BIOC218 Computational Molecular Biology
Course Project, Brad Mills

Towards Virtual Screening: Exploring the Potential (and Pitfalls) of Docking Software

Introduction

For my course project I chose to explore docking software and its application to virtual screening. While there are many good reviews in the literature on docking software (e.g. Schneider, 2002), I wanted to supplement this material with some firsthand experience. I was able to gain access to three commercial docking programs, and while the software is both sophisticated and complex, was able to make reasonable strides toward exploring the potential of the software. My experiments also uncovered some of the challenges in implementing this software in a larger virtual screening environment.

Virtual screening is of interest in the pharmaceutical industry because it provides one mechanism for addressing a key question: which of the compounds at an organization's disposal make the most sense to run in a physical screen (Lyne, 2002). Since it would be both cost and time prohibitive to run what could literally represent hundreds of thousands of compounds, the idea of reducing the potential list through in silico or virtual screens using a computer is attractive. Several approaches have been taken in virtual screening. The concept of a compound being "drug-like" has received a lot of attention, e.g. Lipinski (2000) who initiated what has now become renowned as the "Rule of 5", Veber (2002) who has characterized molecule descriptors that help predict bioavailability, and other descriptor based work exploring the prediction of ADME/Tox properties. These methods are typically based on the physical characteristics of compounds, e.g. molecular weight or number of rotatable bonds, and are often made in the absence of information related to the potential target protein. This type of virtual screening is often applied successfully in concert with docking (Schneider, 2002). Docking, on the other hand, specifically examines the protein target that a small compound ligand or another protein might interact with and attempts to predict the binding affinity. While docking algorithms have a significant Chemistry or Physics bend to them, the concept is absolutely critical in Biology – simply put, proteins function when they are bound to other molecules (Halperin, 2002). Most of the current docking work and reviews have focused on protein-ligand docking as a starting point due to its lower complexity.

Materials and Methods

Selection and Preparation of Test Molecules

Since I wanted to gain some hands-on experience with docking software, I needed a set of molecules with known indications to experiment with. While the concept of virtual screening usually implies large datasets, I was looking for a smaller set of compounds that would be more manageable and would not consume a lot of time or computer resources to evaluate. An article by Stahl and Rarey (2001) provided such a test bed, since SMILES strings (Weininger, 1988) for many of the studied compounds were published as a supplement to the article. I used a scanner to scan in the strings, made a few corrective edits, and then used the Concord program available under the Sybyl package (Tripos) to convert the SMILES strings into single 3D conformations in mol2 format. While Concord complained about the valence of some of the published strings (and many of the other compounds from the paper were not published since they were proprietary), I was able to create a test set of known inhibitors with the following distribution:

Protein Target	**Molecule Count**
Estrogen Receptor	30
MAP Kinase p38	10
Thrombin	50
COX-2	10

This same set of 100 SMILES strings was passed to the Omega software package to generate the multiconformer libraries used by the FRED docking software (OpenEye Scientific Software). Although the SMILES format is a common standard, Omega was not able to process 16 of the strings. The test bed for the FRED program runs, therefore, was comprised of 84 molecules, 49 of which Stahl had collected from the literature as known thrombin inhibitors.

Preparation of Target Structures

While I was going to be dealing with three different docking software packages and three separate protein targets, my preparation of the targets initially followed the same strategy – loading a PDB structure (Berman, 2000) that was bound to an inhibitor into the Sybyl package and isolating an active site box that surrounded the reference molecule by 6.5 angstroms. A mol2 file describing the active site box was fed directly into the FlexX (Tripos) and FRED software. The GOLD software (Tripos) could not work with the mol2 format directly, so I converted the Sybyl file to a simple list of atoms for those runs. Hydrogens also had to be explicitly added to the PDB files for the GOLD runs since it uses an all-atom model, and the bound reference inhibitor was removed using Sybyl since documentation did not state explicitly if it would be ignored. I followed Stahl's lead in selecting PDB files for the estrogen receptor (1ERR) and thrombin (1DWD), although his paper mentions several processing steps that I could not mimic (due to both lack of access to proprietary software or my limited Chemistry knowledge). The p38 MAP Kinase structure used in the literature was proprietary, so I searched PDB for published structures and chose 1OUK. (The COX-2 inhibitors were added to the test molecules simply as negative test cases and to round the number of test molecules up to an even 100.)

