Virtual Screening: Are We There Yet?

Mehran Jalaie* and Veerabahu Shanmugasundaram

Computer Assisted Drug Discovery, Pfizer Global Research & Development, 2800 Plymouth Road, Ann Arbor, Michigan 48105, USA

Abstract: The cost of pharmaceutical development has increased dramatically in recent years, and many assorted approaches have been developed to decrease both the time and costs associated with bringing a drug to the market. Among these methods is the use of *in silico* screening of compound databases for potential new lead compounds, commonly referred to as virtual screening (VS). Virtual screening has become an integral part of the early discovery process in pharmaceutical development, readily observed by the large number of methodologies that have been published to date. Other reviews have been published detailing the various types of virtual screening methods in use**.** This work will review some of the virtual screening approaches and strategies that have been attempted to identify compounds to launch medicinal chemistry campaigns. Understanding trends and drivers in VS should help to set expectations about how and when VS could be used and what it can and cannot deliver and how it can be integrated in a successful screening campaign and used in a complementary fashion to HTS.

Key Words: Virtual screening, structure-based virtual screening, ligand-based virtual screening, similarity searching, hit-directed nearest neighbor searching, hit-lists, data-fusion, data-aggregation.

1. INTRODUCTION

Increasing pressures on the pharmaceutical industry necessitate the ever increasing need to reduce attrition in later stages of drug discovery, improve efficiencies in the discovery and development processes, lower the cost of discovering and developing medicines, increase the speed of drug discovery cycles and produce results in the market. This is especially true in early drug discovery stages where the search for the initial lead compound and subsequent lead optimization begins. The efforts involved here, just by the nature of the drug discovery process, can have important consequences for therapeutic area project teams, where once a compound is identified (or removed) from the purview of a therapeutic project team *it and other structurally-similar compounds* are effectively pursued (or removed) from further consideration. This can have serious impact downstream – they can either set a project team on the right course or completely derail them and send them on a wild goose chase.

At the lead identification stage, large pharmaceutical companies perform any of the following activities – (a) Perform high-throughput screening (HTS) of the entire corporate collection where many hundreds of thousands of compounds are tested in the search for new drug leads, (b) Perform "limited HTS", where a subset of the corporate collection (typically a dissimilarity subset or target-based compound plate sets) is evaluated, (c) A known lead compound is taken and modified through systematic medicinal chemistry to "scaffold hop" to find another lead, (d) Perform virtual screening – either target-based (structure-based VS) or compound-based (ligand-based VS) to "scaffold hop" to find another lead, or (e) Combinations of (a) through (d).

*Address correspondence to this author at the Computer Assisted Drug Discovery, Pfizer Global Research & Development, 2800 Plymouth Road, Ann Arbor, Michigan 48105, USA; E-mail: mehran.jalaie@pfizer.com

However, as a result of the pressure to produce more with less, the old mantra of "screen them all", [approach (a)], as the only effective screening paradigm is no longer viable. As it is impossible to make accurate predictions about the ligand-protein interactions in a high-throughput fashion, researchers across the industry are currently using a variety of VS approaches. Furthermore, a given procedure tends to work better on some targets than on others in ways that are difficult to predict *a priori*. The method of choice in many cases appears to be defined by the familiarity and experience of the computational chemist to a particular computational method, nature and size of the compound collections, knowledge of the target/biological system and the screening capacity. The sections below provide an overview of selected approaches that have been attempted successfully in VS campaigns.

2. STRUCTURE-BASED VIRTUAL SCREENING

Looking for the proverbial "needle in a haystack" is the fundamental issue in finding lead structures that launch successful drug discovery campaigns. This challenge is now routinely addressed using virtual screening methods [1-7]. These methods have shown to be effective at identifying promising chemical entities or leads by employing relevant scoring schemes to identify the most promising candidates. An ever-increasing need in the speed of computing hardware and software underlies a fundamental limitation in virtual screening: very large chemical and conformational search-spaces need to be explored. Nevertheless, taking advantage of available structural information limits these search spaces and allows the Structure-Based Virtual Screening (SBVS) to find its way to the mainstream virtual screening tools [8-10]. SBVS paradigm primarily relies on docking a small molecule to a protein target and quantifying the resulting interaction [11-14]. Promising compounds are then selected from a sorted list of predicted values. Post processing of the list of

1389-5575/06 $50.00+.00

high scoring candidates is also performed to identify structural similarities/dissimilarities responsible for the predicted activity. Availability of so many scoring functions highlights the most problematic area in SBVS: our incapability to reliably calculate the binding affinity of new chemical entities.

The idea of docking and scoring as a virtual screening tool has been around since the dawn of docking methods [11, 12, 15]. The docking challenges arise from the six degrees of translational and rotational freedom as well as the conformational degrees of freedom for each molecule in order to recognize another molecule. The simplest docking algorithms treat molecules as rigid bodies and only explore the translational and rotational space of all molecules. A further simplification made by most docking programs is keeping the receptor molecule rigid and only allowing the conformational space of the ligand molecule to be realized [16-19]. Many of the popular software such as CombiDOCK [20], DOCK [11, 15], LUDI [21, 22], and PRO_SELECT [23] use a Clique-search [24, 25] based approach where two rigid bodies are matched according to specific features in space. As an alternative to the Clique search approach, other software use the Partial Matching [26] or Pattern Recognition algorithms for matching features [27, 28].

However, drug-like molecules are usually floppy and flexible docking of ligands to rigid receptors is considered to be a practical approach in docking [28-30]. Available software range from simulation techniques (QXP [31], ICM [32, 33], PRODOCK [34]), genetic algorithms [28, 29, 35-38] (GOLD [39] and EPDOCK [38, 40]), incremental construction algorithm, (Hammerhead [13], DOCK [15] and FlexX [41, 42]), to conformation database methods (Flexibase/FLOG) [43, 44]. Other docking methods [45], which don't fall into the above categories such as PRO_LEADS [46, 47], LIGIN [48] and other hybrid approaches, have also been proposed by many researchers [49-53]. Several research groups have been using popular docking packages such as Dock, FlexX, GOLD, ConsDock [54] for virtual screening of small datasets in a flexible fashion [5, 12, 55, 56], but none of these approaches are capable of processing a few million compounds, a typical pharmaceutical database, in a reasonable time frame.

