Novel technologies for virtual screening

Thomas Lengauer, Christian Lemmen, Matthias Rarey and Marc Zimmermann

There are several methods for virtual screening of databases of small organic compounds to find tight binders to a given protein target. Recent reviews in *Drug Discovery Today* have concentrated on screening by docking and by pharmacophore searching. Here, we complement these reviews by focusing on virtual screening methods that are based on analyzing ligand similarity on a structural level. Specifically, we concentrate on methods that exploit structural properties of the complete ligand molecules, as opposed to using just partial structural templates, such as pharmacophores. The *in silico* procedure of virtual screening (VS) and its relationship to the experimental procedure, HTS, is discussed, new developments in the field are summarized and perspectives on future research are offered.

Thomas Lengauer
Max-Planck Institute for Informatics
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany
e-mail: lengauer@mpi-sb.mpg.de
Christian Lemmen
BioSolveIT GmbH
An der Ziegelei 75
53757 Sankt Augustin
Germany
Matthias Rarey
Center for Bioinformatics
University of Hamburg
Bundesstraße 43
20146 Hamburg
Germany
Marc Zimmermann
FhG Institute for Algorithms and Scientific Computing
Schloss Birlinghoven
53754 Sankt Augustin
Germany

▼ Virtual (database) screening (VS) is an increasingly important component of the computer-based search for novel lead compounds. There are, fundamentally, two approaches to the general problem: 'VS by docking', which requires knowledge of the 3D structure of the target protein binding site to prioritize compounds by their likelihood to bind to the protein; and 'similarity-based VS', where no information on the protein is necessary – instead, one or more compounds that are known to bind to the protein are used as a structural query. The screening procedure extracts compounds from the database according to an appropriate similarity criterion. In order for the screening procedure to be effective, this criterion should regard molecules that bind tightly to the same proteins as similar.

Docking has been the subject of several recent reviews [1–7]. Here, we complement these reviews with a summary on new developments in VS methods, based on molecular similarity. Methods based on pharmacophore generation and search (recently reviewed in refs [8,9]) have been excluded from the review. Here, we focus on methods that analyze the structure of the complete ligand molecule. In particular, we review methods that have the potential to handle large sets of ligands (thousands to millions).

Small-molecule alignment

The most accurate but also the most comprehensive approach to ligand-based VS is a detailed computational analysis of the structures of the two molecules to be compared. Take the docking paradigm as a starting point: rather than placing the ligand into the binding site, which, in this scenario, we have no access to, a compound that is known to bind to the target protein (such as the natural substrate or another inhibitor) is used as a 'reference molecule'. During screening, the molecules from the compound database (here, called 'test molecules') are superposed onto the reference molecule. In terms of the well-known lock-and-key principle, we compare two flexible keys here.

The superposition places chemically similar parts of the molecules on top of each other with a preference on aligning groups of the molecules that can participate in the same kind of short-range interactions (such as H-bonds). The 'FlexS system' [10] is based on such an approach. As the name suggests, the superposition method is a variant of the sibling docking procedure, FlexX [11]. FlexS keeps the reference molecule rigid and considers the test molecule as flexible. It offers several alternative superpositions for each molecule pair and rank-orders them, according to a similarity score. The software requires

1359-6446/04/$ – see front matter PII: S1359-6446(03)02939-8

Box 1. Benchmark datasets for virtual screening

The only publicly available source of a large dataset with biological assay data are compiled by the National Cancer Institute and available at: http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html

FlexS-77 is a benchmark dataset for ligand-superposition methods, comprising 77 ligands that have been selected from inhibitors of 14 proteins, such that the protein–ligand complex is structurally resolved. The dataset is available at: http://www.biosolveit.de/software/flexs/html/dl-datasets.html

FlexX-200 is a benchmark dataset for protein–ligand docking, comprising 200 protein–ligand complexes. This dataset is available at: http://www.biosolveit.de/software/flexx/html/dl-datasets.html

The Cambridge Crystallographic Data Centre and Astex have provided another curated benchmark dataset for protein–ligand docking, comprising 305 protein–ligand complexes (http://www.ccdc.cam.ac.uk/prods/validation_set/).

about 30 s for a superposition on a single CPU, and returns a reasonably accurate superposition in approximately 70% of the cases on a benchmark dataset (Box 1 and Ref. [10]; for an example, see Figure 1). The generally accepted standard of truth here (i.e. the molecular superposition that we want to rediscover by the computational method) is the superposition that the two molecules attain inside the same binding site of the target protein (available only for test cases). Enrichment factors of between 10- and 55-fold have been reported using FlexS [12,13].