Docking Software and Scoring Functions

Docking is generally viewed as three separate components: the representation used to model the ligand and protein interaction, the algorithm used to search the space of potential interactions, and a scoring function used to evaluate the quality of the interactions (Halperin, 2002). Because the search method and the scoring function can operate independently, you often see papers evaluating various combinations of docking software and scoring functions (e.g. Bissantz, 2000). Another important characteristic of docking programs is rigid vs. flexible modeling. In order to reduce the complexity of the problem, i.e. reduce the allowed degrees of freedom to a computationally manageable number, proteins are typically treated as rigid structures and the smaller ligand molecules are given some limited level of flexibility (Amit Singh, from archived course lecture). In this project, I used the FlexX docking software with both its original and DrugScore scoring functions, the FRED software with its native scoring function, and the GOLD software with both its default (GOLDScore) and ChemScore scoring function. The characteristics of the individual software packages and the scoring functions are discussed briefly below:

FlexX was developed by a research group from Germany (Rarey, 1996) and is distributed commercially by Tripos. The name of the software comes in part from its ability to model flexible ligands. The software uses an incremental construction method that samples the conformation space of the ligand and places it incrementally into the active site. The default scoring function for FlexX is an empirical method based on work by Bh̄n (1994) that estimates the free binding energy of the protein-ligand complex. DrugScore is a knowledge-based scoring function developed later by members of the same German research group (Gohlke, 2000). It is based on statistical analysis of structural information extracted from crystallographically determined protein-ligand complexes.

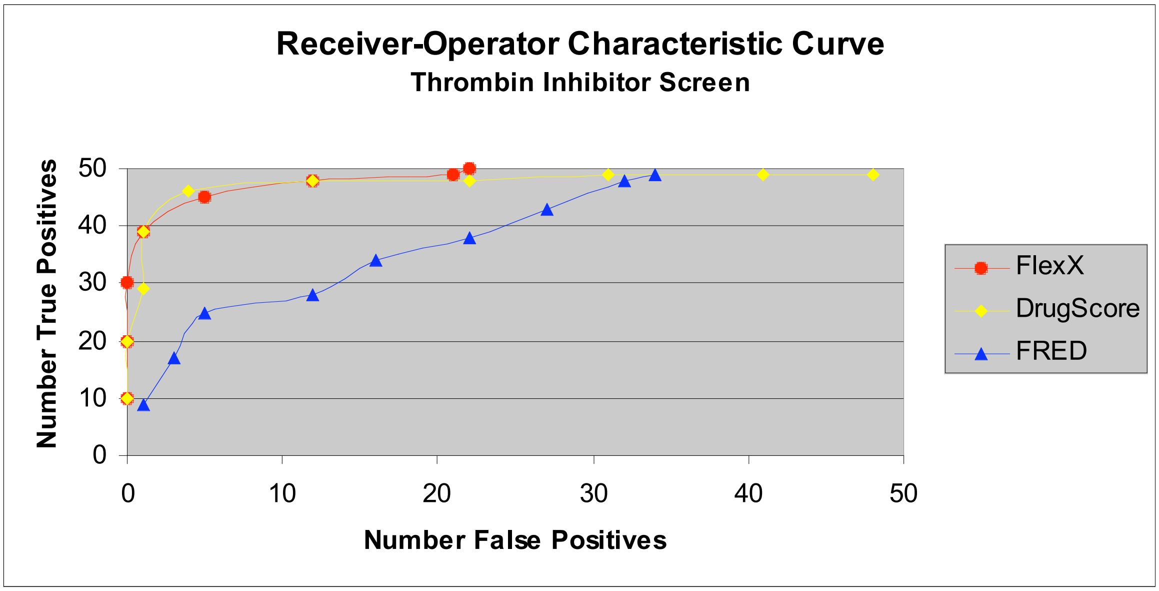
FRED is an acronym for Fast Rigid Exhaustive Docking. It is an implementation of multiconformer docking, meaning that a conformational search of the ligand is first carried out, and all relevant low-energy conformations are then rigidly placed in the binding site. This two-step process allows only the remaining six rotational and translational degrees of freedom for the rigid conformer to be considered. The FRED process uses a series of shape-based filters, and the default scoring function is based on Gaussian shape fitting (Schulz-Gasch, 2003).

