Cytochrome P450 Inhibitor Classification with Statistical Learning

D.Siegle

Abstract

This project compares methods of Cytochrome P450 inhibitor prediction based on compound structures. CYPs are a natural first choice in which to develop *in silico* models because of their central role in drug candidate rejection due to adverse drug-drug interactions. Several non-linear, high-dimensional classification models were built and compared using a large, publicly available high throuhput screening luminescence assay (PubChem AID1851) against five CYP isozymes (1A2, 2C9, 2C19, 2D6 and 3A4). The methods compared are a Bayseian binary QSAR method from the Molecular Operating Environment, and 3 standard machine learning methods implemented in the Python programming language; κ-Nearest Neighbor, Random Forests, and Support Vector Machines. They all performed well, with the methods implemented in freely available software performing as well or better than the one in software that is widely accepted in industry.

Contents

List of Figures

List of Tables

Acknowledgements

I want to thank ...

Chapter 1

Introduction

1.1 Background

The goal of pharmaceutical sciences is to identify and market safe and efficacious drugs. Toxicity is a large contributor to drug candidate attrition in drug development. Inhibition of enzymes in the cytochrome P450 superfamily major source of toxicity in human and animal models because of their role in first pass metabolism. If a compound inhibts cytochrome P450 it is likely to lead to toxic effects when first pass metabolism fails to clear or alters the pharmacokinetics of other therapies. For drug development, it would be useful to know whether a compound of interest inhibits cytochrome P450s as early as possible.

The Cytochrome P450 superfamily is large and varied. Different isozymes metabolize different substrates. Even within and between individuals phar-

macokinetic and pharmacodynamic variability can be high for reasons not entirely characterized. Designing assays for each isozyme and SNP variant and running them against every compound of interest as a general strategy is cost prohibitive. This project was designed to test the predictive power of computational methods for potential cytochrome P450 inhibition based simply on knowledge of chemical structure.

There has been an upward trend in drug development costs for three decades for a variety of reasons. Integrated computational approaches have been proposed as one way to control costs in pharamceutical R&D. [Visser et al., 2014] The foundational skills demonstrated in this thesis are needed to pursue a systems approach to drug development that the industry has recently turned toward as a way to boost R&D productivity.[Visser et al., 2014, Berg, 2014] It is also part of the larger project of *in silico* drug-development, attempting to reduce reliance on exploratory *in vivo* and clinical drug testing and increase the number of effecive treatments for patients and the amount of time to develop them.

There is a tradition of associating compound structure with bioactivity that goes back at least to Hansch[Hansch and Fujita, 1964] and has progressed under the moniker of Quantitative Structure Activity Relationship (QSAR). QSAR started with direct measures of chemical compounds and then derived features and used them to then build expert systems or statistical models that tried to predict biological activity. At the same time the field of machine learning emerged using computation in an analogous and

more general way – to associate features with results, inputs with outputs. Machine learning can be thought of as using algorithms to figure out how to perform important tasks by generalizing from examples. These algorithms are of course usually laborious, which necessitates their execution by computer.

Statistical machine learning builds upon the peer prediction of machine learning. Statistical learning also allows prediction but focuses more on models and methods that can be used by scientists and engineers.[James et al., 2013] Further extension of the statistical learning discipline to the pharmaceutical sciences can lead to important contributions to systems pharmacology. Or rather statitical learning methods are an important precursor to the needs of systems biology and systems pharmacology modeling.

This project compares different techniques for Cytochrome P450 inhibition prediction in the framework of statistical learning.

Reproducibility is one of the main principles of the scientific method. As much as possible, the code and data used in this study is publicly available and source controlled. Reproducibility as a practice is a habit, and a good one to get into. An attempt was made to make the materials and methods for this project reproducible in a completely automated way.

1.2 Review of Literature

1.2.1 Cytochrome P450 Superfamily

One of the largest and most functionally diverse protein superfamilies is the cytochrome P450 family of hemoproteins. From bacteria to humans, the functional breadth of cytochrome P450 activity is far ranging. At the latest count there were significantly more than 2000 identified cytochrome P450 genomic and cDNA sequences that have been divided into a total of 265 different families.[Danielson, 2002] Cytochromes P450 appear in every kingdom from bacteria to higher eukaryotes. Multiple cytochrome P450 genes can be expressed simultaneously as different isozymes and the number of genes per species is highly variable with a tendency for higher eukaryotes to possess more. It is the central role that these ubiquitous proteins play as phase I enzymes in human drug metabolism that makes them so important to the pharmaceutical industry.