Fast flexible docking with PRO_LEAD has been a very successful approach in screening 1.1 million compounds for a single target [6]. However, as the number of pharmaceutically interesting targets [6, 14] and drug-like databases grow [52, 57], much faster docking methods are needed to respond to the avalanche of data. In other words, a balance between accuracy (searching the conformational space) and speed has to be evaluated and achieved in order to take advantage of SBVS. On one hand, docking software such as FRED (Fast Rigid Exhaustive Docking), a high throughput structure-based virtual screening tool for finding leads by using a set of high resolution ligand-protein complexes, are routinely used. The ultimate challenge to tools like FRED, as a lead hunter used in virtual screening, will be measured by its capability of enriching the final list of compounds in comparison to the random selections and its speed. On the other hand, work is ongoing to not only flex a potential ligand, but also account for the flexibility of the receptor site (FlexE [58], a modified version of docking program FlexX, to model receptor flexibility based on crystallographic structures and the Induced Fit within the Glide [59, 60] docking software where it models the conformation of the protein when it binds with the ligand). Treating explicitly the receptor flexibility is computationally expensive and could worsen the result [45], and hence would make this approach a less desirable method for high throughput virtual screening. Despite this recent methods such as fast structure-based screening [61], that utilizes docking and 2D QSAR for rapid scoring is reportedly thousand times faster than regular docking and still demonstrates a comparable performance [62-68].

Schapira *et al*. demonstrated that high-throughput docking (using ICM [33]) could be used to rapidly identify and prioritize lead thyroid hormone antagonists that differ from known ligands. Antagonist molecules of different chemotypes discovered in this work illustrates the power of receptor-based virtual screening and demonstrates that diverse chemical space can be narrowed using structural information. Further lead optimization of Thyroid Receptor (TR) antagonists enhanced the property of the discovered chemical entities [65]. In another investigation [69], ligands were docked into an averaged ensemble of crude homology models of the target protein. Improved homology models were generated using a distance-dependent pair potentials derived from PDB ligand-protein complexes. The modeled receptors were ranked and selected based on the evaluation of the interactions between the ligands and the generated pockets. Force Field minimizations were then applied to the final models. Ultimately, application of this approach to 46 protein–ligand complexes, taken from the Protein Data Bank (PDB), successfully produced near-native binding-site geometries.

Vangrevelinghe and co-workers [70] have identified a substituted indoloquinazoline compound as a novel inhibitor of protein kinase CK2 by virtual screening of a 400 000-compound library against a homology model of human CK2. In this method pharmacophoric knowledge along with multiple scoring (consensus scoring) was applied to enrich the final hit list. The IC_{50} of the best compound is reported to be of 80 nM. This report clearly shows that large-scale database docking procedures in combination with scoring and filtering processes can be useful lead identification procedures. In another example, the crystal structure of cathepsin D was used to select the building blocks for a combinatorial library synthesis. As a result, the library yielded potent inhibitors (K_i = 9-15 nM) of cathepsin D. The success of these studies clearly demonstrates the power of coupling combinatorial chemistry and structure-based design (docking) [71]. Furthermore, this also points to the advantage that VS approaches have over HTS, where evaluation of virtual compound libraries such as the combinatorial libraries in the above example that could be synthesized is possible.

Böhm *et al*. [21, 72, 73] described that random screening provided no suitable lead structures in a search for novel inhibitors of the bacterial enzyme DNA gyrase. However, relying on detailed 3D structural information of the targeted ATP binding site, an approach combining four key techniques (1) *in silico* screening for potential low molecular weight inhibitors, (2) a biased high throughput DNA gyrase screen, (3) validation of the screening hits by biophysical methods, and (4) a 3D guided optimization process, was at-

tempted. When *in silico* screening was performed, the initial data set containing 350 000 compounds was reduced to 3000 molecules. Testing these 3000 selected compounds in the DNA gyrase assay provided 150 hits (5-64 μg/mL) clustered in 14 classes. Seven classes (phenols, 2-amino-triazines, 4-amino-pyrimidines, 2-amino-pyrimidines, pyrrolopyrimidines, indazoles, and 2-hydroxymethyl-indoles) were validated as true, novel DNA gyrase inhibitors that act by binding to the ATP binding site located on subunit B. The 3D guided optimization provided highly potent DNA gyrase inhibitors, e.g., the 3,4-disubstituted indazole (0.03 μg/mL) being 10 times more potent DNA gyrase inhibitor than novobiocin.

Paiva *et al.* [74] compared VS and HTS of the Merck chemical collection, using FLOG [43, 44] docking software, against the tuberculosis target dihydrodipicolinate reductase. The HTS hit-rate was outperformed by 30 fold by VS. The top hits from docking and HTS had K_i values of 7.2 and 35 uM, respectively. Having a rigid receptor site and a rigid bioactive molecule improves the accuracy of docking methods, and also enhances the chance of a successful virtual screening effort. Furthermore, enzymes that are inhibited with rigid ligands are great candidates for fast rigid docking methods [75].

ICM has been used to identify new antagonists of human retinoic acid receptor-alpha even when the relevant crystal structure was not available. New antagonists were discovered with ICM by virtual screening the Available Chemicals Directory (ACD) to a model of retinoic acid receptor-alpha structure [76]. In another scenario, Internal Coordinates Mechanism (ICM) flexible docking procedure was successful in identifying potential binders of the RNA hairpin HIV-1 TAR RNA when a subset of the ACD was screened [77].

DOCK has also been used with various databases to identify potential inhibitors of kinesin and thymidilate synthase [78, 79]. In another investigation, Aronov *et al.* screened a small virtual library and discovered inhibitors of the hypoxanthine-guanine-xanthine phosphoribosyl transferase [80]. Others [81, 82] have also been successful in finding novel inhibitors of FXa and thrombine respectively using docking and scoring methods.

Moreover, in another instance, Burkhard *et al.* [83] employed the molecular docking computer program SANDOCK [84] (the X-ray structure of uncomplexed FKBP, a member of the immunophilin family, was used to provide a template with the binding pocket) to identify more than 20 low micromolar novel FKBP inhibitors. In another investigation, EUDOCK was used to virtual screen a chemical database (ACD) to identify inhibitors (4 out of 21 inhibitors had IC_{50} of 100uM or better) of farnesyltransferase (FT) with zinc present in the active site [85].