GOLD is an acronym for Genetic Optimisation for Ligand Docking. It uses a genetic algorithm to explore the full range of ligand conformational flexibility and allows for partial flexibility of the protein target (Jones, 1997). The original GOLDScore fitness function has four primary energy components: protein-ligand hydrogen bonding, protein-ligand van der Waals, ligand internal van der Waals, and ligand torsional strain. The fitness function parameters are empirical based, and there is also an empirical correction to encourage protein-ligand hydrophobic contact. GOLDScore has been optimized for the prediction of ligand binding positions rather than the prediction of binding affinities (GOLD user's manual). ChemStore was originally developed independently and adapted for docking (Baxter, 1998). It was derived empirically from a set of 82 protein-ligand complexes, and unlike GOLDScore, it is optimized based on measured binding affinities. Free energy changes during ligand binding are estimated from hydrogen-bonding, acceptor-metal, and lipophilic interaction terms. The final ChemScore value also includes a clash penalty and internal torsion term to penalize close docking contacts and poor internal ligand conformations.

Results and Discussion

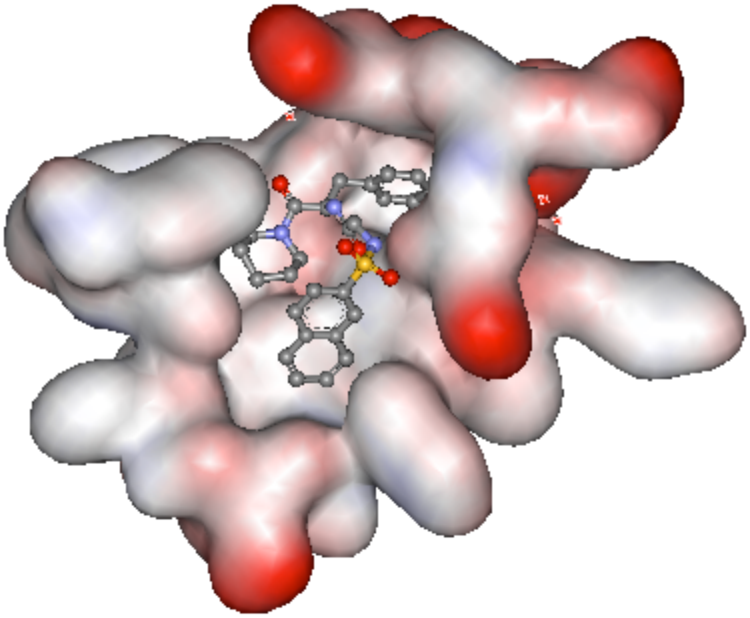
When I set out to explore docking software, it quickly became obvious that both the software and the actual docking problem were quite complex. The project I hoped to implement represents what is often referred to as “unbound” docking (Erickson, 2004). In unbound docking, one simplifies the challenge somewhat by supplying the software with the actual binding site based on a reference molecule and structure. The problem is nonetheless tougher to solve than the “bound” docking problem, where you are “simply” looking to dock the reference molecule back to its binding site. Since I was not dealing with the typical large database scenario found in compound screening, my experiments could be analyzed as a classification problem where the software was evaluated based on its ability to discern members of the set from non-members. I did not evaluate docking accuracy as measured by root-mean-square deviation, but still faced plenty of challenges in setting up the various algorithm runs and looking for biological insight to help guide them toward reasonable discrimination.

The largest set in the compound test bed contained 50 thrombin inhibitors. Thrombin was already of interest to me as I had used it when evaluating quantitative pattern matching in homework 3. It is a biologically interesting enzyme because of its involvement in the blood-clotting cascade with implications for both human victims of stroke and sepsis as well as hemorrhagic disease in cattle fed moldy clover hay (Stryer’s Biochemistry, 2002). Since I had isolated the active site using the Sybyl software, the first docking software that I evaluated was FlexX. Using a Receiver-Operator Characteristic (ROC) curve below as we did in homework 5, one can see that FlexX with its default scoring method did an excellent job of distinguishing the thrombin inhibitors from the other molecules:

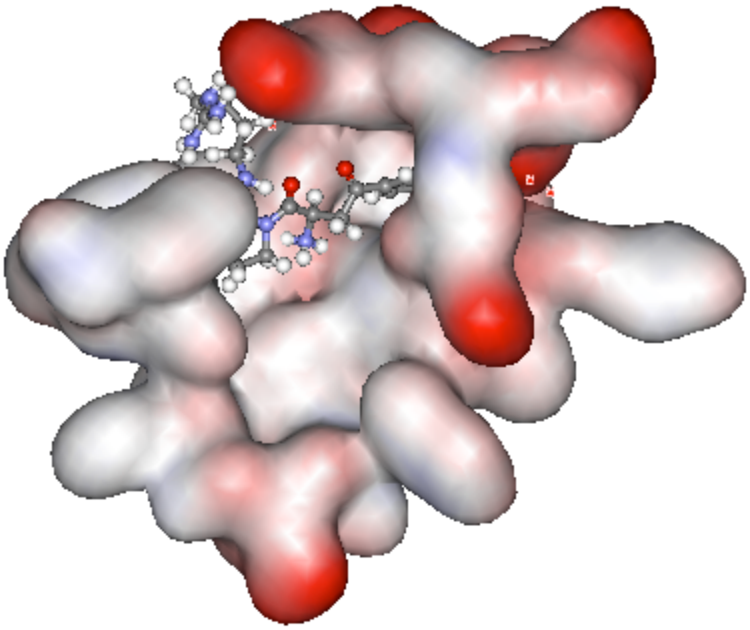


The 30 highest scores from the software are all for thrombin inhibitors, and 45 of the known inhibitors (90%) fell in the upper half of the reported scores. The lowest scoring known inhibitor was 72nd in rank order, so I have not graphed scores below that as false positives. In practice, one would want to pick a cutoff score value that would divide the predictions into an acceptable balance of false positives vs. false negatives. Although this is a small test set from which to make a call, using a cutoff score of –25 (see attached spreadsheet with run details) would give you a precision of 0.93 and a sensitivity/recall of 0.84. This cutoff should be dictated by the capacity of physical screening capabilities and the beginning population size of compounds. If you were starting out with tens of thousands of compounds, one might have to be even more conservative by attempting to eliminate all false positives at the risk of missing good potential leads; using a score of –26.5 would maximize precision at 1 but reduce recall to 0.72. Using the alternate DrugScore scoring function with FlexX also led to good discernment between known thrombin inhibitors and the other compounds. There were a few additional false positives earlier in the ranking of test scores, but one could get both a precision and sensitivity of 0.92 if one had the capacity to physically screen half of the ranked compound population. DrugScore did struggle a little more in picking up a few of the thrombin inhibitors (and in fact listed one of the known inhibitors as undockable), but this would have little relevance to virtual screening in practice since one would not normally expect to be able to set a cutoff score that would include all true positives without being overwhelmed by false positives. FlexX has been studied frequently in the literature, and those findings do not contradict its good performance here: Stahl (2001) notes that FlexX performs well with its default scoring algorithm for those target-ligand combinations that form a significant number of hydrogen bonds, and he includes thrombin in that category. The literature evaluations of DrugScore were a little less consistent, with Gohlke (2000) citing superior performance to the default FlexX scoring function, while Stahl (2001) observed that the knowledge-based method on average performed worse than its empirical counterpart and noted that it could model lipophilic interactions well, but fell well short when modeling hydrogen bond interactions.

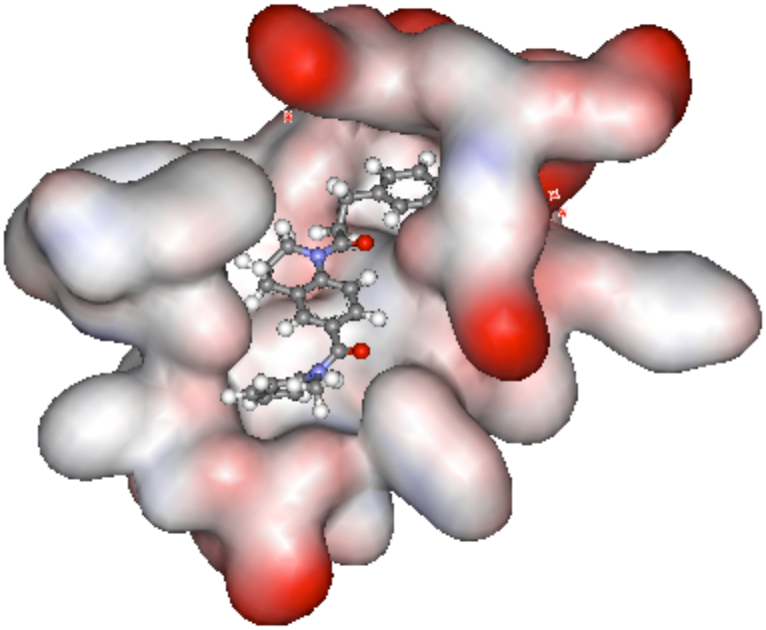
The second software package that I evaluated was FRED. As can be seen from the same ROC graph above, the default settings for FRED fared worse in its ability to separate out the thrombin inhibitors by score. There were more false positives in the initial 30 sequences (a total of 5) and then the slope of the line would indicate that the success rate for the remaining sequences was roughly equivalent to 50/50 random guessing. Since there were only 84 sequences fed from Omega into FRED (see Materials and Methods section), one can also see that several of the positive examples sorted near the bottom of the scores in rank order. Since this was the first time that I had used the FRED software, I suspected that at least some component of its inferior performance was due to my inexperience. To explore this, I wanted to examine some of the docking results for compounds that scored well in all of the methods. This idea of consensus scoring has been used effectively in the literature to increase the effectiveness of docking (e.g. Clark, 2002). The compounds I looked at in detail were designated number 60 and 89 in my test set. Compound 89 had fared well in both the FlexX and FRED runs, receiving the highest or next to highest score in both cases. Compound 60, by contrast, had scored in the top 5 using FRED but outside of the top 5 for both of the FlexX scoring functions. The docking results for these two compounds and the original reference structure isolated from the PDB file are shown below using WebLab ViewerLite (Molecular Simulations Inc.).