The cytochromes P450 (CYPs) constitute the major enzyme family capable of catalyzing the oxidative biotransformation of most drugs and other lipophilic xenobiotics and are therefore of particular relevance for clinical pharmacology. The CYP families are classified based on pairwise amino acid sequence identity among individual members. Families CYP 1-3 are involved in phase I metabolism of human drugs and xenobiotic compounds, whereas other CYP families (CYP 4, 11, 17, 17 and 21) are involved in the metabolism of endogenous compounds such as fatty acids, steroids, eicosanoids, bile acids

and fat soluble vitamins[Singh et al., 2011]

The CYP enzymes that are involved in the oxidative metabolism of drugs play a major part in the activation and elimination of therapeutic drug molecules. CYP inhibition leads to decreased elimination and/or changed metabolic pathways of their substrates, which is the major cause of adverse drug-drug interactions.[Lapins et al., 2013] Adverse side effects of drug-drug interactions is an important consideration, especially during the research phase of drug discovery.[Cheng et al., 2011]

There is the considerable substrate overlap between enzymes of this superfamily. Being broadly specific with respect to their substrates, CYPs are therefore susceptible to inhibition by a large variety of chemical compounds. The results of a large-scale screening against five CYP isoforms identified that the majority of compounds in a typical chemical library cross-inhibited several isoforms, while only a small fraction of the compounds did not inhibit any of the isoforms.[Veith et al., 2009]

1.2.2 Rationale for early compound profiling *in silico*

Drug discovery is a multi-parameter optimization process in which compounds are optimized for interaction with their target while minimizing off-target activities, while at the same time imparting additional drug-like properties on those compounds.[Zlokarnik et al., 2005]

During the last decade, techniques for high-throughput *in vitro* screening of CYP inhibition were developed and implemented on a broad scale in drug

discovery pipelines of pharmaceutical companies and research institutions. Much open data has accumulated from the latter through academic research initiatives (e.g. PubChem Bioassays AID 410 and 1851)

The collected data has enabled development of structure-activity relationship models for *in silico* prediction of CYP inhibition by a much larger pool of researchers than those who designed the assay.

In silico screening has been very appealing as, with the steady increase in computing power, screening cost could in this manner become negligible. The hope is that virtual compounds could be screened for CYP liabilities inorder to realize th savings from reducing the number of candidates of questionable utility that a team would otherwise synthesize.[Zlokarnik et al., 2005]

The ability to predict clinical safety based on chemical structures is becoming an increasingly important part of regulatory decisionmaking. QSAR models are currently used by in industry and by regulators to evaluate safety concerns and possible nonclinical effects of a drug when adequate safety data is absent or inconclusive.[Kruhlak et al., 2012]

QSAR models and their value in informing regulatory decision making will likely increase with the standardization of analytical approaches, more complete and reliable data collection methods, and a better understanding of toxicity mechanisms and the role of disease, and individual susceptibility to adverse clinical events.[Kruhlak et al., 2012]

Many high-throughput technologies are now available to detect P450 inhibitors. High-throughput screening data can be used to guide medicinal

chemists away from these interactions in an early stage. In certain cases it might also identify the inhibition issue by targeted modification of the CYP interacting functionality. To be generally useful, P450 inhibition screens need to be calibrated against standard methods and preferably also tested with a large set of drugs, for which human drug-drug interaction outcome is known. [Zlokarnik et al., 2005] This should decrease the number of withdrawls of novel drugs from the market due to inhibition of major P450 isozymes.

Pubchem Bioassy AID1851 has been the basis for several attempts to advance *in silico* screening development. Recently, Vasanthanathan et al. and Novotarskyi et al. and more, recently developed large-scale single target models for the CYP1A2 isoform, and Cheng created single target QSAR models for all five CYP isoforms.[**?**, **?**, **?**] And Lapins et al. pursured a proteochemometric method for all five isozymes by also taking into account each isozymes protein sequence. [Lapins et al., 2013] These models showed good predictive performances.

