Ligand-Based Design Workflow

Paul Hawkins and Geoff Skillman
OpenEye Scientific Software

Executive Summary

OpenEye Scientific Software has been providing computational chemistry software to the pharmaceutical industry since 1997. OpenEye provides a unique combination of superior cheminformatics and insightful biophysics in its applications, which are designed to fulfill the large-scale modeling needs of professional molecular modelers working in the pharmaceutical and biotech industries. OpenEye is proud to count fourteen of the top fifteen pharmaceutical companies worldwide as customers and as such, its tools have been used to generate novel leads, build cheminformatics infrastructures, and lead-hop numerous times in the "real world" of the pharmaceutical industry.

The applications provided by OpenEye Scientific Software enable users to efficiently and effectively perform virtual screening on large numbers of molecules to identify new structural classes of molecules likely to be active against a protein target of interest. OpenEye applications run on a wide variety of UNIX-based platforms, including Mac OS X, making the applications accessible to the wider pharmaceutical community, many of whom work in multiplatform environments. Development of many of the OpenEye applications has been greatly accelerated by the ease of use and power of the Mac OS X development tools.

This paper introduces a ligand-based approach to virtual screening based on a new method of describing molecular shape and chemistry as embodied in the application ROCS (Rapid Overlay of Chemical Structures). When taken together, ROCS and its supporting applications constitute an effective workflow for virtual screening that is capable of handling hundreds of thousands of molecules, both in the processing and visualization stages of the process.

Evidence from the literature and internal studies shows that the shape-based approach used by ROCS is an effective solution for ligand-based virtual screening. It also shows that the ROCS approach is competitive with structure-based tools for virtual screening.

Contents

Overview

About OpenEye Scientific Software
OpenEye provides software for molecular modeling and cheminformatics to the pharmaceutical industry. It has done so since 1997 in its continuing mission to provide novel software, new science, and better business practices to the industry. Central to its approach is the importance of shape and electrostatics as primary variables of molecular description, platform-independent code for high-throughput 2D and 3D modeling, and a preference for the rigorous rather than the ad hoc. Specific areas of application include chemical informatics, structure generation, docking, shape comparison, charge and electrostatics, and visualization. The software is designed for scientific rigor, as well as speed, scalability, and platform independence. OpenEye makes most of its technology available as toolkits—programming libraries suitable for custom development. Typically, OpenEye software is distributable across multiple processors, supports 64-bit processing, and runs on Mac OS X as well as a variety of other platforms.

Virtual screening is the application of computational models to select or prioritize compounds for experimental screening. Every molecule in a compound database is assigned a score based on some metric, the molecules are ranked on this score, and a very small fraction of the highly ranked molecules is selected for experimental screening. The available methods for scoring molecules range from the relatively simple to the complex. There are two broad categories of virtual screening techniques: ligand-based design and structure-based design.

Ligand-based design methods capitalize on the fact that ligands similar to an active ligand are more likely to be active than random ligands. Ligand-based approaches commonly consider two- or three-dimensional chemistry, shape, electrostatic, and interaction points (e.g., pharmacophore points) to assess similarity. Structure-based design attempts to use the 3D protein structure to predict which ligands will bind to the target. While the structure-based models give the impression of being more realistic, the complexity of protein-ligand complex free energy interactions dictates that many significant approximations must be utilized. In practice, these approximations can severely limit the efficacy of structure-based design and ligand-based design methods can be shown to perform similarly in virtual screening applications.

The amount and quality of information required to apply these techniques varies. Ligand similarity approaches (be they 2D or 3D) require only a single active molecule, which may come from the literature, patents, or in-house experimental data. In these cases, activity might be determined only by an inaccurate high-throughput screen. Ligand-based Quantitative Structure-Activity Relationship (QSAR) approaches require a number of active molecules spanning a wide range of activity against the target receptor (three orders of magnitude is the minimum range). The quality of the QSAR model depends to a large extent on the quality of the activity data, such that reliable QSAR models are usually built based on carefully acquired binding or inhibition data (and not on inaccurate high-throughput screening data).

On the structure-based side, docking requires an experimental structure or a computational model of the protein structure (a homology model) of the target protein. Pharmacophore models that include receptor information require an experimental structure of the complex between an active molecule and its target protein. Further, docking to protein structures that do not have a ligand present or a homology model dramatically reduces the expected performance of structure-based design.[1] An effective virtual screening application that requires only the minimal amount of information (a single active molecule) is desirable, especially for use in the early stages of a discovery program when little is known about the space of active molecules.