3. LIGAND-BASED VIRTUAL SCREENING

Crystal structures of potential drug targets are blueprints for the binding site that needs to be occupied by a ligand. Consequently, with structure-based virtual screening (SBVS) it is easier to rationalize, perk interest and gather buy-in from project teams to launch and execute a SBVS campaign. Many pharmaceutical companies invested heavily in structure-based tools in the 80s and with the advent of HTS and combinatorial chemistry in the 90s, it appeared that ligand-based virtual screening (LBVS) had become passé as the industry focused upon SBVS methods. Now with automation making inroads into macromolecular crystallography in the 00s, thereby enabling parallel crystallization trials to solve crystal structures of difficult protein targets, protein crystallography is quickly establishing itself as a high-throughput technique in its own right and bringing SBVS more frequently to the forefront in early-stage drug discovery efforts. However, a majority of the therapeutic targets still remain difficult to crystallize. There is an immense landscape of proteins that remains to be solved and drugged. One of the biggest gaps is with integral membrane proteins and one such class of membrane proteins, G-protein coupled receptors (GPCRs), are targets for more than 50% of the current drugs on the market and make up the majority of validated targets in biomedical research [86]. Furthermore, using structure-based methods to understand functional activity such as agonism, partial agonism, antagonism and inverse agonism is not relatively straight forward. Sometimes effective binders still lack functional activity and understanding this, using structure-based methods remains a challenging problem. Hence compound screening based on ligand-based methods continues to remain indispensable for a bulk of targets and the primary approach for drug discovery projects over the past several decades. The selection of compounds that are screened depends upon the type, throughput and capacity of the assay. The libraries that are selected to screen are based on a variety of approaches [87] such as HTS of the entire corporate collection, subset screening based on random selection, dissimilarity based selection, filtered collections based on “drug like properties”, target-based plate sets such as kinase libraries, and finally “cherry-picking compounds” based on ligand-based virtual screening approaches.

A well-known principle that is often used in searching for active compounds is that “similar compounds have similar activities”. While not uniformly true, due to the underlying differences in the nature of the activity landscapes, it still holds in enough cases that similarity based searching has become a well-accepted way of finding additional active compounds based on a known lead, hit, or series of hits. However, similarity can be measured in a variety of ways. And like beauty, similarity is in the eyes of the beholder. Ideally one would like to measure similarity through the "eyes of the receptor". Here examples are used to illustrate some of the widely used LBVS methods and processes that define a successful LBVS campaign. These examples are by no means exhaustive and represent only a subset of LBVS techniques used and are discussed within the context of the current literature and some recent trends and developments.

Methods & Applications

There are a variety of LBVS methods that have been developed and applied to select compounds from a database for screening in a particular assay. These methods typically vary in the molecular representation that is used to describe compounds, but essentially perform comparative molecular similarity analysis of compounds. The methods also vary in the use of either a known active compound, and/or a set of active compounds and/or a set of active *and* inactive compounds, to

select compounds from the database and the similarity indices used.

One of the most chemically intuitive and widely used LBVS technique is similarity searching based on the 2D chemical structure of an active compound [88]. Substructure searching based on the active compound is also a form of similarity searching. Typically with similarity searching, databases of compounds are represented as bit string representations or molecular-fingerprints. Bit positions in the fingerprint representation usually encode for the presence or absence of a structural fragment. In similarity analysis, the overlap between bit patterns in corresponding fingerprints is quantified using various similarity coefficients such as the Tanimoto coefficient.

Using molecular fingerprints and similarity values, it was suggested that 85% of compounds that have a Tanimoto coefficient value of 0.85 or greater relative to an active compound should also be active [89]. However, using different data sets, it was recently reported that at a similarity value of 0.85 or greater only about 30% of compounds identified were also active [90]. Thus the relevance of the similarity threshold is influenced by the method used and the dataset under investigation. Nevertheless, similarity searching is usually the first method of choice for any LBVS campaign.

Several other molecular representations, in addition to fingerprint representations, have been used to perform similarity searching. These include topological descriptors, descriptors based on surface-dependent properties, descriptors based on molecular properties, etc. Combinations of these descriptors and structural fragments have also been successfully used in LBVS. For example, a mini-fingerprint (MFP) composed of three numerically encoded 2D descriptors and about 32-40 structural keys was used to identify compounds that contained similar activities of some drug-like molecules for endothelin A antagonists, α1-adrenergic receptor ligands, some serine protease inhibitors and fibrinogen receptor antagonists [91, 92]. In general, MFPs displayed the best overall performance at a Tanimoto cutoff value between 0.65 and 0.7. Within this "similarity interval", MFPs have, on average, an approximately 60% chance of correctly identifying all molecules with biological activity similar to a query compound and recognize only 1-2% false positives. For virtual screening calculations, the authors used a threshold value of 0.6 as an MFP-specific Tanimoto similarity criterion.

LBVS techniques that involve the transformation of continuous descriptors into a binary format based on statistical medians and subsequent definition of a simplified chemistry space have also been used successfully [93, 94]. Identification of consensus positions of specific compound sets in these spaces, and iterative adjustments of the dimensionality of the descriptor spaces is then performed in order to discriminate compounds that share similar activity from others.

Indices such as the molecular equivalence indices [95] classify a molecule with respect to a class of structural features or topological shapes such as its cyclic system or a set of functional groups. These indices have been shown to identify interesting features [96] based on the topological shape of a molecule and its set of functional groups that are strongly linked with activity. A data-shaving study [97] has exploited the presence and absence of a subset of these indices in both active and inactive compounds to "shave" off or deprioritize compounds similar to inactives from LBVS. Similarity searching was shown to improve when compounds predicted to be inactive were deprioritized.

Reduced molecular graph representations [98, 99] have been used for similarity searching and identifying active compounds. Typically in graphs, atoms are represented as nodes and bonds as edges. In reduced graphs, nodes can represent, for example ring systems connected by functional groups. Thus reduced graphs condense molecular representations by emphasizing features. Similarity searching with reduced graphs [100] performed quite comparably to Daylight fingerprints but identified more structurally diverse molecules (broader coverage) having similar activity.

Cluster analysis has been applied for a variety of cheminformatic problems. Typically Jarvis-Patrick clustering has been the method of choice for non-hierarchical clustering and Ward's clustering for agglormerative-hierarchical clustering. Recently, reports where hierarchical clustering has been used as a LBVS tool [101] have been reported. For example, when compounds were selected for testing based on nearest-neighbor analysis using hierarchical clustering methods, average hits rates of 15% or more were observed that were significantly greater than the primary hit rate.

Cell-based partitioning in low-dimensional chemistry spaces using BCUT descriptors [102, 103] and hit-directed nearest neighbor searching in these chemistry spaces has also been widely used as a LBVS technique. BCUTs are molecular descriptors that capture features that are critical for protein-ligand interactions and compounds that have similar BCUT values are hence likely to be similar in activity. Several examples of use of BCUT based searching in a retrospective and prospective fashion have been reported in the literature [104-107].