The United States is about to become the last country to accept QSAR in the drug approval process. The drafting of International Committee on Harmonization (ICH) M7 guideline can be viewed as setting a precedent for possible future, broader regulatory applications of QSAR modeling. ICH M7, will for the first time specify that – under very specific conditions – the results of QSAR computational toxicology predicitions will be considered sufficient for genotoxic contaminants of pharmaceuticals under consideration and eliminate the need for laboratory testing. [Kruhlak et al., 2012]

1.2.3 Quantitative Structure-Activity Relationship

Quantitative structure-activity realtionship (QSAR) modeling is generally accepted as the construction of predictive models of biological activities as a function of structural and molecular information of a compound or compound library.[Nantasenamat, 2009]

QSAR models describe the correlation between molecular features and activity at a given end point of interest. QSAR models are typically defined as those that use statistical methods to analyze the mathematical correlations between molecular features and activity, while SAR models are those that are constructed by using human expert knowledge (expert rule-based).

QSAR models can make it possible to predict the biological activities of a given compound as a function of its molecular structure. It follows from the 'similarity principle' that new and untested compounds possessing similar molecular features as compounds used in the development of QSAR models are assumed to possess similar activities and properties. Several successful models have been published over the years which encompass a wide span of biological and physicochemical properties.

A QSAR model defines the mathematical relationships between descriptors and biological activities of know molecules. This differs from receptor binding-based efficacy prediction by taking into account binding site characteristics as well as molecular docking analysis. In contrast, these methods attempt to predict drug efficacy based on known mechanisms of action and medicinal chemistry by individually studying molecular interactions between

a drug and targets/receptors. [Kruhlak et al., 2012]

Applied QSAR has typically been used for drug discovery and development and was mainly used to correlate molecular information with not only biological activities but also with other physiochemical properties. The later approach has been termed quantitative structure-property relationship (QSPR). Derived molecular parameters account for hydrophobicity, topology, electronic properties, and steric effects. These characteristics of compounds can either be determined empirically through experimentation or theoretically via computational chemistry as needed. [Nantasenamat, 2009] These parameters derived from compound structure are refered to as molecular descriptors.

Molecular Descriptors

Molecular descriptors can be thought of as the mathematical representation of essential information of a molecule in terms of its own physiochemical properties. Depending on the needs of the analysis, proerties considered can be electronic, geometric, hydrophobic, constitutional, lipophilic, steric, solubility, quantum chemical, or topological. From a practical viewpoint, molecular descriptors are chemical information that is encoded within the molecular structures. [Nantasenamat, 2009]

Molecular features can be either experimentally measured or calculated values. They can be in the form of simple physiochemical properties such as logP or logarithmic acid dissociation constant (pKa), numerical repre-

sentaions of substructure fragments, or purely mathematical. Mathematical descriptors are chemical structural features represented in numerical form, and range from simple atom counts to the product of complex equations describing electron distribution across a molecule.[Kruhlak et al., 2012]

Molecular descriptors as predictors used in QSAR modeling are typically less precise than the lock and key relationships that underpin the docking approach to computer-aided drug design. The basic assumptions in QSAR modeling are that similar molecules exhibit similar biological activity and that the physiochemical properties and/or structural properties of a molecule can be encoded as molecular descriptors to predict the biological activity of structurally related compounds.

1.2.4 Modeling

Models can be described as global or local. Global models incorporate chemicals with a range of molecular features acting across the spectrum of chemical pathways, whereas local models are highly focused on a single chemical class and end point. Although local models generally have much higher accuracy, their narrow domain of applicability renders them impractical in most regulatory environments where predictions need to be made for structural classes covering active pharmaceutical ingredients to metabolites, degradants, reagents, and synthetic intermediates. [Kruhlak et al., 2012]

The construction of QSAR models typically follows two main steps:

- Description of molecular structure and derivation of descriptors
- Multivariate analysis correlating molecular descriptors with observed activities.

Additional intermediate steps that are also crucial for sucessful development of such QSAR models include data preprocessing and statistical evaluation. [Nantasenamat, 2009]

A typical QSAR workflow involves chemical structure management, descriptor calculation, and statistical analyses that are treated as separate steps and often performed by non-integrated software packages. This can lead to low throughput and even the lack of possibility of performing predictions for new compounds and updating the models when new data become available, depending on the workflow.