Compound selection and filtering

The number of molecules that is available for screening is large. It is estimated that there are 10 million molecules available for purchase. A large pharmaceutical company might have a corporate compound collection of five or six million compounds. Given the large number of available molecules that might be used in virtual and/or experimental screening, it is useful to be able to remove molecules that possess undesirable properties. Such properties may well vary from project to project and target protein to target protein, but some general rules are widely applied. A short list of properties generally considered undesirable in molecules entering experimental screening could include:

- High likelihood of combining covalently with the target protein
- Possession of toxic functionality
- Low likelihood of oral bioavailability
- Possession of properties likely to interfere with the experimental assay

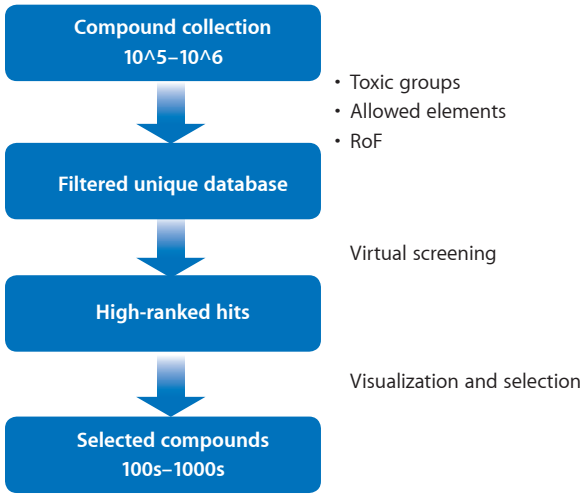
It is advisable to remove these kinds of molecules before the virtual screening step, ensuring that those compounds that are selected by the virtual screen are suitable for submission to the experimental screening protocol. There are a number of well-known examples of each of the four classes of filters mentioned above. Filters to predict oral bioavailability have been explored extensively;[2] the best known of the bioavailability filters is the Lipinski Rule of Five (RoF).

While the application of virtual screening to large compound databases can remove a large number of undesirable compounds in an automated way, there is no substitute for human intervention. An experienced computational chemist can provide valuable quality control to the results of a virtual screen, ensuring that the right compounds have been selected for the right reasons. A tool that allows easy visualization, manipulation, and evaluation of hundreds or thousands of compounds helps the computational chemist to inspect "hit lists" from virtual screens and judge the quality and relevance of the results.

A summary of the stages of a generalized virtual screening workflow can be found in Figure 1. A filtering step first removes inappropriate compounds. The subsequent step assigns a score to each of the molecules that passed the filtering step. The kinds of techniques used to assign scores were outlined briefly in the Overview section. The molecules are then ranked according to this score. The final step is the selection step, wherein the highest-scoring compounds are examined and selected by a modeler based on criteria specific to the project. Such criteria could include the following:

- Is a good visual fit to the model used in the virtual screening
- Possesses interesting functionality—chemistry that might make a molecule a good starting point for elaboration into possibly more potent analogues
- Fits into a known QSAR model
- Tests a unique binding hypothesis
- Has a structural type that is not covered by a competitor's patents

Figure 1. A generalized virtual screening workflow.



Ligand-based virtual screening

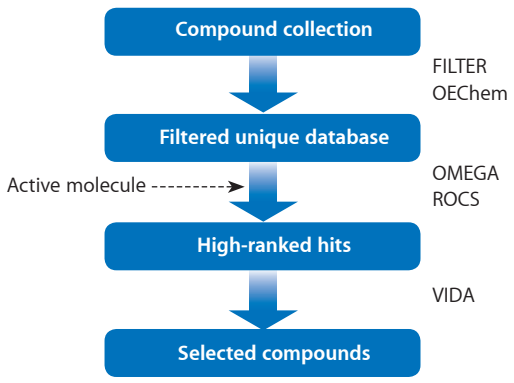
The techniques commonly used in virtual screening can be divided broadly into two parts: structure-based techniques and ligand-based techniques. Structure-based approaches, of which the best known is docking, require a protein structure or homology model as a starting point. On the other hand, ligand-based approaches require much less detailed information. At a minimum, a ligand-based technique requires knowledge of only one active molecule. The virtual screen is then conducted by identifying molecules that share some similarity or properties with that single active molecule.