Statistical partitioning methods such as recursive partitioning (RP), a decision tree based method, have also been used as a LBVS technique. The approach is similar to a divisive-hierarchical clustering but divides data along the decision tree, typically using a two-state (yes/no – present/absent) decision based on molecular descriptors. In a study with monoamine oxidase inhibitors [108], RP models yielded a 15-fold enrichment in hit-rate over random selection and with HTS data sets, up to 10-fold enrichments [109, 110].

Many methods exist for performing LBVS based on 3-D molecular similarity. Lemmen and Lengauer [111] provide a comprehensive review of most of the methods in use today, a large class of which utilizes some form of vector-based representation of 3-D molecular features such as 3-D pharmacophores [112] and various types of 3-D shape descriptors [113]. The components of these vectors can be binary, integer, categorical, or continuous. Most 3-D methods, however, involve some type of direct alignment of the molecules being considered.

LBVS using pharmacophoric queries has also been used extensively. Pharmacophores are derived from conformational explorations of a single active or sets of active and/or

inactive compounds. Based on the complexity of the pharmacophore generated, databases of compound collections are searched to identify virtual hit lists. These methods vary in terms of increasing complexity, information content and success rates [112, 114].

The bulk of the 3-D methods utilize some form of field-based function to represent the fields or pseudo-fields, which can be either continuous or discrete, surrounding molecules. Examples include "steric," electrostatic potential, and lipophilic fields [115]. Several workers have also developed a field-based methodology for directly aligning molecules based upon their electric fields [116, 117], which differs from the usual scalar potential fields that are typically matched, but these approaches have only been implemented as discrete procedures.

Simplified 3D representations of multiple conformations of ligands using clustering techniques that are used in data-mining have been reported recently. For example, Jenkins *et al*. [118] reported that similarity searching with feature point pharmacophores enriches actives taken from HTS datasets as well as those obtained from MDL Drug Data Report (MDDR) activity classes such as COX2 & HIV-RT inhibitors, 5-HT3A & D2 ligands and retinoids. Significantly, this method finds novel scaffold classes compared to 2D (Daylight, MACCS & PipelinePilot molecular fingerprints [119]) & 3D similarity (pharmacophore triplets) methods in the datasets studied.

Topomer similarity searching [120, 121] has been used as a predictor of similarity in biological activity. A topomer is an invariant 3D representation of a molecular fragment [122], derived from its 2D topology by rules that produce absolute coordinates for its constituent atoms. In a prospective study [123], with 308 compounds in 13 assays the LBVS hit rate averaged over all assays was 39%, significantly greater than the control hit rate of 15%.

Shape-based LBVS using ROCS [119, 124] (Rapid Overlay of Chemical Structures) to find new scaffolds for small molecule inhibitors of the ZipA-FtsZ protein-protein interaction, was reported recently [125]. The shape comparisons were made relative to the crystallographically determined bioactive conformation of a HTS hit. The use of ROCS led to the identification of a set of novel, weakly binding inhibitors that were missed using similarity-based approaches such as ISIS 2D fingerprints.

Molecular shape information has also been described in bit string formats for similarity searching based on 3D conformations of test compounds [126]. In these cases, compounds were represented as bit strings [127] capturing a set of shapes and sometimes pharmacophore-like features such as donors, acceptors, aromatic rings, etc. Similarities to knowns actives and inactives were then used to search databases and score candidate compounds.

Recently there has been significant interest in the application of support vector machines for LBVS [128-130]. Typically SVMs have been used as a classification and activity prediction tool. SVMs project compound collections into a space where molecules are represented as vectors and a hyperplane is then constructed based on a linear combination of vectors to differentiate compound sets. Using this approach substrates and non-substrates for different isoforms of UDP-glucuronosyltransferase were classified in a study [131] and was found to be superior (greater than 80% prediction accuracy) to PLS discriminant analysis and Bayesian neural networks.

Self-organizing maps and neural nets have also been used as classifiers and LBVS tools. For example, SOM clustering [132] was used to classify compounds based on screening data of the National Cancer Institute and in combinations with 2D fingerprints neural nets have also been used to virtual screen compound collections to nearly 90% accuracy in identifying CNS active compounds and mining for CYP-3A4 inhibitors [133-135].

4. PRE-FILTERING CORPORATE COLLECTIONS

Pre-filtering corporate collections before a VS run is performed on a regular basis for a variety of reasons. Pre-filtering eliminates compounds that have undesirable properties such as those that contain reactive or toxic functional groups [52], those that do not have drug like character [136], have limited aqueous solubility [137], are likely to have unfavorable ADME properties like stability, permeability, absorption, CNS penetration, etc., and those that will not be considered as initial starting points for a therapeutic project by medicinal chemists because of synthetic issues [138] or unfavorable IP position, even if they were active in the biological assay. Pre-filtering also helps enrich collections with desired compounds and hence improves hit-rates. The complexity of the filtering functions ranges from simple rule-based functions to neural network algorithms or combinations thereof [139, 140].

5. VIRTUAL HIT-LISTS: DATA-FUSION *VS* DATA-AGGREGATION

In SBVS, when shortcomings of scoring functions was realized researchers started to combine scoring functions ranked the final docked list of compounds [141]. The rational behind the consensus scoring was to perform better than the worst individual scoring function in the chosen subset of the scoring functions. For example in a case study using angiogenin (a potent inducer of angiogenesis) [142], the accuracy of the HTS result was improved by virtual screening of the corresponding chemical libraries and selecting hits by HTS/VS consensus. In conjunction with HTS of the National Cancer Institute Diversity Set and ChemBridge DIVERSet E (~18,000 compounds total), VS was performed with two flexible library docking/scoring methods, DockVision/Ludi and GOLD. Analysis of the results revealed that dramatic enrichment of the HTS hit rate can be achieved by selecting compounds in consensus with one or both VS functions. For example, HTS hits ranked as top 2% by GOLD included 42% of the true hits, but only 8% of the false positives; this represents a six fold enrichment over the HTS hit rate. Notably, the HTS/VS method was effective in selecting out inhibitors with micromolar dissociation constants typical of leads commonly obtained in primary screens [142].

Similarly, one can combine several diverse and alike approaches in selecting the final list [143, 144]. The concept of data fusion is concerned about the "how" and "what" meth-

ods should be combined to provide the highest enrichment factor in a given system. Could docking methods be adding extra values to more traditional ligand-based approaches? Should these methods be combined sequentially or in parallel?