Figure 1.1: Typical QSAR Workflow

Because of the complex nature of toxicity, safety prediction is considered more challenging than efficacy prediction. Toxicity mechanisms may be unknown or poorly characterized in higher organisms, and similar pathways and targets may be associated with different toxicities and adverse events. Toxicity prediction must also encompass a number of complex interactions and remain alert to the possibility of finding the unexpected. For instance, toxicities could result from on-target effects due to incomplete knowledge or inadequate target validation, or from off-target effects mediated via unknown

molecules and mechanisms, or even from genetic variation or comorbidities in any of the previously mentioned pathways. Nevertheless, QSAR models have been developed with some success by both industry and regulatory agencies.[Kruhlak et al., 2012]

According to Kruhlak, et al., the successful development of a QSAR model for safety prediction requires a sufficient amount of high-quality data, the appropriate selection of descriptors, the availability of one or more suitable statistical or mathematical models and an effective training and validation strategy.

1.2.5 Machine Learning

Machine Learning algorithms figure out how to perform important tasks by generalizing from examples. Classification is a well understood area of machine learning. A classifier is a system of inputs, typically a vector of discrete and/or continuous feature values which outputs a single discrete value, the class.[Domingos, 2012]

Learning = representation + evaluation + optimization

- Representation = A classifier must be represented in some formal language that the computer can handle. A representation for the learner is tantamount to choosing the set of classifiers that it can possibly learn. This set is called the hypothesis space.

- Evaluation = Often called the evaluation function, objective function, or scoring function - distinguishes good from bad classifiers.
- Optimization = a method to search among classifiers for the highest scoring one.

Most textbooks are organized by representation, but the other components – evaluation and optimization – are equally important.[Domingos, 2012]

The current state of machine learning is fundamentally a subset of optimization and has found its biggest successes in fields where there are far more variables than parameters. The ultimate goal of machine learning in this case is to generalize beyond examples in the training set, because no matter how much data we have, it is unlikely that we will see those exact examples again at test time.[Domingos, 2012]

1.2.6 Statisitcal Learning

A concise definition of statistics is as the applied science that constructs and studies techniques for data analysis (Jan de Leeuw) Statistical learning refers to a set of approaches for estimating a function that describes a dataset as a precursor for prediction or inference.[James et al., 2013] Statistical machine learning constructs that function by generalizing from examples, i.e. data.

Leo Breiman wrote a landmark paper that documented the beginnings of this approach. He said 'There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are

genereated by a given stochastic data model. The other uses algorithmic models and treats the data mechanisms as unknown' [Breiman, 2001] He also claims that the statistical community had traditionally prefered the first view.

This is a wider view than traditionally encountered in most applied science education where simple hypothesis testing predominates. Opening up the field of data analysis like this brings opportunity for exploration along with new concerns, such as the 'curse of dimensionality', 'degrees of freedom of the analyst', 'black box algorithms', difficulty of bias estimation and the 'no free lunch theorem'.[Boulesteix and Schmid, 2014]

In the context of QSAR then, with enough prior knowledge, machine learning may be a practical approach to fill in the gaps of clinical knowledge for any relevant CYP isozyme queried against any known compound of interest so long as the structure is known.

1.2.7 Reproducibility

Science conducted in an open fashion confers the following benefits

- Reproducibility of experiments allows other researchers to use the exact methods to calculate the relations between biological data.
- Faster development of disease models and therapeutic treatments due to the reuse of existing knowledge. Projects can be built upon existing results more easily or extend the research in directions unanticipated

by the original team. First-pass results can be subject to new analysis, and a second look at compounds with interesting side effects can lead to serendipidous discoverys

- Increased quality as a result of having more researchers studying the same topic to provide a layer of assurance that errors will not propagate.

- Long-term availability of data and code. If these resources are not tied to businesses or patents, then they can be posted to multiple repositories to ensure that they are available in the future. [Prli and Lapp, 2012]

High profile case of inability to replicate the effects of major cancer drug findings. [] We don't know which drugs because those findings are not public either.

Information generation can happen far faster and is much more common than data analysis and knowledge creation in the biological sciences.