Given that molecules bind to their target receptor and exert some effect upon that receptor in a three-dimensional manner, there has been a continued interest in 3D techniques in ligand-based virtual screening. The best known among these techniques is the pharmacophore approach, which attempts to abstract features of an active molecule (or shared features among a set of active molecules) that are likely to be important in binding to the target receptor. The virtual screen is then conducted by identifying molecules that could potentially match this set of features.

Other 3D ligand-based approaches include shape similarity, which attempts to score database molecules based on their overall shape similarity to a query molecule, rather than just on the molecule's ability to match an abstracted set of features, as pharmacophore tools do. Further discussion of the shape concept and its implementation in the OpenEye tool ROCS can be found on page 10.

An example workflow for ligand-based virtual screening using tools from OpenEye is shown in Figure 2. Preprocessing steps remove undesirable and duplicated compounds, shape-based similarity scores are computed with ROCS, and the highly ranked hits are visually inspected with VIDA, the OpenEye visualization tool.

Figure 2. Use of OpenEye tools in ligand-based virtual screening.



Ligand-based virtual screening with OpenEye tools

Software development at OpenEye Scientific Software

All fourteen of OpenEye's molecular modeling applications run in the Mac OS X environment and support both Intel and PowerPC hardware. To reliably generate this kind of cross-platform support, more than seventy percent of OpenEye developers use Mac OS X regularly in their development process.

A basic tenet of OpenEye's development policy is cross-platform compatibility. Apple's dedication to industry standards and the UNIX foundation of Mac OS X provide easy portability to and from other UNIX environments. As an added benefit, with the introduction of Intel processor–based machines, GCC on Mac OS X is one of the fastest compilers the company uses. Developer productivity is also enhanced by rapid compilation and testing of new code.

Many of the applications of molecular modeling in the pharmaceutical industry are computationally intensive and must be run on clusters of hundreds of processors. In such an environment, computational efficiency is essential. OpenEye regularly relies on Shark, part of the CHUD developer tool set, to inform and guide its optimization processes. Shark allows analysis of memory, cache, and processor bottlenecks of either optimized or debugged code without the need to link to any external libraries. Shark provides an intuitive graphical user interface to explore problems in a top-down or bottom-up manner at either the source code or assembly level. Shark highlights trouble spots and provides useful hints and tips for more efficient coding. In one recent short session, Shark helped to identify a subtle cache-miss that, when fixed, resulted in a significant increase in speed in one of OpenEye's flagship applications.

In this section, a set of OpenEye tools useful in ligand-based screening will be discussed. Specifically, four tools will be introduced: FILTER for removal of undesirable compounds, OMEGA for the generation of conformers, ROCS for ranking by three-dimensional shape similarity, and VIDA for 3D visual inspection.

FILTER

As noted above, the first step in any virtual screening protocol is the removal of undesirable compounds from the compound database to be searched. The OpenEye tool FILTER provides the ability to filter compounds based on a number of criteria:

- Possesses given functional groups or elements
- Has properties that exceed set limits, e.g., molecular weight, solubility
- Is incorrectly rendered in the database, i.e., has 5-valent carbon atoms
- Has combinations of properties likely to result in low oral bioavailability, e.g., is not compliant with the Rule of Five.

FILTER utilizes filter files—text files that define the range of properties and functional groups that are permissible in molecules. Two such files are provided with FILTER, one aimed at identifying molecules that could be suitable as drugs (filter_drug) and one that attempts to identify compounds that could be lead molecules (filter_lead). The drug-like filter permits larger and more highly functionalized molecules, while the lead-like filter only allows smaller molecules, which are more suitable as starting points for an optimization program.

FILTER can also set a consistent (neutral pH) protonation state on all molecules in a database, for example, ensuring that all carboxyl groups are deprotonated and all alkylamines are protonated. This consistent approach can compensate for the different rendering of the same functional group in different molecules that often occurs in vendor databases and can sometimes occur in corporate collections. Handling functional groups in this way ensures that downstream applications will score molecules based on a charge or protonation state that is consistent. This will ensure that differences in scoring result from meaningful differences between molecules, not simply from differences in charge state arising from different protonation of the same functional group.

The problem of duplicated compounds in the screening dataset arises frequently, especially when compounds in the dataset have been acquired from different sources. FILTER can be used to eliminate duplicate compounds rapidly as part of the filtration process. This way, there is no time wasted on processing duplicates later in the workflow.