Since it is difficult to establish which method may perform best given a particular instance, data fusion methods, also referred to as consensus scoring, have been recently employed for merging virtual screening results from different similarity methods. Successful data fusion approaches could reduce the uncertainty involved in selecting the appropriate VS method. Given a set of plausible VS results from various methods for a particular query, it is unlikely that a user will know *a priori* which method's results are most appropriate. In addition, useful data fusion approaches could potentially result in a merged result that is superior to any of the individual input method results and thus is able to extract useful information from all input lists, including inferior methods. Among the published fusion methods, the Sum-rank method has been shown to be among the most successful [145-149].

In LBVS, it is well-known that chemistry spaces are representation dependent. As a result, relationships among compounds in one chemistry space are not necessarily preserved in another chemistry space. Thus, an intrinsic chemistry space does not exist [150], and this has important consequences with regard to the distribution of compounds in these spaces. For example, it is entirely possible that clusters of compounds in one chemistry space may become uniformly spread out in another chemistry space and vice-versa. Therefore, Nearest Neighbor (NN) relationships may not be the same in the different spaces. These confounding factors, namely that different chemistry-space representations lead to different distributions of compounds and that significant violations of the similarity principle occur, have led to the realization that the quest for the best computational technique in NN searching of compound databases may be a futile exercise. Several researchers in the field have hence developed and applied a variety of novel computational tools to mitigate some of these representation-dependent and similarity or distance-biased views of chemistry space [104, 150, 151].

Holliday *et al*. [148] studied different types of similarity coefficients and data fusion methods to combine and optimize similarity measures between molecules. Data fusion methods such as Sum-rank, best-in-n fusion, best single etc. were used to combine and compare several similarity coefficients to obtain an overall estimate, which was later used for similarity searching. Results indicate that combinations such as Sum-rank fusion method improved hit-rates but no single combination gave a consistently high performance.

Raymond *et al*. [152] developed a new consensus scoring approach for merging the results of different virtual screening methods based on conditional probabilities. This technique was experimentally evaluated using several ligand-based virtual screening methods and compared to two variations of the Sum-rank fusion method where the conditional probability (CP) method performed as well or better than the Sum-rank methods. The individual VS methods chosen for the study were Daylight and BCI fingerprints, 2D MCS and 3D MCS, Tripos's dbtop and Open Eye's ROCS. These results indicate that consensus scoring enriches the hit-rate or the number of active compounds retrieved with respect to the best individual methods on average.

Post-processing virtual screening hit-lists with data aggregation methods have also been reported in the literature. In a seminal study, Sheridan and Kearsely [151] presented results from a retrospective analysis of various historical virtual screening studies with in-house methods on several different therapeutic targets. They concluded that the effectiveness of any similarity method varies greatly from one biological assay to another in a way that is difficult to predict *a priori*. Also, any two methods tend to select different subsets of actives from a database, so it is better to use several search methods where possible.

Shanmugasundaram *et al*. [104] recently described a data-aggregation strategy used in a prospective fashion at Pharmacia for identifying compounds for follow-up screening based on several ligand-based virtual screening hit-lists. This approach took explicitly into account different representations of chemistry space and identified compounds for follow-up screening that are likely to provide the best overall coverage of the chemistry spaces considered. The representations included 3-D, 2-D, 2-D topological BCUTs (2-DT) and molecular fingerprints derived from substructural fragments. The LBVS hit-lists that were obtained had little overlap. Moreover, in all of the four chemistry space representations, a minimum of 3- to 4-fold enrichment in actives over random screening was observed. The set of assays examined in this work covered a range of therapeutic-area projects, from CNS to anti-fungal to antibacterial and contained both cell-based and target-based assays, some of which were functional assays and some of which were binding assays.

6. ANALYSIS OF VS RESULTS – ENRICHMENT *VS* COVERAGE

Typically in a virtual screening campaign, after the selection of the virtual hit-lists and post-processing of these virtual hit-lists, a selection is made to finalize the list of compounds that are actually ordered, plated and tested in the biological assay of interest. Activity values of these compounds in the biological screen are then reported back. Two different measures, enrichment and coverage, are widely used to assess the virtual screening results.

Enrichment can be defined as the ratio of the proportion of actives in the VS campaign to the proportion of actives obtained in a primary screen or random selection of compounds. This gives a measure of how much better the VS campaign performed when compared to a normal primary screen or just a random selection of compounds.

Coverage is defined as the ratio of the number of actives in the VS to the number of actives in the entire collection, expressed as a percent. This gives a crude measure of how many different active compounds (or series) were identified or missed.

Typically it has been noted that enrichment and coverage behave in an approximately complementary fashion. As the similarity threshold increases, enrichment increases but cov-

erage decreases and vice versa. Use of these measures and the appropriate selection of the similarity threshold is where the "art of VS campaign" lies.

SUMMARY

In several studies, different VS procedures have yielded different subsets of active compounds for the same biological target. Furthermore, a given procedure tends to work better on some targets than on others in ways that are difficult to predict *a priori*. Thus, based upon a substantial amount of studies in the literature and from the arguments advanced earlier concerning the lack of invariance of different representations and their associated chemistry spaces, it does not appear that any single approach to VS can unequivocally identify compounds that are similar to active compounds obtained in screening studies. Thus, several researchers have taken different approaches to combine the virtual hit-lists obtained from different VS methods.

Using combinations of VS methods provides a practical approach to balance enrichment of hit-rates and coverage of chemistry space. Although it might be extremely satisfying to see high enhancements in hit-rates compared to primary screening rates (such as 100-fold or 200-fold), the ultimate success of any VS approach is the ability to cover different chemical classes of molecules and the ability to provide options to therapeutic area project teams. There is always a nagging fear (the sword of Damocles) hanging over any project team's head that the next blockbuster drug was actually a close analog of a compound in the corporate collection, but was a borderline hit which the project team had overlooked. The ability to combine multiple VS approaches, lowers that risk of not finding the next Lipitor or a close analog of it in your screen.

REFERENCES

[1] Lewis, R. A.; Pickett, S. D.; Clark, D. E. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B. Ed.; Wiley-VCH: New York, **2000**, Vol. *16*, pp. 1-51.

[2] Walters, W. P.; Ajay; Mucko, M. A. *Curr. Opin. Chem. Biol.,* **1999**, *3*, 384.

[3] Podlogar, B. L.; Muegge, I.; Brice, L. J. *Curr. Opin. Drug Discov. Devel.,* **2001**, *4*, 102.

[4] Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.,* **1995**, *38*, 1431.

[5] Stahl, M.; Rarey, M.; Klebe, G. In *Bioinformatics*; Lengauer, T. Ed.; Wiley-VCH: New York, **2001**, Vol. *2*, pp. 137-170.