Reference compound from 1DWD PDB file isolated in an active site box.

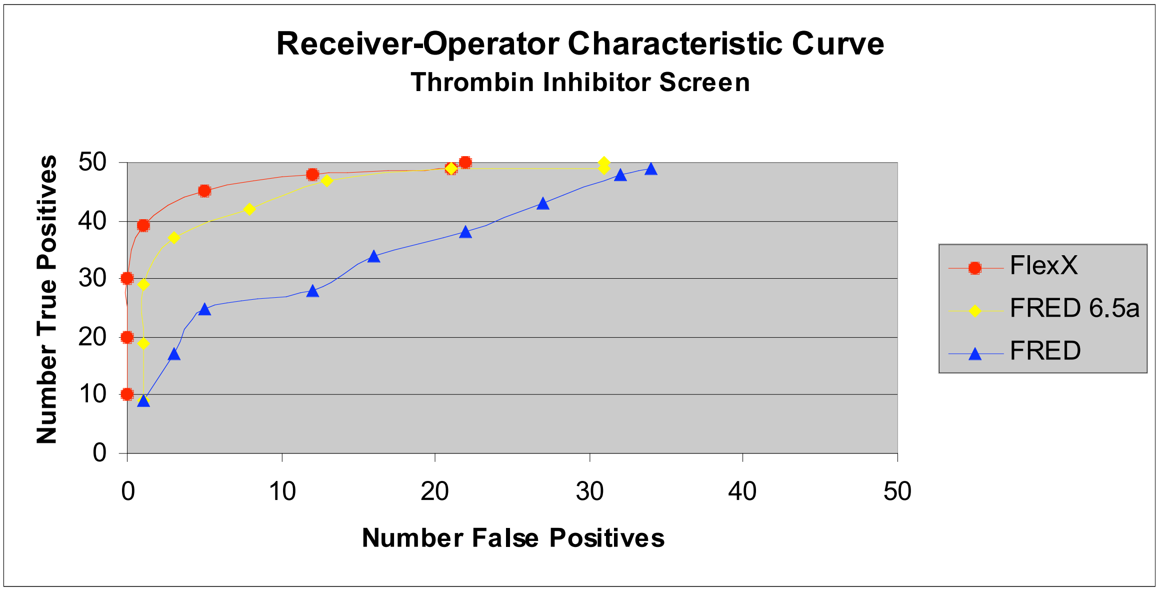


Test compound 60 docking result in active site box.