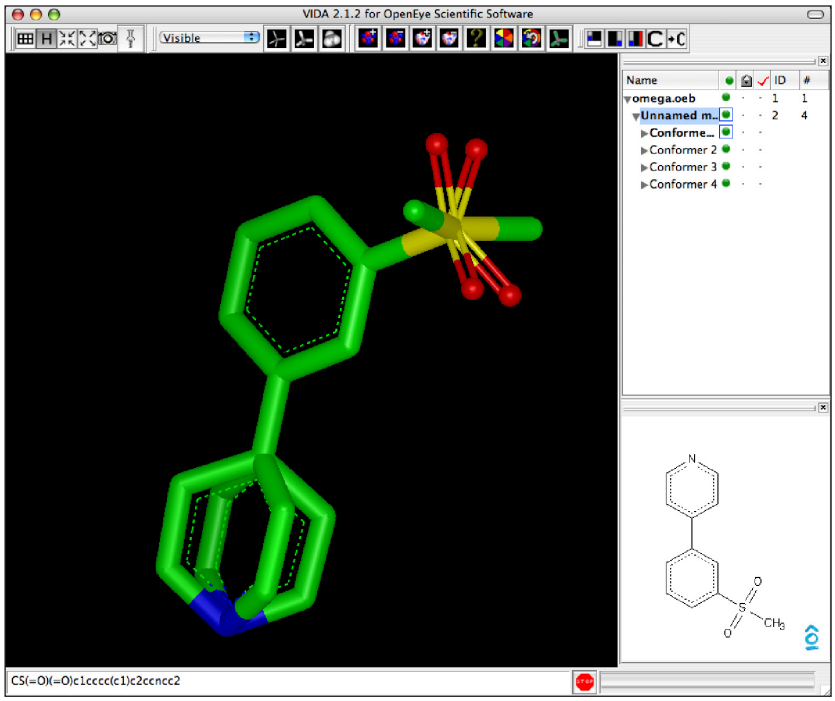
OMEGA

In order to provide optimal efficiency, many OpenEye tools in the virtual screening arena operate on pregenerated conformer ensembles rather than by manipulating conformers during the scoring process. OpenEye tools such as ROCS require that the database of molecules to be screened contain multi-conformer ensembles for each candidate molecule.

OMEGA constructs conformers for a molecule using a systematic, rules-based approach. The rules are derived from analyses of the conformations of molecules found in experimental structural databases such as the CSD and the PDB. OMEGA ensures that the conformers produced are low in energy by use of the Merck Molecular Force Field (MMFF94). The combination of systematic, rules-based conformer production with an appropriate energy cutoff ensures that conformers from OMEGA are both diverse and explore low-energy regions of conformational space, including those close to the experimental conformation.

An illustration of the rules-based nature of conformers from OMEGA is shown in Figure 3.

Figure 3. Conformations generated by OMEGA.