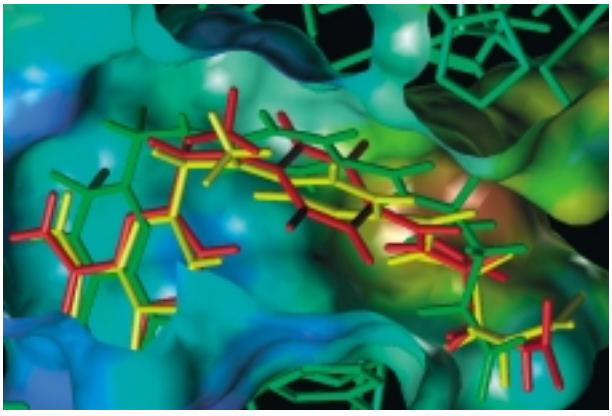

Figure 1. Superposition of methotrexate (MTX) and dihydrofolate (DHF) inside the binding pocket of dihydrofolate reductase. Green: MTX (crystal),Yellow: DHF (crystal) , Red: Superposition of DHF onto crystal MTX by FlexS.

Jones *et al.* have developed GASP, a molecular superposition procedure that is based on a genetic algorithm [14]. It is slower than FlexS but has the advantage that it can handle both the reference and the test molecule, flexibly. This can be of help in cases where no conformational preferences can be assigned to either of the molecules. Wild *et al.* optimized the alignment of molecular electrostatic potential (MEP) fields describing the two molecules [15]. For this they also used genetic algorithms.

In their programme MIMIC, Mestres *et al.* [16] follow a relatively typical approach, in which molecules are represented as sets of Gaussian functions, modeling fields of properties. The alignment is derived by gradient-based optimization of a scoring function that assesses the overlap of the respective fields. The most widely used functions for this purpose are the Carbo- [17] and the Hodgkin-Indices [18], respectively.

Other field-based approaches use a 3D grid around a molecule and calculate certain location-specific properties for each grid point. Peter Goodford's GRID program [19] is certainly the pioneering approach in this direction. It calculates an interaction potential with a virtual probe atom at the grid location. To perform a comparison, either the fields of two molecules or the molecules themselves have to be aligned in 3D space.

More recently, Krämer *et al.* and Pitman *et al.* reported on fFlash, a method for 3D database screening that is based on molecular superposition [20,21]. The approach is fragment-based. Adjacent pairs of fragments in the two molecules to be compared are conformationally sampled and the resulting variety of presenting binding features is stored in a lookup table. Virtual screening using a query compound then constitutes an on-the-fly reassembly of the fragmented compounds, using the wide-spread, graph-based procedure of clique detection on the feature patterns in the lookup table.

Hahn *et al.* describe ligand-based virtual screening that is performed purely on the basis of shape-filtering [22]. Similarly, a more recent approach by Putta *et al.* [23] rapidly detects shape matches, using a rough alignment that is derived from second-order moments of the conformer shape, followed by a binary comparison of steric occupancy on a grid. In the latest incarnation of this program, partial shape matches can be found via a more elaborate scheme of local moments and feature-type annotation of the shapes [24]. More detailed recent reviews on small-molecule alignment can be found elsewhere [25,26].

Descriptor-based screening: molecular topology as an efficient descriptor

Ligand superposition is computer-intensive: a single molecular comparison takes at least several seconds. To facilitate

searching through large chemical databases for molecules that are similar to a given query molecule, representations that allow for a much more time-efficient comparison (or even indexing of the database) are of central importance. The selection of molecular descriptors that are suitable for this purpose can either be structured by the type of molecule information used (macroscopic, topological or 2D, or 3D) or by the structure of the final descriptor (scalar, linear or non-linear). Owing to the wide application of quantitative SAR (QSAR), an enormous number of different descriptors are available.

Perhaps the oldest way of expressing similarity between molecules is on the basis of values such as molecular weight, log P and so on. The industry-wide success of Lipinski's 'Rule of 5' [27], using four whole-molecule properties (weight, log P, number of H-bond donors and H-bond acceptors), is an impressive example of how useful whole molecule descriptors can be in drug design.

A more detailed view of the molecular structure is captured by linear descriptors that encode many structural properties of the molecule in a binary- or real-valued vector. Each position in the vector stands for a property, such as the presence of some specific functional group or the occurrence of two specific atoms a specific number of bonds apart. (e.g. for binary descriptors, a 1 or a 0 at a particular position of the vector signifies that the molecule either does or does not have the corresponding property, respectively). Frequently used descriptors, such as MACCS structural keys (MAACS II, MDL Information Systems; http://www.mdli.com), Daylight (DAYLIGHT Software Manual, Daylight; http://www.daylight.com) or Unity Fingerprints (UNITY, Chemical Information Software 4.0, Tripos; http://www.tripos.com), are based on this concept. Instead of binary descriptors, numerical properties, such as the number of occurrences of an atom in a molecule, might be used. Such numerical vectors are also called holograms [28]. There are many such properties, leading to descriptor vectors that have lengths of many thousands, even millions, in the case of some pharmacophore descriptors (e.g. Ref. [29]). Once a vector representation is created, the similarity between two molecules can be expressed by functions like the Tanimoto or cosine coefficient, or even the Euclidean distance [30]. Owing to the simplicity of these functions, only a few CPU cycles are necessary for a similarity calculation.