[6] Waszkowycz, B.; Perkins, T. D. J.; Sykes, R. A.; Li, J. *IBM Syst. J.,* **2001**, *40*, 360.

[7] Lahana, R. *Drug Discov. Today,* **1999**, *5*, 43.

[8] Greer, J. E.; J. W.; Baldwin, J. J.; Varney, M.D. *J. Med. Chem.,* **1994**, *37*, 1035.

[9] Bohacek, R. S.; McMartin, C.; Guida, W.C. *Med. Res. Rev.,* **1996**, *16*, 3.

[10] Shoichet, B. K.; Bussiere, D.E. *Curr. Opin. Drug Discov. Devel.,* **2000**, *3*, 408.

[11] Kuntz, I. D. *Science,* **1992**, *257*, 1078.

[12] Shoichet, B. K.; Stroud, R. M.; Santi, D.V.; Kuntz, I.D.; Perry, K. M. *Science,* **1993**, *259*, 1445.

[13] Welch. W.; Ruppert. J.; Jain, A. N. *Chem. Biol.,* **1996**, *3*, 449.

[14] Gruneberg, S.; Stubbs, M. T.; Kelebe, G. *J. Med. Chem.,* **2002**, *45*, 3588.

[15] Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.,* **1982**, *161*, 269-288.

[16] Muegge, I.; Rarey, M. In *Reviews in Computaional Chemistry*, Lipkowitz, K. B.; Boyd, D. B. Ed. Wiley-VCH: New York, **2001**, Vol. *17*, pp. 1-60.

[17] Wermuth, C. G. *The Practice of Medicinal Chemistry*, Academic Press: San Diego, **1999**.

[18] Grootenhuis, P. D.; Kollman, P. A.; Seibel, G. L.; DesJarlais, R.L.; Kuntz, I. D. *Drug Des.,* **1990**, *5*, 237.

[19] Allen, F. H.; Bellard,S.; Brice, M.D.; Cartwright, B.A.; Doubledy, H.; Higgs, H.; Hummelink-Peters, T.; Kennard, O.; Motherwell, W.D.S.; Rodger, J.R.; Watson, D.G. *Acta Crystallogr.,* **1979**, *B35*, 2331.

[20] Sun, Y.; Ewing, T. J. A.; Skillman, A. G.; Kuntz, I. D. CombiDOCK: *J. Comput. Aided Mol. Des.,* **1998**, *12*, 597.

[21] Böhm, H.-J. *J. Comput. Aided Mol. Des.,* **1992**, *6*, 593.

[22] Böhm, J. H. *J. Comput. Aided Mol. Des.,* **1992**, *6*, 61.

[23] Murray, C. W.; Clark, D. E.; Auton, T.R.; Firth, M. A.; Li, J.; Sykes, R. A.; Waszkowycz, B.; Westhead, D. R.; Young, S. *J. Comput. Aided Mol. Des.,* **1997**, *11*, 193.

[24] Kuhl, F. S.; Crippen, G.M.; Friesen, D. K. *J. Comput. Chem.,* **1984**, *5*, 24.

[25] Bron, C.; Kerbosch, J. Finding All Cliques of an Undirected Graph. *Communi. Associ. Comput. Machin.,* **1973**, *16*, 575.

[26] Fischer, D.; Norel, R.; Wolfson, H.; Nussinov, R. *Struct. Funct. Genet.,* **1993**, *16*, 278.

[27] Fischer, D.; Lin, L. S.; Wolfson, H.; Nussinov, R. A. *J. Mol. Biol.,* **1995**, *248*, 459.

[28] Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. *J. Comput. Aided Mol. Des.,* **1995**, *9*, 113.

[29] Clark, K. P.; Jay, *J. Comput. Chem.,* **1995**, *16*, 1210.

[30] Leach, A. R. *Molecular Modelling :Principles and Applications*, Addison Wesley Longman Limited: Essex England, **1996**.

[31] McMartin, C.; Bohaceck, R. S. *J. Comput. Aided Mol. Des.,***1997**, *11*, 333.

[32] Abagyan, R.; Tortov, M.; Kuznetsov, D. *J. Comput. Chem.,* **1994**, *15*, 488.

[33] Totrov, M.; Abagyan, R. *Proteins,* **1997**, *1*, 215.

[34] Trosset, J. Y.; Scheraga, H. A. *J. Comput. Chem.,* **1999**, *20*, 412.

[35] Judson, R. In *Reviews in Computational Chemistry*, Lipkowitz, K. B.; Boyd, D. B. Ed. Wiley-VCH: New York, **1997**, Vol. *10*, pp. 1-73.

[36] Jones, G.; Willett, P.; Glen, R.C. *J. Comput. Aided Mol. Des.,* **1995**, *9*, 532.

[37] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.,* **1998**, *19*, 1639.

[38] Gehlhaar, D. K.; Verkhivker, P.A.; Rejto, D.B.; Fogel, L.J.; Feer, S.T. In *Docking Conformationally Flexible Small Molecules into a Protein Binding Site Through Evolutionary Programming. Proceeding of the Fourth Annual Conference on Evolutionary Programming*; McDonnell, J. R.; Reynolds, R.G.; Fogel, D. B. Ed.; MIT Press, Cambridge, MA.; **1995**, pp. 615-627.

[39] Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.,* **1997**, *267*, 727.

[40] Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. *Chem. Biol.,* **1995**, *2*, 317.

[41] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.,* **1996**, *261*, 470.

[42] Rarey, M.; Kramer, B.; Lengauer, T. *J. Comput. Aided Mol. Des.,* **1997**, *11*, 369.

[43] Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. *J. Comput. Aided Mol. Des.,* **1994**, *8*, 153.

[44] Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. *J. Comput. Aided Mol. Des.,* **1994**, *8*, 565.

[45] Abagyan, R.; Totrov, M. *Curr. Opin. Chem. Biol.,* **2001**, *5*, 375.

[46] Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. *Struct. Funct. Genet.,* **1998**, *33*, 367.

[47] Baxter, C. A.; Murray, C. W.; Waszkowycz, B.; Li, J.; Sykes, R. A.; Bone, R. G. A.; Perkins, T. D. J.; Wylie, W. *J. Chem. Inf. Comput. Sci.,* **2000**, *40*, 254.

[48] Sobolev, V.; Wade, R.C.; Vriend, G.; Edelman, M. *Proteins Struct. Funct. Genet.,* **1996**, *25*, 20.