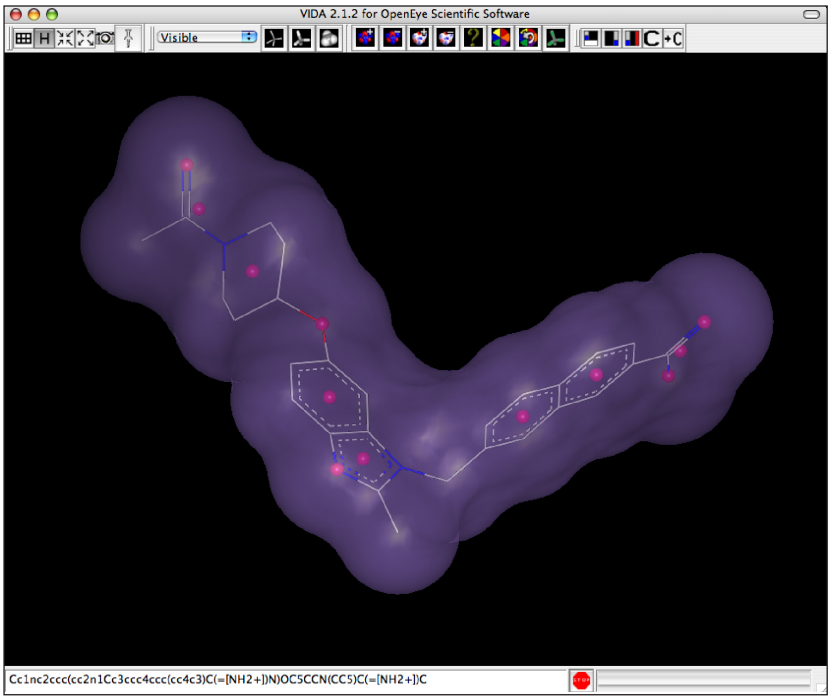


ROCS

The OpenEye tool ROCS moves beyond the abstraction approach of pharmacophores, utilizing the entire molecular shape of the query molecule, along with its chemical features. Database molecules are scored relative to a query molecule using a physically rigorous measure of three-dimensional similarity, along with a measure of the level of matching of appropriate chemical functionality (e.g., a hydrogen-bond donor in the database molecule matching spatially with a hydrogen-bond donor on the query molecule). Figure 4 shows a molecule contained within its molecular shape, with its chemical features denoted by pink spheres. The features recognized by ROCS include:

- Hydrogen-bond donors and acceptors
- Positive and negative charges
- Rings
- Hydrophobic groups

Figure 4. Shape and features of a molecule as perceived by ROCS.



As alluded to in the OMEGA discussion above, ROCS requires a database of conformers that it compares to the query molecule. Each conformer of each database molecule is overlaid rigidly on the query molecule, and the overlap of molecular volume between the query and the database conformer is optimized. Then, a measure of shape similarity between the query and the database conformer (the shape Tanimoto coefficient) is calculated. Once all conformers of the database molecule have been overlaid and the shape Tanimoto calculated, the conformer with the highest shape Tanimoto (highest shape similarity) is saved, along with the overlay of that conformer with the query molecule. Figure 5 shows an example of an overlay obtained from ROCS of a query molecule (shown with its molecular surface), with another molecule of a very different chemical structure.

Figure 5. ROCS overlay of two molecules active against Factor Xa. The query molecule is shown in green; the molecular surface is that of the query molecule.

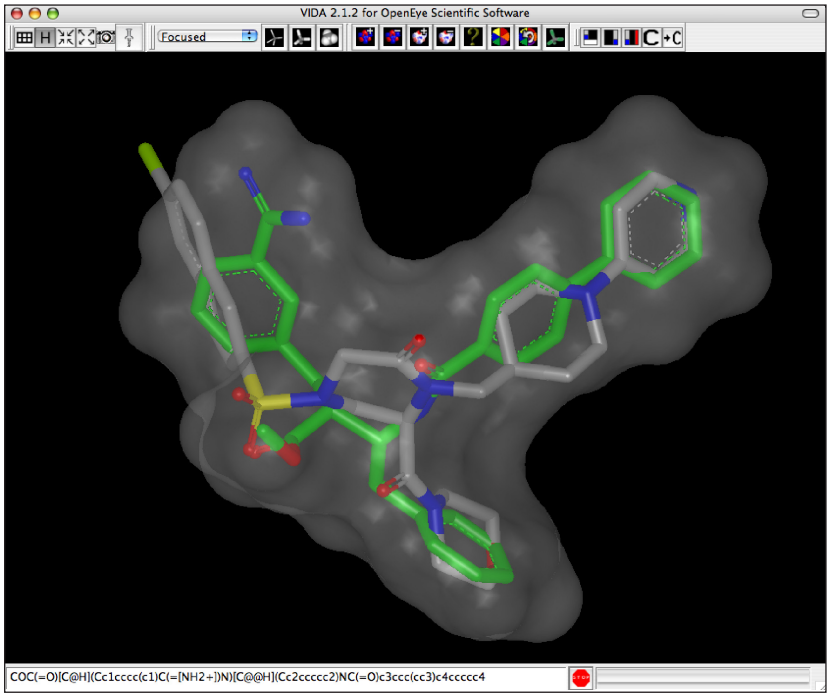


Figure 5 shows that the ROCS application's combination of a rigorous approach to shape matching and a simple estimation of chemical similarity imparts high scores not only to molecules that are structurally similar to the query, but also to molecules from a chemical class very different from the query.

Despite the rigorous approach to shape taken by ROCS, the molecular overlay and computation of the shape similarity metric are extremely rapid (1200 overlays per second on an iMac with 1.83GHz Core Duo), enabling the processing of large databases at rates of up to 15 molecules per second. This is much faster than many other ligand-based approaches.

The 3D shape-based approach implemented in ROCS excels at identifying compounds with chemical structures that may be very different from that of the query molecule, but that can interact with a protein in a similar manner. This is a critical task that arises often in the drug design process.