There are organizational issues to overcome. More biologists trained in quantitative and statistical methods for analyzing large data sets are needed.

Large -omics data collections can be expensive to generate and also to manage. Efforts are often distributed and decentralized, leading to duplications of effort and lack of common standards. For pharmaceutical drug discovery researchers, the long-term commitment and infrastructure required to support these approaches is not compatible with current restructuring trends in the industry.[Berg, 2014]

New research findings, supporting data and methods should, therefore, be made publicly available for independent verification and replication in order not to delay medical advances.

1.3 Aims

The aim of this project is to quantify the risk of off target effects for the candidate molecules after building statistical machine learning models that make binary classification of compounds based on their structure, as inhibitors or non-inhibitors of Cytochrome P450.

This project also compares a well accepted, commercial method of binary classification with a variety of open source implementaions of common machine learning algorithms acording to the following plan.

- Build Binary QSAR models in the Molecular Operating Environment (MOE).
- Develop and implement comparable methods in open source software.
- Evaluate and compare results from all models.
- Perform this analysis as reproducible research.

Chapter 2

Experimental Details

2.1 Review of PubChem Assay 1851

PubChem BioAssay 1851 contains data for inhibition of five major CYP isoforms (CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4) by 17,143 chemical compounds. [Veith et al., 2009]

The assay used low low-fluorescence substrates, which are converted to more-fluorescent metabolites. The progression of the reaction is measured by an increase in fluorescence intensity during CYP metabolism of the substrate. Inhibitors of a particular CYP reduce the rate of metabolism of the substrate and this results in a decreased fluorescence signal. [Zlokarnik et al., 2005]

The most recent technology developed for CYP inhibition is based on substrates that release luciferin as the metabolite. This is a coupled assay system in which addition of luciferase and ATP converts the freed luciferin to

des-carboxyluciferin with light emission. The format is similar to fluoresence methods, requiring addition-only manipulations and employing luminescence plate readers. [Zlokarnik et al., 2005]

This assay required recombinant sources of CYP enzymes, because the probes are not sufficiently P450 isoform-selective to be used with those derived from human liver microsomes. Another potential source for false positives in these assays can be compounds that interfere with light generation directly, such as compounds that interfere with luciferase enzymatic activity. [Zlokarnik et al., 2005]

In the dataset, compounds are classified as active or inactive inhibitors for each CYP with an activity cutoff set to AC50 = 10uM (AC50, activity concentration 50, refers to the concentration that is required to elicit half-maximal effect). However, in cases where the dose-response curve for a compound showed poor fit or the inhibition efficacy was below 60%, the assay results were regarded as inconclusive.[Lapins et al., 2013]

Compounds were characterized by their Activity Score and regarded as inhibitors if their activity score ranged between 40 and 100. PubChem Acitivity Score is assigned based on an AC50 values, which was combined with a confidence measure. Combining measures for completeness of a dose-response curve and efficacy of inhibition, resulted in the Activity Score, where a larger value indicates higher inhibitory activity and/or higher confidence in inhibitory assay result. Compounds with an activity score equal to zero are considered as non-inhibitors while compounds with activity scores above 0

and up to 40 are considered inconclusive. [Lapins et al., 2013]

The the tested compounds were all drugs or drug-like compounds. The chemical space revealed that the majority of compounds had molecular weight below 500 daltons and logP below 5. [Lapins et al., 2013]

All compounds were obtained as SMILES strings.

2.2 Feature generation and molecular descriptors

Data preprocessing - The preliminary steps in data preprocessing typically requires data cleaning as raw data often contain anomalies, errors, or inconsistencies such as missing data, incomplete data, and invalid character values which may cause trouble in data analysis if left untreated. It is more complicated when data are collated from many formats requiring harmonization and elimination of redundancies.[Nantasenamat, 2009]

There exists a great deal of variability in the range and distribution of each variable in the data set. However, this may pose a problem for algorithms using distance measurements in the learning step. These situations are handled by applying statistical techniques such as min-max normalization or z-score standardiation.

