Investigating Computational Methods for Evaluating Putative Substrate Conformations in Cytochrome P450s

Zachary Zappala, ACM # 0271324
Dr. Andrea Salgian and Dr. Leeann Thornton, Mentor
Department of Computer Science, The College of New Jersey
Department of Biology, The College of New Jersey

Problem and Motivation

The motivation for this project is the need to automate the analysis of potential substrate conformations in models that predict the structure of cytochrome P450 enzymes. Specific interest was set on the CYP734A family of cytochrome P450s, which is composed of five members: A1, A2, A4, A5, and A6. The A1 enzyme is found in the plant *Arabidopsis thaliana,* and is known to play a role in the metabolism of brassinosteroid growth hormones – by degrading hormones, the enzyme regulates plant growth. The other enzymes are found in *Oryza satvia*, more commonly known as rice, and are suspected to function similarly to A1. Enzymes work by catalyzing specific chemical reactions; to do this, they must bind to at least one of the molecules involved in the reaction – such molecules are known as substrates. Substrates bind to cytochrome P450s by entering a pocket region called the active site, which is part of the internal structure. Computer predictions of substrate orientation are of great interest in predicting biologically relevant interactions or explaining known reactions. There are current computational approaches that predict how a given substrate molecule will orient within the active site of an enzyme; these approaches generate potential conformations (also known as “docks”). Once these conformations have been generated, analysis is typically carried out by hand; specifically, conformations were being categorized into one of seven different categories representing different general conformations within the active site. This process requires a significant amount of time, effort, and runs a large risk of human inconsistency; this project addresses these concerns by automating analysis of substrate orientation with computational methods.

Background and Related Work

In biological systems, the chemical activities of cells (their “work”) are carried out by hundreds of thousands of micro-machines called proteins (which are further classified into different types, such as enzymes). Enzymes in particular catalyze chemical reactions, reducing the time necessary for them to take place and allowing life to thrive. In order for enzymes to carry out this procedure, they have a specific region on (or within) their three-dimensional surface known as the active site that binds to one or more chemical molecules (the substrates). Each enzyme catalyzes a specific chemical modification of its substrate. Understanding the millions of chemical reactions catalyzed by enzymes is thus a fundamental branch of biological studies that is key to understanding the natural world.

Based on the success of computational methods such as homology modeling [Kirton et. al. 2002] and substrate docking [Kaufmann et. al. 2009], it is thought that these strategies can be applied to predictive substrate docking. Such an algorithm would be able to take potential substrate conformations and evaluate them for various descriptive criteria, such as vertical alignment, horizontal alignment, bond strain, distances been specific atoms, and general positioning within an active site. The need for this program is derived from the complexities of biological computing: discoveries made by computers must be verified experimentally in a wet lab, but it is difficult, time consuming, and expensive to do this arbitrarily [Kemp et. al. 2004]. The Molecular Operating Environment (MOE) can be used to generate potential substrate conformations (Molecular Operating Environment, Chemical Computing Group, Montreal). Previously, our lab had been analyzing these conformations by visual analysis, the process of which is complicated for a number of reasons. For example, visual analysis is a time intensive process, requiring an evaluator to view each conformation individually, process the shape of the substrate molecule (where its different components are), rotate the three-dimensional representation of the molecule, and finally assign a category. For the human eye, potential categories must be descriptive and numerous enough to be useful while also general enough and few enough to be recognized by hand; potential conformations often do not reflect the ideal description of the classification but are nonetheless relevant. Finally, categorizations can vary from evaluator to evaluator and even from sampling to sampling due to subjectivity in human analysis. These limitations all arise from the lack of a mathematical analysis of the potential conformations, as the problem is essentially a geometric classification problem.

Accessory programs such as the one described in this paper overcome the limitations of the human evaluation and give novel uses to data generated with MOE. Such a tool can, amongst other things, be used to help direct wet lab research efforts by providing the tools to predict not only how a substrate will conform but also whether or not it is likely a substrate for the enzyme at all. Similar computational work has been done, but not to the same degree; in this situation, homology models of cytochrome P450 enzymes and their active sites are being tested, whereas most other studies have known at least something for sure about the molecules under investigation [Song et. al. 2007; Schormann et. al. 2008; Bjelic & Aqvist 2004].