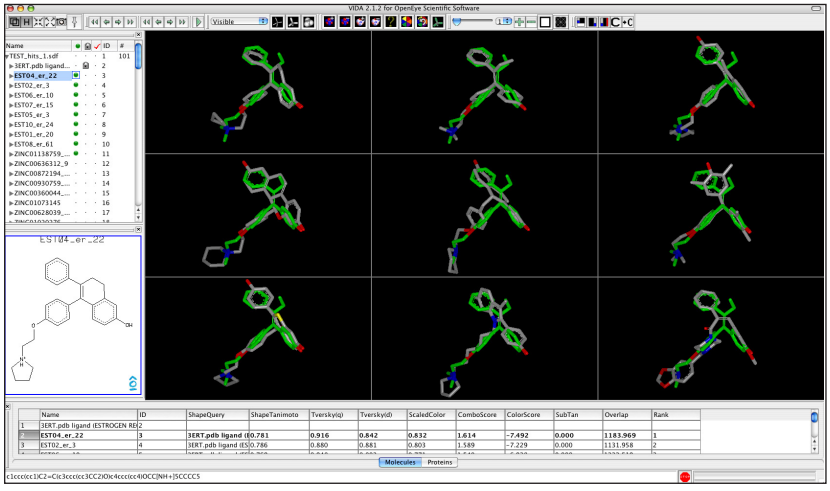
VIDA

Whether structure-based or ligand-based design is used to propose molecules of interest, the final stage in virtual screening is often 3D visual inspection. For many modelers, this is the best method to bring their professional expertise, the specific requirements of the project, personal preferences of the medicinal chemists on the project, and subtle aspects of the known structure-activity relationships of the project to bear.

Modelers may examine hundreds or thousands of ligands in a day, either alone or in a protein environment. On Mac OS X, VIDA offers facile comparison and annotation of ligands. A user can examine the molecules one at a time or many at a time, in either a single pane or multiple panes, and render them as a 3D object or depict them as a 2D chemical structure familiar to chemists. To make these graphical investigations feasible, OpenEye recommends at least a Mac Pro or MacBook Pro with 128MB or 256MB of video memory and a high-end graphics card.

When molecular modelers and chemists use structure-based design, a critical phase of the workflow is to gain a detailed understanding of the overlap of shape and chemical features between putative ligands and the query molecule. Visualization of the protein binding site (if it is available) with 3D stereo is an indispensable tool during this familiarization process. VIDA offers hardware stereo 3D images on the Mac Pro or Power Mac G5 with the NVIDIA Quadro FX 4500 graphics card, compatible Stereographics glasses, and running Mac OS X 10.4 or later. With this type of setup, a modeler can rapidly gain familiarity with a protein binding-site structure and gain deeper insights into the relative position and potential interactions of a potential binding region than are readily available with conventional 3D visualization.

Figure 6. Vida integrates 2D, 3D, and spreadsheet views of molecular data and allows simultaneous comparison of multiple molecular overlays.



Results from ROCS

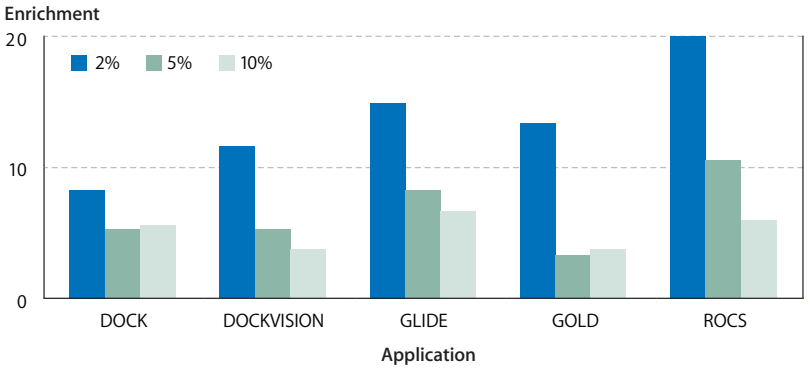
In this section, results from the literature and from recently generated data on the use of ROCS in virtual screening and lead-hopping will be discussed.

Virtual screening

A few examples of the utility of ROCS in virtual screening will be presented. In each case, a dataset from a published paper on virtual screening will be used to assess the performance of ROCS. Since the same compound sets were used in the publication and in the ROCS study, the published results and those of ROCS can be compared directly.