Instead of considering topology only, the 3D properties of a molecule can be encoded in a linear descriptor (such as the distance between functional groups, see [31,32], or elementary shape properties, like shortest and longest distances along PCA axes [20]). 3D information usually pertains to a fixed conformation. However, molecules are generally flexible and the conformation that they adopt in the binding site of the receptor protein might be quite different from the one that was taken to compute the 3D descriptor information. One approach to solving this problem is to generate multiple conformations and combine the 3D information [29]; however, this has its disadvantages, because the 3D features in a descriptor are assumed to exist simultaneously – an erroneous assumption, if they originate from different conformations. This is one reason why 2D descriptors frequently outperform 3D descriptors [33–35].

One disadvantage of the 2D molecular descriptors that have been described so far, is that they cannot adequately gauge the molecular similarity of pairs of compounds that are structurally different but behave alike in binding to a protein [36]. Such a similarity measure is needed if 'scaffold hopping' (i.e. moving from a given compound in one structural class to one that has similar binding behaviour but is structurally significantly different) is to be performed. Molecules with a different scaffold to the query molecule are the typical starting point for a new lead series. Therefore, scaffold hopping is a basic requirement for VS procedures, not least for circumventing patent claims.

In the search for such a similarity measure, Rarey and Dixon [37] developed a descriptor that, besides properties of the 2D formula, also captures the relative arrangement of functional groups in the molecule. Their 'Feature Tree' (FTree) descriptor constitutes a graph (tree) that captures the overall topology of the molecule. In contrast to the reduced graph approach, proposed earlier [38–40], FTrees produce a more detailed description of physicochemical properties and are combined with a matching procedure that calculates optimal assignments at a higher resolution [37]. The nodes of the tree represent functional groups of the molecule; edges connect nodes as in the chemical structure. Several physico-chemical properties, or features, are associated with each node. The similarity of two molecules is determined by matching of the two corresponding trees (Figure 2). A numerical similarity score quantifies the quality of the matching, ranging from zero (completely dissimilar) to one (identical). The matching relates the nodes in one tree onto nodes in the second tree, preserving the topology of the trees and maximizing the similarity score.

Tree mapping is computationally more difficult than vector comparison; thus, the throughput of FTrees is two to three orders of magnitude slower (a few hundred per second) than that of conventional linear descriptors (tens of thousands per second). However, the Feature Tree descriptor is more accurate in describing properties that are

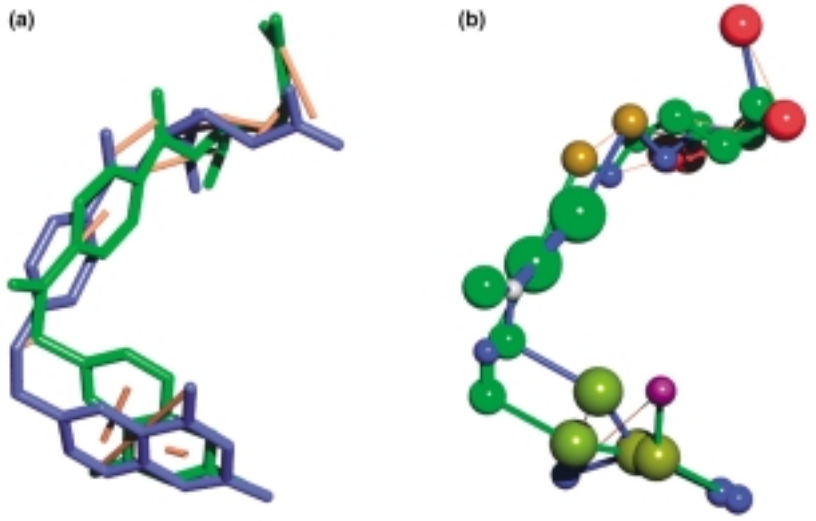

Figure 2. (a) The molecules methotrexate and dihydrofolate superposed in the binding pocket of dihydrofolate reductase. **(b)** the corresponding superposed FTrees. Straight lines represent the mapping of corresponding features in both molecules and FTrees, respectively.

relevant for binding and is consequently less dependent on the molecular topology [41]. This increases the potential for identifying new structural classes of active compounds.

