with MAD and DMC. *J Chem Inf Model* 2006;**46**:1623–34.
111. Bendel A, Mussa HY, Glen RC, *et al*. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J Chem Inf Comput Sci* 2004;**44**:1708–18.
112. Eckert H, Bajorath J. Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *J Chem Inf Model* 2006;**46**: 2515–26.
113. Rush TS 3rd, Grant JA, Mosyak L, *et al*. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* 2005;**48**: 1489–95.
114. Haigh JA, Pickup BT, Grant JA, *et al*. Small molecule shape-fingerprints. *J Chem Inf Model* 2005;**45**:673–84.
115. Saeh JC, Lyne PD, Takasaki BK, *et al*. Lead hopping using SVM and 3D pharmacophore fingerprints. *J Chem Inf Model* 2005;**45**:1122–33.
116. Cherkasov A, Shi Z, Fallahi M, *et al*. Successful in silico discovery of novel nonsteroidal ligands for human sex hormone binding globulin. *J Med Chem* 2005;**48**:3203–13.
117. Marrero-Ponce Y, Iyarreta-Veitía M, Montero-Torres A, *et al*. Ligand-based virtual screening and in silico design of new antimalarial compounds using nonstochastic and stochastic total and atom-type quadratic maps. *J Chem Inf Model* 2005;**45**:1082–100.
118. Tetko IV, Bruneau P, Mewes HW, *et al*. Can we estimate the accuracy of ADME-Tox predictions? *Drug Discov Today* 2006;**11**:700–7.
119. Gardiner SJ, Begg EJ. Pharmacogenetics, drug-metabolizing enzymes, and clinical practice. *Pharmacol Rev* 2006;**58**: 521–90.
120. Palm K, Stenberg P, Luthman K, *et al*. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharma Res* 1997;**14**:568–71.
121. Lipinski CA. Drug-like properties and the causes of poor solubility and poot permeability. *J Pharmacol Toxicol* 2000;**44**: 235–49.
122. Ghose AK, Viswanadhan VN, Wendoloski JJ. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery: Part 1. A qualitative and quantitative characterization of known drug databases. *J Comb Chem* 1999;**1**:55–68.
123. Oprea T. Property distribution of drug-related chemical databases. *J Comput Aided Mol Des* 2000;**14**:251–64.
124. Wenlock MC, Austin RP, Barton P, *et al*. A comparison of physicochemical property profiles of development and marketed oral drugs. *J Med Chem* 2003;**46**:1250–6.
125. Congreve M, Carr R, Murray C, *et al*. A 'rule of three' for fragment-based lead discovery? *Drug Discov Today* 2003;**8**: 876–7.
126. Hou T, Wang J, Zhang W, *et al*. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr Med Chem* 2006;**13**:2653–67.
127. Moda TL, Torres LG, Carrara AE, *et al*. PK/DB: database for pharmacokinetic properties and predictive *in silico* ADME models. *Bioinformatics* 2008;**24**:2270–71.
128. Wessel MD, Jurs PC, Tolan JW, *et al*. Prediction of human intestinal absorption of drug compounds from molecular structure. *J Chem Inf Comput Sci* 1998;**38**:726–35.
129. Wegner JK, Fröhlich H, Zell A. Feature selection for descriptor based classification models: Part 2. Human intestinal absorption (HIA). *Chem Inf Comput Sci* 2004;**44**: 931–9.

130. Agatonovic-Kustrin S, Beresford R, Yusof AP. Theoretically derived molecular descriptors important in human intestinal absorption. *J Pharmaceut Biomed Anal* 2001;**25**:227–37.

131. Xue Y, Li ZR, Yap CW, *et al*. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J Chem Inf Comput Sci* 2004;**44**: 1630–8.

132. Klopman G, Stefan LR, Saiakhov RD. A computer model for the prediction of intestinal absorption in human. *Eur J Pharm Sci* 2002;**17**:253–63.

133. Norinder U, Osterberg T, Artursson P. Theoretical calculation and prediction of intestinal absorption of drugs in humans using MolSurf parametrization and PLS statistics. *Eur J Pharm Sci* 1999;**8**:49–56.

134. Fujiwara S, Yamashita F, Hashida M. Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network. *Int J Pharm* 2002;**237**:95–105.

135. Ajay, Bemis GW, Murcko MA. Designing libraries with CNS activity. *J Med Chem* 1999;**42**:4942–51.

136. Clarke DE. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena: Part 2. Prediction of blood–brain barrier penetration. *J Pharm Sci* 1999;**88**:815–21.

137. Grüneberg S, Stubbs MT, Klebe G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J Med Chem* 2002; **45**:3588–602.

138. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* 1963;**7**:95–9.

139. Hooft RW, Vriend G, Sander C, *et al*. Errors in protein structures. *Nature* 1996;**381**:272.

140. Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. *Nature* 1992;**356**: 83–5.

141. Hassan M, Brown RD, Varma-O'brien S, *et al*. Cheminformatics analysis and learning in a data pipelining environment. *Mol Disvers* 2006;**10**:283–99.

142. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 2000;**299**:499–520.

143. Karchin R, Cline M, Mandel-Gutfreund Y, *et al*. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* 2003; **51**:504–14.

144. Guha R, Howard MT, Hutchison GR, *et al*. The blue obelisk—interoperability in chemical informatics. *J Chem Inf Model* 2006;**46**:991–998.

145. Murray-Rust P, Rzepa HS, Wright M. Development of Chemical Markup Language (CML) as a system for handling complex chemical content. *New J Chem* 2001;**4**: 618–34.

146. Dalby A, Nourse JG, Hounshell D, *et al*. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J Chem Inf Comput Sci* 1992;**32**:244–55.

147. Weininger D. SMILES: a chemical language and information: Part 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6.

148. Ash S, Cline MA, Homer W, *et al*. SYBYL Line Notation (SLN): a versatile language for chemical structure representation. *J Chem Inf Comput Sci* 1997;**37**:71–9.

149. Schueller A, Haehnke V, Schneider G. SmiLib v2.0: a Java-based tool for rapid combinatorial library enumeration. *QSAR Comb Sci* 2007;**26**:407–10.

150. Sadowski J. A hybrid approach for ring flexibility in 3D database searching. *J Comput-Aided Mol Design* 1997;**11**: 53–60.

151. Song CM, Bernardo PH, Chai CL, *et al*. CLEVER: pipeline for designing in silico chemical libraries. *J Mol Graph Model* 2009;**27**:578–83.

152. Hendlich M. Databases for protein-ligand complexes. *Acta Crystallogr D* 1998;**54**:1178–1182.

153. Laskowski RA, MacArthur MW, Moss DS, *et al*. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* 1993;**26**:283–91.

154. Gopalakrishnan K, Sowmiya G, Sheik SS, *et al*. Ramachandran plot on the web (2.0). *Protein Pept Lett* 2007;**14**:669–71.

155. Ananthula RS, Ravikumar M, Pramod AB. Strategies for generating less toxic P-selectin inhibitors: pharmacophore modeling, virtual screening and counter pharmacophore screening to remove toxic hits. *J Mol Graph Model* 2008;**27**: 546–57.

156. Jacq N, Breton V, Chen H-Y, *et al*. Virtual screening on large scale grids. *Parallel Comput* 2007;**33**:289–301.