In min-max normalization, the minimum and maximum value of each variable is adjusted to a uniform range between 0 and 1 according to the following equation: xxxxxx In z-score standardization, essentially the variable of

interest is subjected to statistical operation to achieve mean center and unit variance according to the following formula: xxxxxx [Nantasenamat, 2009]

In situations where the data does not have a Gaussian (normal) distribution, simple mathematical functions can be applied to achieve normality or symmetry in the data distribution. A commonly used approach is to apply logarithmic transformation on the the variable of interest in order to achieve distribution approaching normality. This is typically performed on dependent variables such as the modeled biological/chemical properties of interest whereby IC50 may be transformed to logIC50 or -logIC50. Practically, such mathematical operation is applied to each individual value of a given variable of interest. [Nantasenamat, 2009]

Feature or variable selection – Typically data sets often contain redundant or noisy variables which make it more difficult for learning algorithms to discern meaningful patterns from the input data set. ... Such multicollinearity of the variables ... treated ... in order to reduce unnecessary computational resources that are required in model construction. [Nantasenamat, 2009]

Statistical evaluation In QSAR modeling it is essential to validate the model as well as apply statistical parameters to evaluate its predictive performance. The predictive performance of a data set can be assessed by dividing it into a training set and a testing set. The training set is used for constructing a predictive model whose predictive performance is evaluated on the testing set. Internal performance is typically assessed from the predictive performance of the training set while external performance can be assessed

from the predictive performance of the independent test set that is unknown to the training model. A commonly used approachfor internal validation is known as N-fold cross-validation where a data set is partitioned into N number of folds. For example, in a 10-fold cross-validation 1 fold is left out as a testing set while the remaining 9 folds are used as the training set for model construction and then validated with that fold left out. In situations where the number of samples is limited, leave-one-out cross-validation is the preferred approach. Analogously, the number of folds is equal to the number of samples present in the data set. [Nantasenamat, 2009]

The standard measures of performance used for QSAR models are sensitivity, specificity, positive predictivity, negative predictivity, and concordance. Additional measures such as the Matthews coefficient can be used as a single metric for comparing one model to another while correcting for bias. Because most stat models can be modified to have greater specificity vs sensitivity (or vice versa) by adjusting the training set or applying predictive filters, the Matthews coefficient can be used to determine when the modification is so extreme that it leads to overall degradation of the models performance.[Kruhlak et al., 2012]

Statistical parameters - Pearson's correlation coefficient (r) is a commonly used parameter to describe the degree of association between two variables of interest. Calculated r values have values between -1 and +1 which indicate direct (positive) and indirect (negative) correlation. For describing the relative predictive performance of a QSAR model, r is used to measure

the correlation between experimental (x) and predicted (y) values of interest in order to observe the variability that exists between variables. This is calculated according to the following equation: ...

Root mean squared error (RMS) is another commonly used parameter for assessing the relative error of the QSAR model. RMS is computed according to the following formula: ...

Degrees of freedom take into consideration the number of compounds and the number of independent variables that are present in the data set. This can be calculated using the equation n - k - 1 where n = # of compounds k = # of descriptors. The higher the value, the more reliable the QSAR model is. [Nantasenamat, 2009]

Mathematically speaking, an outlier is essentially a data point which has a high standard residual in absolute value when compared to other samples in the data set. A commonly used approach for detecting outliers is performed by calculating the standard residuals of all compounds in the data set of a QSAR model. [Nantasenamat, 2009]

2.3 Data analysis consideration

Balance number of positives and negatives

Split dataset into training and test set before looking at it.

Cross-validation

Molecular Operating Environment

SciPy ecosystem

Figure 2.1: Supervised Learning Scheme for Clasification

Modeling in Python

The Python programming language is a dynamically-typed, object-oriented interpreted language. Its primary strength lies in the ease with which it allows a programmer to rapidly prototype a project, although it also has a powerful and mature set of standard libraries that can facilitate large-scale production-level software engineering projects as well. Python has a very shallow learning curve and excellent online learning resources.

Methods of Correlation of Compound and CYP Descriptors to Activity Data

κ-Nearest Neighbor (κNN

The kNN algorithm predicts the class of a test set object based on the class membership of its κ most similar training set objects. [Lapins et al., 2013]

Random Forest (RF)

Random Forest is a classifier that consists of multiple decision trees. A decision tree is made of nodes and branches. At each node the dataset is split

based on the value of some attribute that is selected so that the instances of different classes are predominately moved to different branches.