Tree-like descriptors are not the only non-linear descriptors. To describe the relative orientation of interacting groups, field-based approaches are in use. The associated alignment problem (as discussed previously) is certainly the major obstacle for large-scale applications. To avoid the alignment step, procedures are used to convert the nonlinear descriptor back into a linear one; one possibility is to extract characteristic features of the field description [36,42].

In principle, any kind of measure can be used to create a linear descriptor by using a reference panel of compounds or target proteins. An interesting example of this technique is affinity fingerprints, which can be either experimental [43] or virtual [44]. Here, the affinity of a molecule to a reference set of proteins is measured or calculated, resulting in an affinity vector. If the binding behaviour of the molecules to the reference panel is similar, it can be expected that they also behave similar in binding to an unknown protein.

From virtual screening to virtual searching: exploring combinatorial chemistry spaces

Compound databases can have up to several millions of compounds, however, they reflect only a tiny portion of the universe of compounds that can be synthesized, in principle. One possible way of covering a broader range of compounds is by definition of molecules, based on molecular fragments and rules, describing how these fragments can be combined to result in a valid molecule. Owing to the combinatorial nature of such chemistry spaces, a few thousand fragments, in combination with a dozen linkage rules, are sufficient to create a virtual compound space that contains more than 10^{20} molecules – a trillion times more than that currently contained in the largest compound collections, such as Beilstein (Crossfire Beilstein Database, Beilstein Chemiedaten und Software GmbH and MDL Information Systems; http://www.beilstein.com). Although it is difficult to predict whether a compound can actually be synthesized, a careful selection of fragments and linkage rules provides chemistry spaces that are reasonably accessible via synthesis [45]. Several approaches, for example, focused or diverse combinatorial library design, or structure-based *de novo* design, explore such spaces.

To enable similarity searching in a chemistry space, completely new algorithmic approaches were required, because the traditional method of enumeration and search was not compatible. If, however, the problem is formulated as a combinatorial optimization problem, tackling it becomes feasible.

Schneider *et al.* [46] employ a genetic algorithm to construct molecules from a given chemistry space, with high similarity to a given query. Such a method is fast and can be used in combination with arbitrary similarity measures. However, it covers only a small portion of the search space and gives no guarantee of arriving at the global optimum. Another possibility involves calculating molecular properties at the fragment level and combining only those fragments that show sufficient partial similarity to the query. The ChemSpace approach [47] follows this idea. For each fragment, so-called 'topomeric' shape descriptors are generated that represent the shape of the fragment in a standardized conformation. By splitting a single bond or a pair of bonds, the query molecule is divided into two or three pieces. Each piece is compared independently with the fragments in the chemistry space. Well-fitting fragments are combined to result in a molecule with high shape similarity to the given query.

The Feature Tree descriptor (discussed previously) is also well suited to chemistry space similarity searches [48]. The descriptor represents building blocks of a molecule as the nodes of a tree, therefore, fragments can be converted to small Feature Trees and then combined to

form larger trees. A specially designed algorithm, FTrees-FS, based on a double-dynamic programming paradigm, exploits this feature and enables efficient searching in chemistry spaces. The algorithm not only allows for an arbitrary topology of connected fragments, it also performs a complete search with an optimality guarantee within minutes. Furthermore, the targeted level of similarity between query and resulting molecules is under the control of the user; setting this level to, say, 0.9, instead of its maximum value 1.0, enables effective scaffold hopping. Figure 3 shows an example of scaffold hopping between two structural classes of cyclooxygenase-2 (COX-2) inhibitors. A detailed analysis of similarity levels and several examples of scaffold hopping can be found in [48].