Approach and Uniqueness

Until now, no one has automated the process of analyzing potential conformations of substrates docked in a computer model of enzyme structure. Previous analysis by hand had led to categorizations of potential conformations into one of seven different broad categories, as enumerated in Table 1. It was decided that these conformational categories would also be examined by the computational approach as well.

#	Qualitative Description
1	Unrealistic distortion, improbable conformation
2	Substrate is horizontal, with the chain over the I-helix
3	Substrate is horizontal, with the rings over the I-helix
4	Substrate is vertical, with rings over heme and chain up
5	Substrate is vertical, with chain over heme and rings up
6	Substrate is horizontal, with the chain over the B-sheet
7	Substrate is horizontal, with the rings over the B-sheet

Table 1. Substrate Conformational Categories

# refers to the categorical ID assigned to each pose; the qualitative description describes what was looked for when analysis was performed.

In brief, these categories describe a number of potential orientations of the substrate molecules in a frame of reference to a planar heme region (defining the “bottom” of the active site) and bordered on two sides by an alpha helix (the I-helix) and a beta-pleated sheet (the B-sheet), two types of enzyme secondary structures.

To categorize substrate conformations, it was necessary to describe the active site in more mathematical terms; the MOE program produces a set of three-dimensional points that define the active site when anchoring substrates. These points were taken and expanded into a series of overlapping spheres, ultimately creating the globular shaped structure seen in Figure 1 that defines the active site.

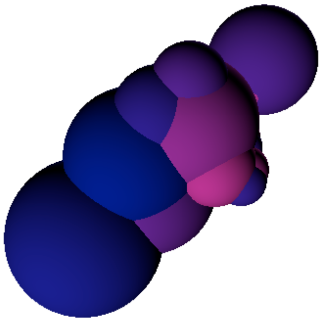

Figure 1. Active site spheres representing the active site for A2 – colors are aesthetic.

With the active site now defined by a series of spheres, all of the spheres were divided into one of four different regions regarding their positioning in relation to the heme; those closest were labeled as "heme" or "bottom" spheres, while those farther away are labeled as "top" spheres. Those in the middle of these two regions were labeled as "middle" spheres and those on the side, "side" spheres. To determine these regions, a best-fit line was determined through the active site points using random sampling and consensus (RANSAC). RANSAC is an algorithm that randomly samples from a provided data set for a certain number of iterations to fit a model and figure out the consensus model [Fischler et. al. 1981]. After this line was fit, the spheres closest to the heme were annotated as "heme" regions, proximity to the line and distance from the heme were used to annotate the rest of the spheres. These annotations are shown in Figure 2.

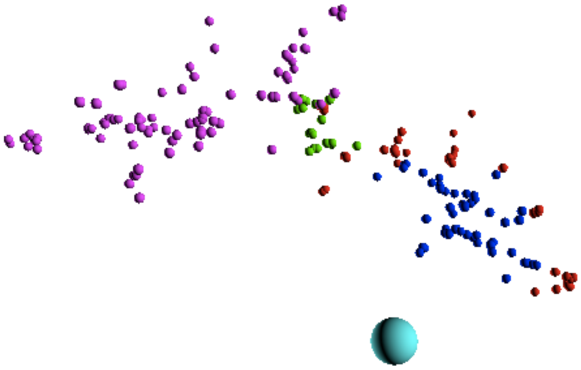

Figure 2. Division of the substrate anchors into categorized regions: cyan sphere is heme atom; magenta spheres represent the top region; green spheres represent the middle region; blue spheres represent the heme region; red spheres represent the side region.

With the active site fully described mathematically, the substrate molecules needed to be annotated as well in order to figure out their orientation relative to the active site. All of the substrate molecules that were investigated belong to the brassinosteroid biosynthetic pathway and all have carbon atoms that belong to either a *ring* region or *chain* region. The substrate data, when exported from MOE, does not contain this data so it was necessary to algorithmically determine it. When importing the substrate molecules, each atom was treated as the vertex of a graph and each chemical bond as an edge, linking the molecule together. Using cyclic detection algorithms, it was possible to determine which carbon atoms belong to the *ring* group and those that belonged to the *chain* group, as illustrated in Figure 3.