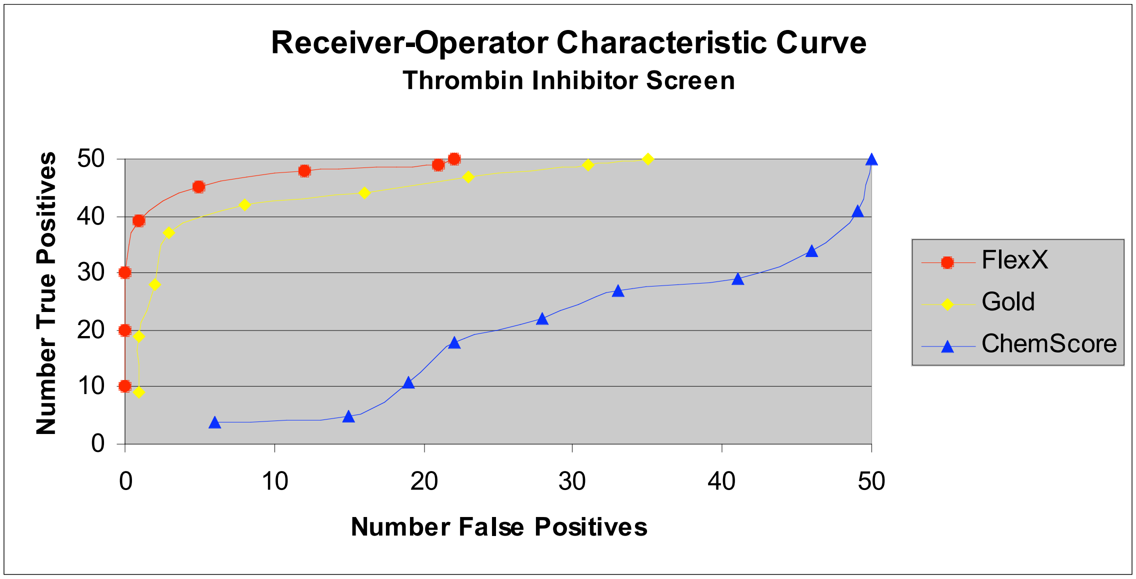
Test compound 89 docking result in active site box.

These views help illustrate a couple of key points: 1) the docking arrangement of compound 89 looks quite similar to that of the original reference compound and would seem to justify its high score from the docking algorithms; 2) the docking arrangement of compound 60 looks less like the original reference compound - in fact it looks a little suspect that the southern most cavity is empty and a portion of the compound is pointing north outside of the active site into an area that may be occupied by other portions of the target protein. Further investigation of FRED's default parameters showed that while I was passing in the same representation of the active site box, it was only using a 5 angstrom diameter in its own calculation of the active site. Setting an option to increase this to 6.5 angstroms (the same as the default for FlexX), produced the following ROC curve:



The blue line above shows the initial FRED results and the yellow line is the improvement achieved by increasing the size of the active site to accurately reflect collisions that would occur between compounds and the protein. While the results are not as strong as the FlexX results shown in red, there is still a marked improvement. There is only a single false positive among the top 30 ranked scores, and a cut-off score of –18200, for example, would achieve a precision of 91% while correctly partitioning 80% of the true actives. In practice, after some further refinements by someone with a better Chemistry background than myself, I might be tempted to use FRED over even the FlexX results due to its efficiency. FRED running on a cluster of Linux boxes took literally only seconds to run while Sybyl/FRED running on an SGI server took several hours. An interesting application of virtual screening where this time difference might really pay dividends is in screening virtual libraries. In this application, not only is the target docking done via computer, but the compounds that are being docked are being generated virtually on the fly by combining potential pharmacophores onto scaffolds of interest (Schneider, 2002). Setting the scoring threshold very high to limit as many false positives as possible might give you the potential to pull out some interesting compounds for synthesis from literally millions of simulated compounds. These thoughts on FRED's efficiency and reasonable results are also supported by evaluations in the literature. Schulz-Gasch (2003) suggests that FRED is a good alternative for general use in virtual screening because of its speed and notes that it works particularly well when the binding mode of the ligand is determined by the overall shape of the binding pocket rather than hydrogen bonding. She notes the estrogen receptor and thrombin to a lesser degree as falling into this category.