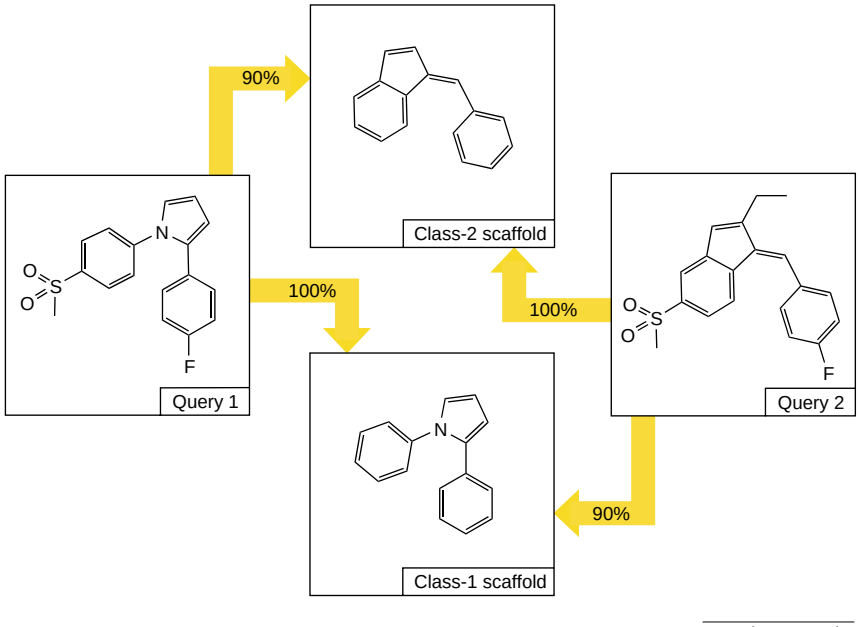

Figure 3. Scaffold hopping with fragment spaces: COX-2 inhibitors of class 1 (query 1) detect class 2 inhibitors and vice versa when the search is performed on 90% similarity level. At 100% similarity level, compounds of the same class are discovered.

Similarity-based analysis of high-volume screening data

Similarity-based methods are now widely accepted as a useful tool in the drug design process. Virtual screening (VS), an *in silico* procedure, and HTS, an experimental procedure, are regarded as complementary approaches in the lead identification process [49]. Owing to the rapid progress in the field of combinatorial chemistry, large sets of diverse structures are available. Therefore, the most frequently applied virtual screening methods rely on fast 2D descriptors. The identified screening hits are compared to each other to generate a hypothesis on the underlying lead structure. This can be accomplished by similarity-based methods.

Similarity-based methods are used for identifying structural classes around the detected screening hits, by fast clustering or partitioning algorithms. After such grouping of similar ligands, SAR models can be generated that relate biological activity to the presence or absence of substructures or functional groups. These models can be used to prioritize molecules for further testing.

Two main approaches to the computer-aided detection of active compounds are described in the literature: 'iterative screening' and 'one-shot screening' (Figure 4). In iterative (or sequential) screening, a medium-sized initial sample of, say, several thousand compounds is proposed for experimental testing. The measured activities are used to construct a SAR model. By applying the model to untested compounds, new molecules with a high probability of being active, can be proposed for further experimental testing. Iterative screening aims to reduce the amount of compounds that have to be tested. Machine learning approaches have been suggested that couple especially well with the sequential screening strategy [50].

To follow are a few examples of similarity-based methods that have been applied to extremely large datasets to find lead candidates.

CerBeruS [51,52] is a method developed for iterative screening. CerBeruS is based on Daylight fingerprints. First, all compounds are grouped by their structural similarity. An initial sample of compounds (e.g. cluster representatives) is selected for iterative screening. Then, in each round of screening, all members of 'active' clusters (i.e. clusters that contain at least one active molecule) are selected for retesting. CerBeruS proposes only highly similar molecules for testing. This strategy results in a high hit rate but is unlikely to identify new scaffolds or lead series. Therefore, each active structural class has to be present in the first random sample. As a point in case, in the example originally reported [51], CerBeruS would have missed more than half of the actives.

Whereas, in iterative screening experimental screens of moderate size are performed successively, in one-shot screening (Figure 4b), all accessible compounds are tested