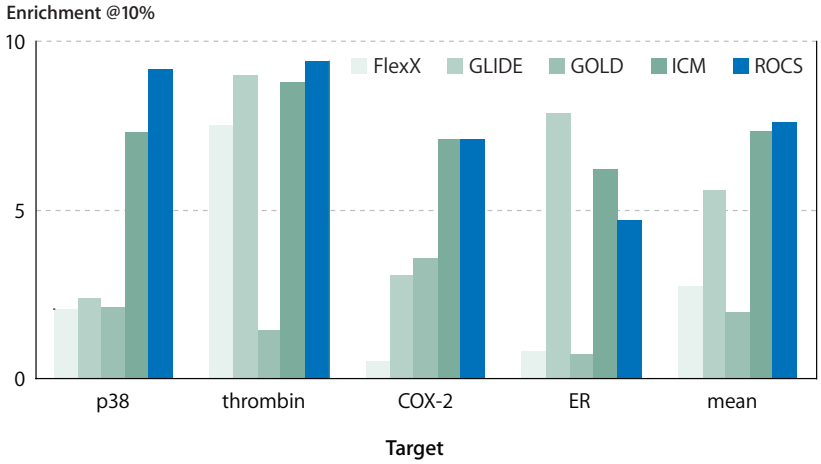
The first study was published by staff at Johnson and Johnson.[3] They compared the performance of a number of well-known docking tools (DOCK, DOCKVISION, GLIDE, and GOLD) when used for virtual screening against three proteins (PTP-1B, thrombin, and HIV-1 PR). The results from the paper and the results when using ROCS for screening the same compounds are shown in Figure 6: ROCS outperforms the four docking tools at almost every point.

Figure 7. Average enrichments for virtual screening across PTP-1B, thrombin, and HIV-1 PR proteins when using docking tools and ROCS.



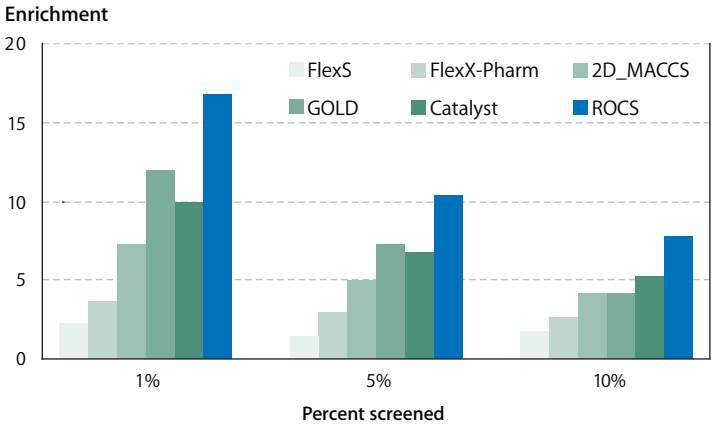
A related study was published by workers from Astra-Zeneca,[4] in which a different set of docking tools (FlexX, GLIDE, GOLD, and ICM-Dock) were used in virtual screening against a different set of protein targets (COX-2, the estrogen receptor, p38 kinase, and thrombin). As can be seen in Figure 7, ROCS performs as well as ICM-Dock and better than all the other docking tools when averaged across the four targets studied.

Figure 8. Average enrichments for virtual screening across four targets, using docking tools and ROCS.



In the cases shown above, docking was carried out into experimental crystal structures. There are a large number of receptors of current medicinal interest that do not have available crystal structures, e.g., the GPCRs and ion channels. To perform docking into GPCRs, a homology model must first be built and validated. As mentioned above, one should expect significant degradation of structure-based design results when using a computational model of this nature. A virtual screening experiment was undertaken by staff at Sanofi-Aventis[5] using docking into homology models. They compared the success of docking into homology models for four GPCRs (5HT2A, A1A, D2, and M1) with virtual screening using ligand-based tools. Two docking tools were used in the study (FlexX-Pharm and GOLD), along with two 3D ligand-based tools (Catalyst and FlexS). Figure 8 compares the performance of these four tools with the performance of ROCS and the performance of the MACCS fingerprints, a two-dimensional similarity measure. ROCS outperforms both the structure-based (docking) and the ligand-based tools by a significant margin.