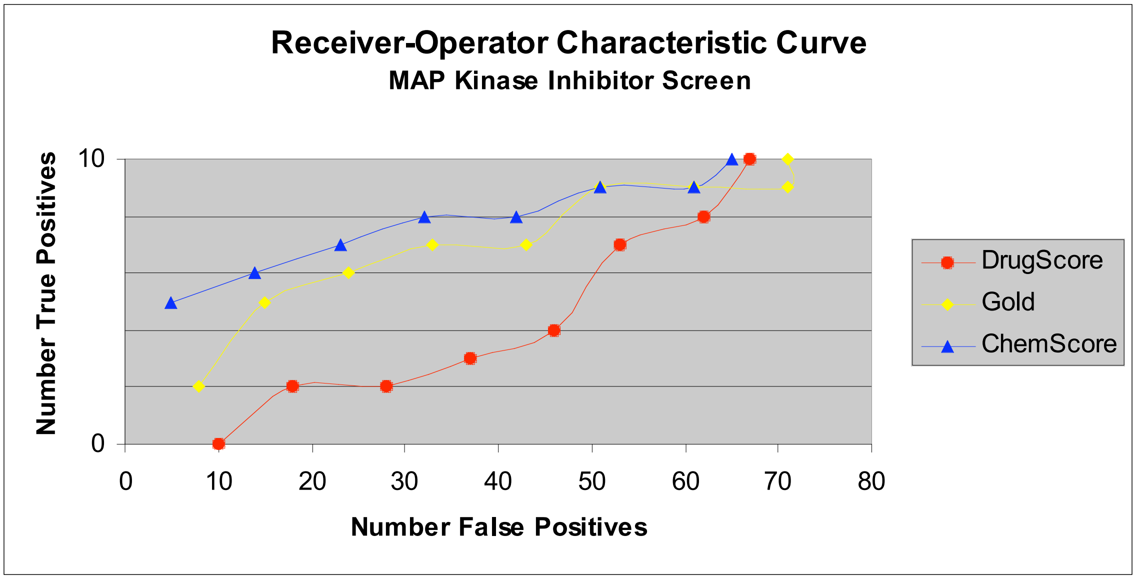
The final docking software that I evaluated was GOLD, which uses a genetic algorithm based search mechanism to examine potential docking combinations. Results using both the default GOLDScore scoring function and ChemScore for the thrombin inhibitor example are compared to FlexX below:



The yellow line representing the default GOLDScore function displays decent results. There are only two false positives among the 30 top scores and choosing a cutoff score of 51 yields a precision of 0.93 while recalling 82% of the positive examples. FlexX would clearly be the algorithm of choice to use here, given also that GOLD took almost 24 hours to run. This performance was running on a single Linux processor, and there are ways to speed up the performance, e.g. running on a cluster of Linux processors or reducing the number of cross-over and mutation operations that ran within each generation of the genetic algorithm, i.e. reducing the search space. Neither of these operations, however, is likely to give the GOLD algorithm an advantage over FlexX. My choice of using FlexX over GOLD is not supported by the literature. Bissantz's (2000) evaluation of multiple docking and scoring function combinations found GOLD to have the best docking accuracy, and Kontoyianni (2004) also found GOLD to behave better overall than FlexX and its other competitors. It is very interesting to note, however, that both authors mention that the best docking program/scoring function combination is highly dependent on the target in question. While their evaluations did show GOLD to be the best overall choice across a broad array of protein targets, FlexX did perform favorably for thrombin due to its aptitude for handling hydrogen bonding (Stahl, 2001).

Using the ChemScore function with GOLD produced surprisingly poor results. One would be better off randomly guessing whether each compound was a thrombin inhibitor rather than using this model. I don't have a firm explanation for why the performance was so poor, but if you look at the components of the ChemScore scoring function, my suspicion is that it has something to do with the collision term. Since GOLD uses an all atom model where hydrogens are explicitly required, perhaps our binding site is slightly off and ChemScore is simply less forgiving than the other studied scoring functions; this possibility is indirectly studied in the final example.

While I was happy from the outset with the precision that the docking algorithms were displaying for the thrombin inhibitor example, the other two examples that I tried were less turnkey. Stahl mentions that the PDB structure 1ERR that he chose to work with for the estrogen receptor was selected because of its open confirmation that could accommodate both agonists and antagonists. This "open confirmation" maybe should have served as a warning to me that docking attempts might act promiscuously, for I had little luck in my initial attempt with FlexX at achieving anything other than what appeared to be no better than random behavior. Since the estrogen receptor forms a fairly large dimer structure, I chose to mark that one up to experience and move on to a MAP Kinase example. My initial attempt at MAP Kinase p38 was as equally unimpressive as the estrogen receptor example. The DrugScore scoring function yielded the best results with FlexX, but even that (as can be seen by the red line in the ROC curve below) approximated the 1/10 guessing that one would expect by random chance.



The negative result with ChemScore at the end of the thrombin inhibitor study actually led me to wonder if the issue might fall largely back to not having precisely specified the active site. While FlexX and FRED required an active site box to be fed into the software, GOLD offered a few additional ways to define the active site. A SwissProt (Boeckmann, 2003) search for the p38 MAP Kinase annotation identified the active site as ASP residue 168. Specifying this residue as the center of the active site produced the above improved GOLD and ChemScore results. While still not as clear cut as the thrombin examples, the improvement to 5 true positives in the top 9 scores (vs. a random expectation of 1) demonstrates that even a little biological information can increase the effectiveness of the model. The literature findings on ChemScore are generally luke warm. Verdonk (2003) notes that its speed is significantly faster than the default GOLD scoring function, but finds the end results to be quite similar. Schulz-Gasch (2003) also focuses mainly on its speed, and Bissantz (2000) finds it to be poor performer in the very type of classification/ranking problem that I am evaluating.