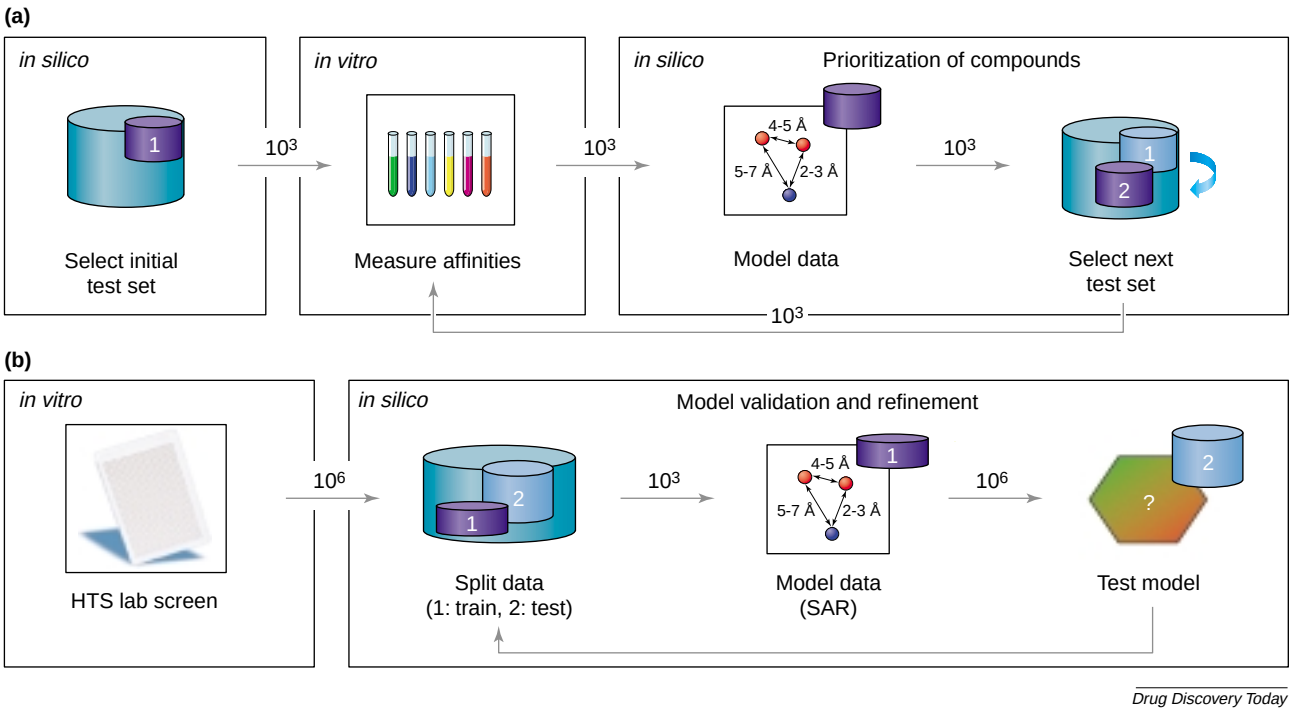

Figure 4. (a) Iterative screening: starting with a small set of compounds, in each experimental screening step, an SAR model is constructed, which is then used to select new compounds for testing. **(b)** One-shot HTS screening: in each virtual screening step, an SAR model is constructed from a training set, which is then validated on the test set.

experimentally at once, and the iterative construction of a model for an active compound is then carried out in the computer, using computational learning techniques. Here, the classical train-and-test paradigm is used.

The tools MCASE (Multi Computer Automated Structure Evaluation) [53] and PGLT [54] (Phylogenetic-Like Tree Algorithm) use recursive partitioning (RP) for the analysis of one-shot screening data (RP is a statistical method for the classification of large datasets). In each step, the statistically best variable is used to split the dataset into smaller and more homogenous subsets. MCASE is a QSAR expert system that allows for binary activity classification by correlation of substructure descriptors with biological activity. MCASE is one of the earliest methods. Its main restriction, which the field has since transcended, is that it only considers the binary classification as active or inactive. PGLT combines different data mining methods around the RP paradigm. RP methods are particularly well suited for problems where adequate information is available. In HTS data analysis, only the first splits have sufficient statistical support. Outliers, unbalanced and noisy data are not handled well by RP methods.

As an alternative to RP, neural networks can be applied to screening data. An example is 3D MIND (Mining Information for Novel Discoveries) [55], which is based on non-linear mapping. This method enables the grouping of compounds in a high-dimensional space by projecting them to 2D self-organizing maps (SOMs). 2D structural keys are used as descriptors. Approximately 20 000 compounds from the National Cancer Institute (NCI) tumor cell panel were grouped by two different SOMs: one for structural similarities and one for comparing biological readouts of different cell lines. The analysis showed that structurally similar compounds share the same cellular activity [55].

Predefined structural families can be used to group similar compounds of a HTS screen. An example is LeadScope™ [56], which uses a predefined database of hierarchically ordered substructures. The ligands of the test set are grouped into structurally homogeneous classes, according to this template database of ~27 000 substructures. For each subset, a so-called p-value is computed, describing the probability that the molecules show a given average activity. Because of the predefined hierarchy of structural classes, LeadScope™ has no computationally expensive training phase. Thus, it can be applied to extremely large screening sets, but will not find previously unknown structural motifs.

The only publicly available source of a large dataset with biological assay data is compiled by the NCI (Developmental Therapeutics Program, National Cancer Institute; http://dtp.nci.nih.gov/docs/3d_database/structural_information/structural_data.html). Most of the described methods have been validated on one of the NCI datasets: HIV or cancer. To some extent, previously described structural motifs could be identified, for example, a collection of nucleosides by PGLT and topoisomerase inhibitors by LeadScope™. A thorough comparison of the results and an example of a real test case is missing for all algorithms. However, for MCASE only, an experiment using ten compounds of unknown activity is reported. This method was able to predict five out of seven active molecules as active and identified all three inactives. There are further publications providing an analysis of the NCI HIV or cancer data [57–60].