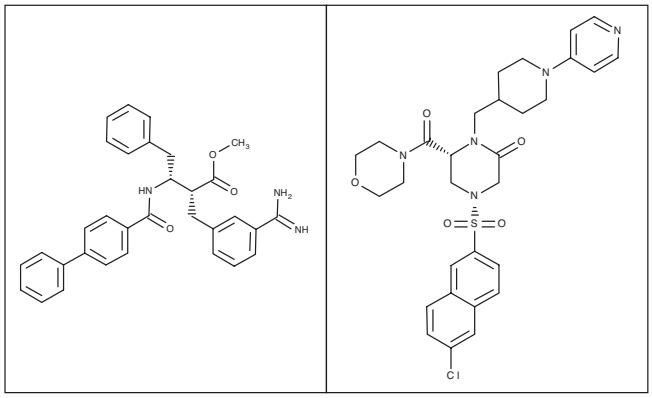
Figure 9. Average enrichments for virtual screening on GPCRs (5HT2A, A1A, D2, and M1) from a variety of 3D and 2D approaches.



Lead-hopping

The shape-based approach embodied in ROCS allows rapid identification of hit compounds possessing the appropriate shape to match a protein binding site. While the hit compounds may derive from very different chemical series, they show significant similarity in their shapes. As such, ROCS can be used with good results in lead-hopping, or moving from one class of active compound to another class with a different chemical structure. A simple example is shown in 3D in Figure 5 above, while the 2D structures of the same molecules are shown below in Figure 9.

Figure 9. Lead-hopping with ROCS. Query molecule is shown at left, hit is shown at right. Note the radically different chemical scaffolds.



In all the previous examples, the experiments have been retrospective, meaning they have assessed efficacy in ranking known ligands. Clearly, prospective applications of a tool provide a more powerful validation of the technology under investigation. For a successful prospective use of ROCS in virtual screening that was recently published, see Bologa et al.[6] Two other recent publications detail the use of ROCS in prospective virtual screening and lead-hopping (Rush et al[7] and Muchmore et al[8]).

Adding flexibility to the workflow

Though the OpenEye tools discussed in the workflows above function seamlessly together in the Mac OS X environment, many experienced users desire to integrate these OpenEye workflows into a larger research environment. This can require individualized molecular manipulation, file-format interconversion, molecular tagging, and data analysis. All these tasks and many more fall under the umbrella of OEChem, OpenEye's cheminformatics programming toolkit. On Mac OS X, OEChem is available in the C++, Python, and Java languages. OEChem uses multiple models of chemistry in order to achieve the most reliable file-format conversion in the industry, backed up by industry-leading molecular data integrity that is required in the molecule-centric pharmaceutical world.

OEChem is regularly used in the pharmaceutical industry for a variety of cheminformatics tasks, from simple one-off molecular manipulation to enterprise-level infrastructure building. Perhaps the most visible and large-scale deployment has been the National Institutes of Health's extensive use of OEChem to build the cheminformatics infrastructure behind the PubChem project.

Summary

Appropriate prefiltering of databases (using tools such as FILTER with rules crafted specifically for the project in question) and ranking of these filtered databases using ROCS is a powerful combination for ligand-based virtual screening. Comparison experiments with data from the literature show that the ROCS solution is superior to many structure-based approaches in virtual screening. Prospective experiments support the utility of ROCS in virtual screening and provide powerful examples of the use of the shape-based approach for lead-hopping.

The power and flexibility of the OEChem cheminformatics toolkit allow users to easily and rapidly develop and customize their own workflows that integrate tools developed by OpenEye and many other vendors.

[1]McGovern, S.L., Shoichet, B.K., J. Med. Chem. 46, 2895 (2003).
[2]Martin, Y.C., J. Med. Chem., 48, 3164–3170 (2005). Veber, D.F., et al, J. Med. Chem., 45, 2615–2623 (2002). Egan, W.J., et al, J. Med. Chem., 43, 3867–3877 (2000). Lipinski, C., et al, Adv. Drug Deliv. Rev., 23, 3 (1997).
[3]Cummings et al, J. Med. Chem., 48, 962 (2005).
[4]Chen et al, J. Chem. Inf. Model., 46, 401 (2006).
[5]Evers et al, J. Med. Chem., 48, 5448 (2005).
[6]Bologa et al, Nature Chem. Biol., 2, 207 (2006).
[7]Rush et al, J. Med. Chem., 48, 1489 (2005).
[8]Muchmore et al, Chem. Biol. Drug Des., 67, 174 (2006).