[49] Wang, J.; Kollman, P. A.; Kuntz, I. D. *Proteins Struct. Funct. Genet.,* **1999**, *36*, 1.

[50] Hoffmann, D.; Kramer, B.; Washio, T.; Steinmetzer, T.; Rarey, M.; Lengauer, T. *J. Med. Chem.,* **1999**, *42*, 4422.

[51] Broughton, H. B. *J. Mol. Grap. Model.,* **2000**, *18*, 247.

[52] Walters, W. P.; Stahl, M. T.; Murcko, M. A. *Drug Discov. Today,* **1998**, *3*, 160.

[53] David, L.; Luo, R.; Gilson, M. K. *J. Comput. Aided Mol. Des.,* **2001**, *15*, 157.
[54] Paul, N.; Rognan, D. ConsDock: *Proteins Struct. Funct. Genet.,* **2002**, *47*, 521.
[55] Kramer, B.; Rarey, M.; Lengauer, T. *Proteins Struct. Funct. Genet.,* **1999**, *37*, 228.
[56] Stahl, M.; Rarey, M. *J. Med. Chem.,* **2001**, *44*, 1035.
[57] Martin, Y. C. *Perspec. Drug Discov. Des.,* **1997**, *7/8*, 159.
[58] Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. *J. Mol. Biol.,* **2001**, *308*, 377.
[59] Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.,* **2004**, *47*, 1739.
[60] Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. *J. Med. Chem.,* **2004**, *47*, 1750.
[61] Yoon, S.; Smellie, A.; Hartsough, D.; Filikov A. *J. Comput. Aided Mol. Des.,* **2005**, *19*, 483.
[62] Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discov.,* **2004**, *3*, 935.
[63] Rizzo, R. C.; Wang, D.-P.; Tirado-Rives, J.; Jorgensen, W. L. *J. Am. Chem. Soc.,* **2000**, *122*, 12898.
[64] Rosenfeld, R. J.; Goodsell, D.S.; Musah, R.A.; Morris, M. G.; Goodin, D.B.; Olson, A.J. *J. Comput. Aided Mol. Des.,* **2003**, *17*, 525-36.
[65] Schapira, M.; Raaka, B. M.; Das, S.; Fan, L.; Totrov, M.; Zhou, Z.; Wilson, S. R.; Abagyan, R.; Samuels, H. H. *PNAS,* **2003**, *100*, 7354.
[66] Schapira, M.; Raaka, B. M.; Samuels, H. H.; Abagyan, R. In silico discovery of novel Retinoic Acid Receptor agonist structures. *BMC Struc. Biol.,* **2001**, 1, www.biomedcentral.com/1472-6807/1/1
[67] Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. *Curr. Opin. Chem. Biol.,* **2002**, *6*, 439.
[68] Schneidman-Duhovny, D.; Nussinov, R.; Wolfson, H. J. *Curr. Med. Chem.,* **2004**, *11*, 91.
[69] Evers, A.; Gohlke, H.; Klebe, G. *J. Mol. Biol.,* **2003**, *334*, 327.
[70] Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. *J. Med. Chem.,* **2003**, *46*, 2656.
[71] Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J.; Sun, Y.; Kuntz, I. D.; Ellman, J. A. *Chem. Biol.,* **1997**, *4*, 297.
[72] Böhm, H.-J.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller, N.; Mueller, F. *J. Med. Chem.,* **2000**, *43*, 2664.
[73] Böhm, H.-J. *J. Comput. Aided Mol. Des.,* **1992**, *6*, 593.
[74] Paiva, A. M.; Vanderwall, D. E.; Blanchard, J. S.; Kozarich, J. W.; Williamson, J. M.; Kelly, T. M. *Biochim. Biophys. Acta Prot. Struct. Mol. Enz.,* **2001**, *1545*, 67.
[75] Schulz-Gasch, T.; Stahl, M. *J. Mol. Model.,* **2003**, *9*, 47.
[76] Schapira, M.; Raaka, B. M.; Samuels, H. H.; Abagyan, R. *PNAS,* **2000**, *97*, 1008.
[77] Filikov, A. V.; Mohan, V.; Vickers, T. A.; Griffey, R. H.; Cook, P. D.; Abagyan, R. A.; James, T. L. *J. Comput. Aided Mol. Des.,* **2000**, *14*, 593.
[78] Tondi, D.; Slomczynska, U.; Costi, M. P.; Watterson, D. M.; Ghelli, S.; Shoichet, B. K.*Chem. Biol.,* **1999**, *6*, 319.
[79] Aronov, A. M.; Munagala, N. R.; Kuntz, I. D.; Wang, C. C. *Antimicrob. Agents Chemother.,* **2001**, *45*, 2571.
[80] Aronov, A. M.; Munagala, N. R.; De Montellano, P. R.; Kuntz, I. D.; Wang, C. C. *Biochemistry,* **2000**, *39*, 4684.
[81] Liebeschuetz, J. W.; Jones, S. D.; Morgan, P. J.; Murray, C. W.; Rimmer, A. D.; Roscoe, J. M. E.; Waszkowycz, B.; Welsh, P. M.; Wylie, W. A.; Young, S. C.; Martin, H.; Mahler, J.; Brady, L.; Wilkinson, K. *J. Med. Chem.,* **2002**, *45*, 1221.
[82] Makino, S.; Kayahara, T.; Tashiro, K.; Takahashi, M.; Tsuji, T.; Shoji, M. *J. Comput. Aided Mol. Des.,* **2001**, *15*, 553.
[83] Burkhard, P.; Hommel, U.; Sanner, M.; Walkinshaw, M. D. *J. Mol. Biol.,* **1999**, *287*, 853.
[84] Burkhard, P.; Taylor, P.; Walkinshaw, M. D. *J. Mol. Biol.,* **1998**, *277*, 449.
[85] Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. *J. Med. Chem.,* **2000**, *43*, 401.
[86] Drews, J. *Science,* **2000**, *287*, 1960.
[87] Bajorath, J. *Nat. Rev. Drug Discov.,* **2002**, *1*, 882.
[88] Willett, P. In *Similarity and Clustering in Chemical Information Systems*. Research Studies: Letchworth, UK, **1987**.
[89] Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.,* **1996**, *39*, 3049.
[90] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. *J. Med. Chem.,* **2002**, *45*, 4350.
[91] Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 394.
[92] Xue, L.; Godden, J. W.; Bajorath, J. *Environ. Res.,* **2003**, *14*, 27.
[93] Godden, J. W.; Furr, J. R.; Xue, L.; Stahura, F. L.; Bajorath, J. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 21.
[94] Godden, J. W.; Stahura, F. L.; Bajorath, J. *J. Med. Chem.,* **2004**, *47*, 5608.
[95] Xu, Y.-J.; Johnson, M. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 181.
[96] Xu, Y.-J.; Johnson, M. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 912.
[97] Schreyer, S. K.; Parker, C. N.; Maggiora, G. M. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 470.
[98] Gillet, V. J.; Willett, P.; Bradshaw, J. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 338.
[99] Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 346.
[100] Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 2145.
[101] Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. *J. Chem. Inf. Comput. Sci.,* **1999**, *39*, 21.
[102] Pearlman, R. S.; Smith, K. M. *Perspect. Drug Discov. Des.,* **1998**, *9-11*, 339.
[103] Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.,* **1999**, *39*, 28.
[104] Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. *J. Med. Chem.,* **2005**, *48*, 240.
[105] Gao, H. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*, 402.
[106] Beno, B. R.; Mason, J. S. *Drug Discov. Today,* **2001**, *6*, 251.
[107] Pirard, B.; Pickett, S. D. *J. Chem. Inf. Comput. Sci.,* **2000**, *40*, 1431.
[108] Rusinko, A.; Young, S. S.; Drewry, D. H.; Gerritz, S. W. *Comb. Chem. High Throughput Screen,* **2002**, *5*, 125.
[109] Jones-Hertzog, D. K.; Mukhopadhyay, P.; Keefer, C. E.; Young, S. S. *J. Pharmacol. Toxicol. Methods,* **1999**, *42*, 207.
[110] Van Rhee, A. M.; Stocker, J.; Printzenhoff, D.; Creech, C.; Wagoner, P. K.; Spear, K. L. *J. Comb. Chem.,* **2001**, *3*, 267.
[111] Lemmen, C.; Lengauer, T. *J. Comput. Aided Mol. Des.,* **2000**, *14*, 215.
[112] Güner, O. F. *Pharmacophore: Perception, Development, and Use in Drug Design.* IUL Biotechnol. Ser: International University Line, **2000**.
[113] Mansfield, M. L.; Covell, D. G.; Jernigan, R. L. *J. Chem. Inf. Comput. Sci.,* **2002**, 42, 259.
[114] Van Drie, J. H. Pharmacophore discovery: A critical review. In *Computational Medicinal Chemistry for Drug Discovery;* Bultinck, P. Ed.; Marcel Dekker Inc., New York, N.Y. **2004**, pp. 437-460.
[115] Du, Q.; Arteca, G. A.; Mezey, P. G. *J. Comput. Aided Mol. Des.,* **1997**, *11*, 503.
[116] Petke, J. D. *J. Comput. Chem.,* **1993**, *14*, 928.
[117] Hodgkin, E. E.; Richards, W. G. *Int. J. Quantum Chem.,* **1987**, *14*, 105.
[118] Jenkins, J. L.; Glick, M.; Davies, J. W. *J. Med. Chem.,* **2004**, *47*, 6144.
[119] Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. *J. Med. Chem.,* **2004**, *47*, 6777.
[120] Cramer, R. D.; Jilek, R. J.; Andrews, K. M. *J. Mol. Graph. Model.,* **2002**, *20*, 447.
[121] Jilek, R. J.; Cramer, R. D. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 1221.
[122] Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. *J. Med. Chem.,* **1999**, *42*, 3919.
[123] Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. *J. Chem. Inf. Model.,* **2005**, *45*, 673.
[124] Rush, T. S. III; Grant, J. A.; Mosyak, L.; Nicholls, A. *J. Med. Chem.,* **2005**, *48*, 1489.
[125] Grant, J. A.; Gallard, M. A.; Pickup, B. T. *J. Comput. Chem.,* **1996**, *17*, 1653.
[126] Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 1230.
[127] Byvatov, E.; Schneider, G. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 993.