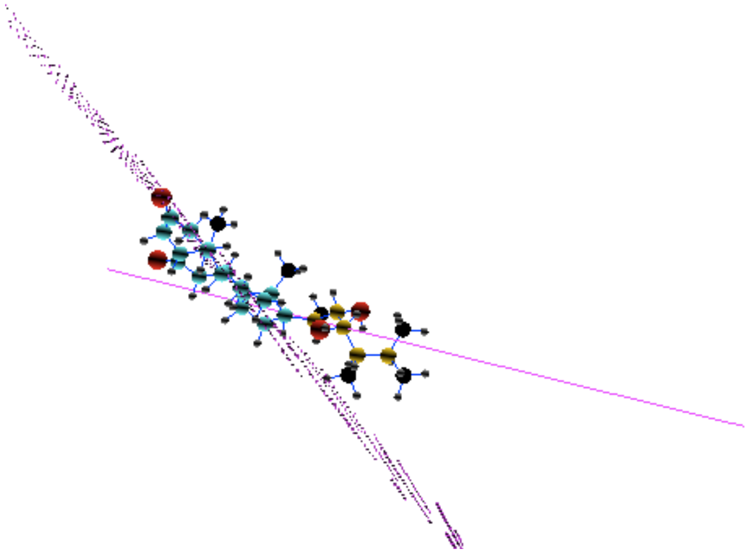

Figure 3. Ring and chain carbons of brassinolide, the last member of the brassinosteroid biosynthesis pathway; *ring* carbons are highlighted in cyan while *chain* carbons are highlighted in yellow. An approximate plane running through the *ring* region and an approximate line running through the *chain* region are shown in purple for reference.

The program then observes the orientation of these two separate groups of carbons in relation to each other and the several distinct regions within the acting site, as well as the general shape of the potential conformation. It determines the percentage of ring or chain carbons in each distinct region of the active site. With this quantitative data in hand, it is capable of assigning each conformation into one of the qualitative categories listed in Table 1. In order to refine the categorizations being made, thresholds were determined experimentally for each carbon group per region percentage that matched up with visually confirmed categorizations.

For a secondary approach, support vector machine (SVM) techniques were employed through the SVM*light* tool [Joachims 2010]. Positive examples of each category were visually determined and fed into the SVM*light* learning module along with negative examples of each category (negative conformations being the positive conformations for other categories). Once SVM models were built, adapter programs that automate classification via the SVM tools were written to synthesize the SVM classifications.

Results and Contributions

Figure 4 shows the classification performance of the two computational approaches when compared to visual analysis, particularly for P450s A1, A4, and A5. The fact that the classifications do not completely agree with visual analysis is acceptable, since the purpose of developing these approaches was to overcome the limitations of visual analysis. In general, the

SVM and threshold hold approach were comparable in classification ability, although the SVM approach performs better when classifying substrate conformations in the A1 enzyme. That the computational methods do approach agreement with visual analysis in A1, A4, and A5 suggests that the computational methods are successfully categorizing these potential substrate models.