Conclusions

This brief foray into exploring virtual screening via docking has demonstrated its potential, but it has also exposed some pitfalls that illustrate the complex nature of the challenge. While several of the programs achieved a high level of both precision and sensitivity in identifying known inhibitors to the thrombin target, the reality of the drug screening environment would require scanning a much larger database to obtain even a few promising hits. Given a much larger number of sequences to dock, one most likely has to make some tradeoffs in quality to increase the quantity of docking that can be done, and one would expect the number of false positives to increase, probably significantly. Struggling to get even a classification level problem working across several programs and different protein targets illustrates that a sophisticated level of both Chemistry and Biology insight are required to function successfully in a real virtual screening environment and that these algorithms are by no means “plug-and-play”. To quote Gohlke (2000) directly, “Definition of an appropriate reference state and accounting for inaccuracies inherently present in experimental data is required to achieve good predictive power.” Probably the biggest take away from this project for me is that no one algorithm or scoring function is currently the best for modeling all protein-ligand interactions. Any debate over whether empirical or knowledge-based scoring functions are best seems futile since, as pointed out by Schulz-Gasch (2003), Kontoyianni (2004), and several others, the success of a particular program/scoring function combination is highly dependent on the nature of the target protein in question, and performing preliminary evaluation test runs to pick your best option would be highly recommended. And finally, given the large literature base studying what characterizes a drug-like molecule, performing virtual screening in isolation from this information and any other additional filters would seem unwise.

References

Books

Biochemistry. Berg, Jeremy M.; Tymoczko, John L.; and Stryer, Lubert. New York: W. H. Freeman and Co.; 2002.

Journals

Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD (1998). Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins*, 33, 367-382.

Bissantz C, Folkers G, Rognan D (2000). Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.*, 43, 4759-4767.

Bh̄n HJ (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Design*, 8, 243-256.

Clark RD, Stizhev A, Leonard JM, Blake JF, Matthew JB (2002). Consensus scoring for ligand/protein interactions. *Journal of Molecular Graphics and Modelling*, 20, 281-295.

Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M (2004). Lessons in Molecular Recognition: The Effects of Ligand and Protein Flexibility on Molecular Docking Accuracy. *J. Med. Chem.*, 47, 45-55.

Gohlke H, Hendlich M, Klebe G (2000). Knowledge-based Scoring Function to Predict Protein-Ligand Interactioins. *J. Mol. Biol.*, 295, 337-356.

Halperin I, Ma B, Wolfson H, Nussinov R (2002). Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. *Proteins*, 47, 409-443.

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997). Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.*, 267, 727-748.

Kontoyianni M, McClellan LM, Sokol GS (2004). Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.*, 47, 558-565.

Lipinski CA (2000). Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods*, 44, 235-249.

Lyne PD (2002). Structure-based virtual screening: and overview. *Drug Discovery Today*, 7, 1047-1055.

Rarey M, Kramer B, Lengauer T, Klebe G (1996). A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261, 470-489.

Schneider G, Bh̄n HJ (2002). Virtual screening and fast automated docking methods. *Drug Discovery Today*, 7, 64-70.

Schulz-Gasch T, Stahl M (2003). Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.*, 9, 47-57.

Singh AP (1998). Protein Docking. Archived guest lecture for Biochemistry 218, Stanford University.

Stahl M, Rarey M (2001). Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.*, 44, 1035-1042.

Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD (2002). Molecular Properties that Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.*, 45, 2615-2623.

Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003). Improved Protein-Ligand Docking Using GOLD. *Proteins*, 52, 609-623.

Weininger D (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28, 31-36.

Software and Databases

Concord, Tripos Incorporated.

FlexX, Tripos Incorporated.

FRED, OpenEye Scientific Software.

GOLD, Tripos Incorporated.

OMEGA, OpenEye Scientific Software.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank. *Nucleic Acids Research*, 28, 235-242.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*earch, 31, 365-370.

Sybyl, Tripos Incorporated.

WebLab ViewerLite, Molecular Simulations Inc.