Classification starts at the root node and is performed by passing the instances along the tree to leaf nodes. [Lapins et al., 2013]

To introduce diversity between the trees of a random forest, a small subset of all attributes is randomly selected to take decisions at each node of each tree. The classification decision is performed by considering results of all trees by majority vote. The optimal size of the forest and the number of attributes to consider at each node were found by performing five-fold cross-validation. [Lapins et al., 2013]

Support Vector Machines (SVM)

SVM is a machine learning technique for classification or regression that uses linear or non-linear kernel-functions to project the data into a high-dimensional feature space. [Lapins et al., 2013]

Correlation is performed in this hyperspace based on the structural risk minimization principle i.e., aiming to increase the generalization ability of a model. [Lapins et al., 2013]

They applied the commonly used Gaussian radial basis function kernel optimal gamma (width of the kernel function) and error penalty parameter C were found after performing grid search on fivefold cross validation. [Lapins et al., 2013]

Binary QSAR in MOE

Chapter 3

Results: Classification

3.1 Methods

Dataset preparation

The data from Pubchem Assay 1851 is available for download from the NIH website. The interface changes from time to time, but I downloaded two files which comprised the entire dataset for the experiment. A structures file contained the structural information encoded in SMILES format for each tested compound with corresponding structure ID and compound ID as assigned by whoever. Another file, also organized by SID and CID, contained all of the luminescent responses from the high-throughput screen and the fitted parameters that are summarized by an activity score.

Both files were downloaded as comma separated value (.csv) files and merged together based on the SID column using functions from Python's

pandas library.

First the merged file was loaded into a database in the Molecular Operating Environment. MOE functionality was used to obtain the washed configuration of compounds by removing the salts and finding an energy-minimized conformation. MOE was then used to calculate descriptors based on the molecular structures. The entire suite of MOE 2-D descriptors was selected for descriptor generation, resulting in 186 additional columns of nominal, ordinal and continuous vales appended to the database. The resulting master file was saved in .csv format.

The dataset was then split by isozyme into 5 separate files that each contained only the SID, the Activity Score, and all 186 of the MOE 2-D descriptors using a script in the Python programming language.

For each isozyme, the number of active inhibitors was far outnumbered by the number of inactive inhibitors. The datafile for each isozyme was subjected to a script that separated the inactives from actives. Then the order of the inactives was randomly shuffled and the column trimmed to the length of the activity column, thereby balancing the number of inactives and actives as required by some of the statistical methods.

Next a script randomly shuffled the balanced datasets and split them into a training set and a test set. The training set held 80% of the original values and the test set held 20% of the original values. The split was not based on activity score. The ratio of active/inactive in each split was inspected to check if they were still acceptably balanced.

For each use of randomness in computation, a seed number was set for the random number generator to ensure reproducible results.

The balanced and split datasets are saved to Figshare.com for permanent, free and open access. DOI All subsequent analyses use these same splits for comparability.

Figure 3.1: General Data Analysis Workflow

Modeling in the Molecular Operating Environment

PLS regression is a qunatitavtive method available in the Molecular Operating Environment that has shown poor predictive accuracy with this dataset as demonstrated by previous attempts in Dr. Zheng's lab.

The Binary QSAR approach outlined by Labute [**?**] and included in the MOE version, takes a Bayesian and probabilistic approach to classification of activity based on a reduction of the total number of descriptors to principle components.

To carry out this analysis, first load the training data into MOE and used a menu driven interface to initiate the Binary QSAR methods. A threshold value of 39 was selected; all activity values 40 and above will be considered active and all values 39 and below are considered inactive. The smoothing parameter was left at the default value of 0.25. MOE automatically performs principle component analysis on high-dimensional datasets.

For each isozyme a number of models were built, each using a different

number of principal components. Each principal component is orthogonal and uncorrelated to the rest, and each one captures a portion of the total variance inherent in the dataset. The assumption is that inclusion of more principle components leads to more of the variance being accounted for in a classification decision. However, since each principle component is a linear combination of all the variables, the benefits of dimensionality reduction comes at the cost of interpretability. For comparison, models with 2, 5, 10, 15, 20, 30, and 44 principle components were constructed.

MOE models were written to .fit files and the model report saved as a .txt file.