[128] Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 1882.
[129] Warmuth, M. K.; Liao, J.; Raetsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 667.
[130] Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 2019.
[131] Rabow, A. A.; Shoemaker, R. H.; Sausville, E. A.; Covell, D. G. *J. Med. Chem.,* **2002**, *45*, 818.
[132] Keseru, G. M.; Molnar, L.; Greiner, I. *Comb. Chem. High Throughput Screen.,* **2000**, *3*, 535.
[133] Keseru, G. M. *J. Comput. Aided Mol. Des.,* **2001**, *15*, 649.
[134] Molnar, L.; Keseru, G. M. *Bioorg. Med. Chem. Lett.,* **2002**, *12*, 419.
[135] Gao, H.; Shanmugasundaram, V.; Lee, P. *Pharm. Res.,* **2002**, *19*, 497.
[136] Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. *J. Med. Chem.,* **2004**, *47*, 4891.
[137] Muegge, I. *Med. Res. Rev.,* **2003**, *23*, 302.
[138] Ajay, A.; Walters, W. P.; Murcko, M. A. *J. Med. Chem.,* **1998**, *41*, 3314.
[139] Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. *J. Mol. Graph. Model.,* **2002**, *20*, 281.
[140] Jenkins, J. L.; Kao, R. Y. T.; Shapiro, R. *Proteins,* **2003**, *50*, 81.
[141] Raymond, J. W.; Jalaie, M.; Bradley, M. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 601.
[142] Gohlke, H.; Klebe, G. *J. Med. Chem.,* **2002**, *45*, 4153.
[143] Stahl, M.; Rarey, M.; Klebe, G. Screening of Drug Databases. In *Bioinformatics*, Lengauer, T. Ed.; Wiley-VCH: New York, **2001**, Vol. *2*, pp. 137-170.
[144] Belkin, N. J.; Cool, C.; Croft, W. B.; Callan, R. K. Effect of Multiple Query Representations on Information System Performance. *In Proceedings of SIGIR,* **1993**, pp 339-346.
[145] Raymond, J. W.; Willett, P. *J. Comput. Aided Mol. Des.,* **2002**, *16*, 59.
[146] Salim, N.; Holliday, J. D.; Willett, P. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 435.
[147] Whittle, M.; Willett, P. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 449.
[148] Maggiora, G. M.; Shanmugasundaram, V. Molecular similarity measures. In *Methods in Molecular Biology,* Bajorath, J. Ed.; Clifton: NJ, **2004**, Vol. *275*, pp. 1-50.
[149] Sheridan R. P.; Kearsley, S. K. *Drug Discov. Today,* **2002**, *7*, 903.
[150] Raymond, J. W.; Jalaie, M.; Bradley, M. P. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 601.

Received: January 18, 2006 **Revised: May 08, 2006** **Accepted: May 09, 2006**