VS and HTS techniques have been applied successfully in the drug design process but they are just beginning to be used in combination. Further progression in this direction seems to be promising. Specifically, VS can be used to analyze the growing number of noisy data points from HTS experiments, and HTS can be used to validate virtual screening results.

Perspectives

Owing to the involvement of ever-increasing numbers of compounds in today's drug design projects, there is a high demand for efficient computational screening tools. Clearly, progress has been made in the quality and speed of VS methods; however, there is still a much room for improvement.

Descriptors need to be developed that incorporate 3D features but are conformation-independent. *De novo* design or other methods that navigate through virtual compound spaces must begin to take synthesizability into account. Scoring functions need to be refined. The ability of relating biologically similar but chemically diverse compounds requires further improvement. These are merely a few areas of intensive current research.

An example is the extension of the fuzzy FTrees descriptor, called HTSview, for the analysis of HTS data, by means of multiple FTree alignments. By combining the information of remotely related actives and inactives into a single super tree, efficient database searches with molecule ensembles are possible. Multiple FTree models capture local and global properties of the compounds and give statistics with only a minor loss of quality compared to the much more labor- and compute-intensive 3D comparative molecular field analysis (COMFA) models [61].

A general observable trend is the increasing use of algorithmic technology in improving the quality and throughput of VS. Several goals might still be out of reach in the short- or medium- term: as computational screening procedures go beyond lead identification and incorporate secondary drug properties, such as ADME/Tox, we have to deal with several tens or hundreds of putative targets. Furthermore transport processes and side effects are hard to capture, statistically, and are difficult to model. Much effort must be invested in understanding the functional mechanisms of cells before ADME/Tox can go beyond quantitative structure–property relationship predictions.

References

1 Schneider, G. and Böhm, H.J. (2002) Virtual screening and fast automated docking methods. *Drug Discov. Today* 7, 64–70
2 Abagyan, R. and Totrov, M. (2001) High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* 5, 375–382
3 Bissantz, C. *et al.* (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* 43, 4759–4767
4 Taylor, R.D. *et al.* (2002) A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* 16, 151–166
5 Brooijmans, N. and Kuntz, I.D. (2003) Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* 32, 335–373
6 Lyne, P.D. (2002) Structure-based virtual screening: an overview. *Drug Discov. Today* 7, 1047–1055
7 Halperin, I. *et al.* (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* 47, 409–443
8 Good, A.C. *et al.* (2000) High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discov. Today* 5 (Suppl 1), 61–69
9 Patel, Y. *et al.* (2002) A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J. Comput. Aided Mol. Des.* 16, 653–681
10 Lemmen, C. *et al.* (1998) FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.* 41, 4502–4520
11 Rarey, M. *et al.* (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 261, 470–489
12 Lemmen, C. and Lengauer, T. (1999) Fragment-Based Screening of Ligand Databases. In *Molecular Modelling and Prediction of Bioactivity*, (Gundertofte, K. and Jorgensen, F.S., eds), p. 169–174, Plenum Press, New York
13 Lemmen, C. *et al.* (2000) Multiple molecular superpositioning as an effective tool for virtual database screening. *Perspect. Drug Des. Discovery* 20, 43–62
14 Jones, G. *et al.* (1995) A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput. Aided Mol. Des.* 9, 532–549
15 Wild, D.J. and Willett, P. (1996) Similarity searching in files of three-dimensional chemical structures. Alignment of molecular electrostatic potential fields with a genetic algorithm. *J. Chem. Inf. Comput. Sci.* 36, 159–167
16 Mestres, J. *et al.* (1997) A molecular field-based similarity approach to pharmacophoric pattern recognition. *J. Mol. Graph. Model.* 15, 114–121
17 Carbo, R. *et al.* (1980) How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* 17, 1185–1189
18 Good, A.C. *et al.* (1992) The utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* 32, 188–191
19 Goodford, P.J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 28, 849–857
20 Krämer, A. *et al.* (2003) Fast 3D molecular superposition and similarity search in databases of flexible molecules. *J. Comput. Aided Mol. Des.* 17, 13–18

21 Pitman, M.C. *et al.* (2001) FLASHFLOOD: a 3D field-based similarity search and alignment method for flexible molecules. *J. Comput. Aided Mol. Des.* 15, 587–612

22 Hahn, M. (1997) Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci.* 37, 80–86

23 Putta, S. *et al.* (2002) A novel shape-feature based approach to virtual library screening. *J. Chem. Inf. Comput. Sci.* 42, 1230–1240

24 Putta, S. *et al.* (2003) A novel subshape molecular descriptor. *J. Chem. Inf. Comput. Sci.* 43, 1623–1635

25 Bures, M.G. (1997) Recent techniques and applications in pharmacophore mapping. In *Practical Application of Computer-Aided Drug Design* (Charifson, P.S., ed.), pp. 39–72, Marcel Dekker, New York