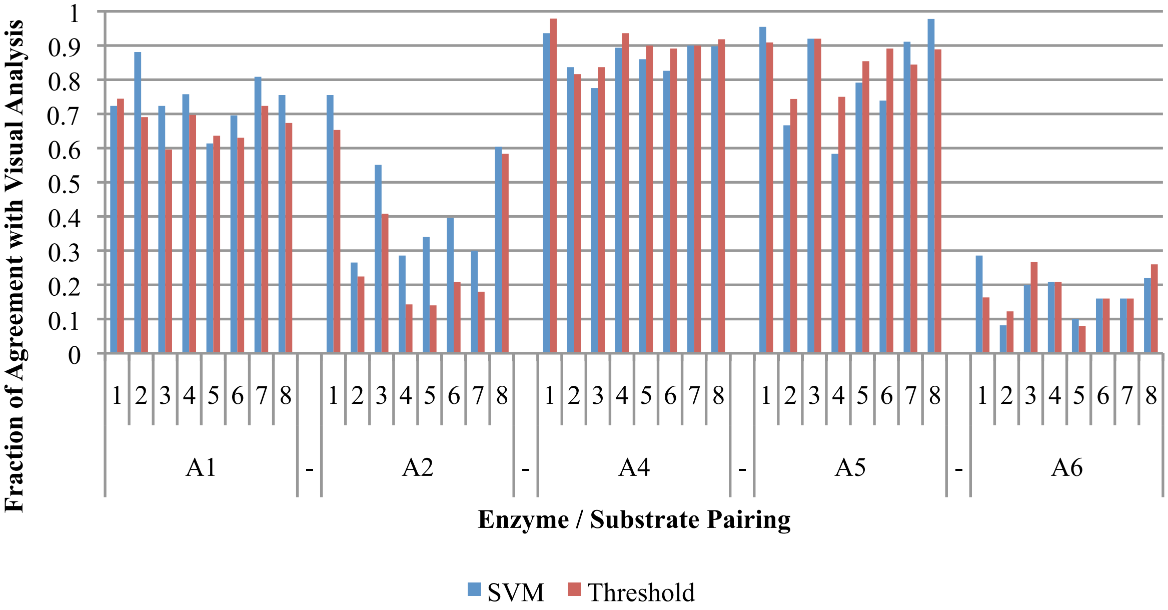

Figure 4. Compares the SVM and threshold approaches side-by-side and shows their agreement with visual analysis. Substrates are: (1) 3-dehydroteasterone; (2) 6-oxocampestanol; (3) brassinolide; (4) campestanol; (5) castasterone; (6) cathasterone; (7) teasterone; and (8) typhasterol.

The failure to correctly classify A2 and A6 is corroborated with difficulties that were experienced when categorizing them visually. There are several potential explanations for this unexpected result; for one, it is possible that the A2 and A6 enzymes are significantly different from the rest of the CYP734A family, precluding the use of the exact same methods. Another possibility is that the homology models being examined are not biologically relevant. To resolve this issue, refining the A2 and A6 models will allow testing of a new batch of substrate docks. Other future work includes improving the SVM model system, since only one set of SVM models have been tested with, and extending these applications to a more general setting. Ideally the program suite would accept any combination of enzyme, active site information, and potential substrate docks to report what archetypal conformations are observed without any human intervention whatsoever.

References

1. Annalora, A., Goodin, D., Hong, W. X., Zhang, Q., Johnson, E., and Stout, C. 2010. Crystal Structure of CYP24A1, a Mitochondrial Cytochrome P450 Involved in Vitamin D Metabolism. Journal of Molecular Biology 396(2): 441-451.
2. Bjelic, S., and Aqvist, J. 2004. Computational prediction of structure, substrate binding mode, mechanism, and rate for a malaria protease with a novel type of active site. Biochemistry 43(46): 14521-14528.

3. Juhl, P., Trodler, P., Tyagi, S., and Pleiss, J. 2009. Modeling substrate specificity and enantioselectivity for lipases and esterases by substrate-imprinted docking. BMC Structural Biology 9(1): 39.
4. Kaufmann, K., Dawson, E., Henry, L., Field, J., Blakely, R., and Meiler, J. 2009. Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies. Proteins 74(3): 630-42.
5. Kemp, C., Flanagan, J., van Eldik, A., Maréchal, J. D., Wolf, C., Roberts, G., et al. 2004. Validation of model of cytochrome P450 2D6: an in silico tool for predicting metabolism and inhibition. Journal of Medicinal Chemistry 47(22): 5340-5346.
6. Kirton, S., Kemp, C., Tomkinson, N., St-Gallay, S., and Sutcliffe, M. 2002. Impact of incorporating the 2C5 crystal structure into comparative models of cytochrome P450 2D6. Proteins 49(2): 216-231.
7. Kortagere, S., Krasowski, M., and Ekins, S. 2009. The importance of discerning shape in molecular pharmacology. Trends in Pharmacological Sciences 30(3): 138-147.
8. Schormann, N., Senkovich, O., Walker, K., Wright, D., Anderson, A., Rosowsky, A., et al. 2008. Structure-based approach to pharmacophore identification, in silico screening, and three-dimensional quantitative structure-activity relationship studies for inhibitors of Trypanosoma cruzi dihydrofolate reductase function. Proteins 73(4): 889-901.
9. Song, L., Song, L., Kalyanaraman, C., Kalyanaraman, C., Fedorov, A., Fedorov, A., et al. 2007. Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nature Chemical Biology 3(8): 486.
10. O'Conner, J. 2006. Scripting for the Java Platform. Sun Developer Network. (http://java.sun.com/developer/technicalArticles/J2SE/Desktop/scripting/)
11. StdDraw Javadoc: (http://www.cs.princeton.edu/introcs/stdlib/javadoc/StdDraw.html)
12. Fischler, M. A., and R. C. Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM 24: 381–395.
13. Joachims, T. 2010. SVM*light*. http://svmlight.joachims.org/.
14. Chemical Computing Group. Molecular Operating Environment (MOE). http://www.chemcomp.com/.