26 Lemmen, C. and Lengauer, T. (2000) Computational methods for the structural alignment of molecules. *J. Comput. Aided Mol. Des.* 14, 215–232

27 Lipinski, C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26

28 Tong, W. *et al.* (1998) Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.* 38, 669–677

29 Mason, J.S. *et al.* (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* 42, 3251–3264

30 Willett, P. *et al.* (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 38, 983–996

31 Sheridan, R.P. *et al.* (1996) Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* 36, 128–136

32 Good, A.C. and Kuntz, I.D. (1995) Investigating the extension of pairwise distance pharmacophore measures to triplet-based descriptors. *J. Comput. Aided Mol. Des.* 9, 373–379

33 Matter, H. (1997) Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* 40, 1219–1229

34 Brown, R.D. and Martin, Y.C. (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* 36, 572–584

35 Brown, R.D. and Martin, Y.C. (1997) The information of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* 37, 1–9

36 Cruciani, G. *et al.* (2002) Suitability of molecular descriptors for database mining. A comparative analysis. *J. Med. Chem.* 45, 2685–2694

37 Rarey, M. and Dixon, J.S. (1998) Feature trees: a new molecular similarity measure based on tree matching. *J. Comput. Aided Mol. Des.* 12, 471–490

38 Gillet, V.J. *et al.* (1991) Computer storage and retrieval of generic chemical structures in patents. 13. Reduced graph generation. *J. Chem. Inf. Comput. Sci.* 31, 260–270

39 Gillet, V.J. *et al.* (2003) Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* 43, 338–345

40 Barker, E.J. *et al.* (2003) Further development of reduced graphs for identifying bioactive compounds. *J. Chem. Inf. Comput. Sci.* 43, 346–356

41 Matter, H. and Rarey, M. (1999) Design and diversity analysis of compound libraries for lead discovery. In *Combinatorial Organic Chemistry*, (Jung, G., ed.), pp. 409–439, Wiley-VCH

42 Cruciani, G. *et al.* (2000) VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* 11 (Suppl 2), S29–S39

43 Kauvar, L.M. *et al.* (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* 2, 107–118

44 Briem, H. and Lessel, U. (2000) In vitro and in silico affinity fingerprints: finding similarity beyond structural classes. *Perspect. Drug Discov. Des.* 20, 231–244

45 Lewell, X.Q. *et al.* (1998) RECAP–retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* 38, 511–522

46 Schneider, G. *et al.* (2000) De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided Mol. Des.* 14, 487–494

47 Andrews, K.M. and Cramer, R.D. (2000) Toward general methods of targeted library design: topomer shape similarity searching with diverse structures as queries. *J. Med. Chem.* 43, 1723–1740

48 Rarey, M. and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *J. Comput. Aided Mol. Des.* 15, 497–520

49 Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1, 882–894

50 Warmuth, M.K. *et al.* (2003) Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* 43, 667–673

51 Engels, M.F. *et al.* (2000) CerBeruS: a system supporting the sequential screening process. *J. Chem. Inf. Comput. Sci.* 40, 241–245

52 Engels, M.F. and Venkatarangan, P. (2001) Smart screening: approaches to efficient HTS. *Curr. Opin. Drug Discov. Dev.* 4, 275–283

53 Klopman, G. and Tu, M. (1999) Diversity analysis of 14 156 molecules tested by the National Cancer Institute for anti-HIV activity using the quantitative structure-activity relational expert system MCASE. *J. Med. Chem.* 42, 992–998

54 Nicolaou, C.A. *et al.* (2002) Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.* 42, 1069–1079

55 Rabow, A.A. *et al.* (2002) Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. *J. Med. Chem.* 45, 818–840

56 Roberts, G. *et al.* (2000) LeadScope: software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* 40, 1302–1314

57 Chen, X. and Reynolds, C.H. (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.* 42, 1407–1414

58 Shi, L.M. *et al.* (2000) Mining and visualizing large anticancer drug discovery databases. *J. Chem. Inf. Comput. Sci.* 40, 367–379

59 Tamura, S.Y. *et al.* (2002) Data analysis of high-throughput screening results: application of multidomain clustering to the NCI anti-HIV data set. *J. Med. Chem.* 45, 3082–3093

60 Voigt, J.H. *et al.* (2001) Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* 41, 702–712

61 Zimmermann, M. *et al.* (2003) Extracting knowledge from high-throughput screening data: towards the generation of biophore models. In *14th European Symposium on Quantitative Structure-Activity Relationships*, (Van de Waterbeemd, H. ed.), Blackwell Publishing