The test sets were loaded and the washed structures were appended to the .csv files. All models were evaluated using the menu driven workflow in MOE and the classification probabilities were appended to the database file and saved as a .csv.

The resulting file was loaded into a MS Excel spreadsheet and the predictions classified as actives or inactives. Predicted probabilities of ≥ 0.5 were evaluated as active inhibitor predictions. Confusion matrices were then tabulated for predicted vs actual actives and inactives within the spreadsheet. And from the confusion matrices accuracy scores were calculated - total accuracy, specificity, and accuracy of inactive predictions. These results were saved and the results reported below.

Results from MOE binary classification						
PCs		2c19	2c9	2d6	1a2	3a4
2	train	0.585	0.621	0.589	0.638	0.644
2	test	0.593	0.609	0.590	0.632	0.627
5	train	0.685	0.683	0.670	0.704	0.67
5	test	0.683	0.675	0.667	0.705	0.656
10	train	0.699	0.699	0.685	0.734	0.698
10	test	0.691	0.682	0.662	0.746	0.677
15	train	0.700	0.702	0.686	0.737	0.701
15	test	0.687	0.686	0.665	0.752	0.680
20	train	0.705	0.710	0.704	0.741	0.705
20	test	0.699	0.685	0.683	0.748	0.686
30	train	0.717	0.711	0.703	0.735	0.761
30	test	0.699	0.690	0.686	0.739	0.686
44	train	0.708	0.712	0.690	0.725	0.698
44pc	test	0.694	0.688	0.669	0.720	0.676

Table 3.1: Results from MOE binary classification

Modeling in Python

Toy Problem - κNN on first 2 PCs

I don't know if I want to include this section. It is a gentle intro ending with a fairly easy to grasp visualization before the dimensionality explodes. So I'll probably include it.

After downloading the 2c19 dataset, the training data is loaded into memory. A script was written to perform a PCA on the training data, then select the first two principle components and construct a classification model using the κ-nearest neighbor algorithm. The major benefit of this exercise was the ease of visualization of results. Subsequently the two principle components used in model construction are plotted and overlayed with the model

predictions.

The steps in this process script are as follows; the training data is loaded into a dataframe and Activity Score is identified as the response variables. All 186 molecular descriptors are selected as the input variables and then individually scaled to a mean of 0 and standard deviation of 1. PCA is performed on the scaled and normed descriptors. The first two PCs are used to train a κ-NN model that treats any ActivityScore over 40 as an active. The κ-NN function automatically performs a five-fold cross-validation of the model and creates a model object, which can be used to evaluate similarly formated data, such as the test set.

Next the test set is loaded into a dataframe, the response variable identified, the descriptors scaled and normed, and subsequently evaluated by the model object. Accuracy in this case is calculated in the script, as well as a confusion matrix. These results are then plotted in the Ipython notebook with the principle components as axes, thpredicted values from the model are calculated for every point and overlayed with the model results.

κ-NN utilizing on the full seet of 2-D descriptors

The full models follow a similar procedure to the previous workflow, except they omit data reduction by principle component analysis and use the full descriptor set to consruct and evaluate models.

For each isozyme separately, training data is loaded into a dataframe. The Activity Score is assigned the role of response variable and all 186 are

identified as predictor variables. Predictor variables are scaled and normed to mean 0 and standard deviation of 1. kNN.fit method is called from the scikit-lean library to train a model, and then reports a the confusion matrix and an accuracy score for the training model by five-fold cross-validation.

Random Forest classification utilizing the full set of 2-D descriptors

For each isozyme separately, training data is loaded into a dataframe. The Activity Score is assigned the role of response variable and all 186 are identified as predictor variables. Predictor variables are scaled and normed to mean 0 and standard deviation of 1. RF.fit method is called from the scikit-lean library to train a model, and then reports a the confusion matrix and an accuracy score for the training model by five-fold cross-validation.

Support Vector Machine classification utilizing the full set of 2-D descriptors

For each isozyme separately, training data is loaded into a dataframe. The Activity Score is assigned the role of response variable and all 186 are identified as predictor variables. Predictor variables are scaled and normed to mean 0 and standard deviation of 1. SVM.fit method is called from the scikit-lean library to train a model, and then reports a the confusion matrix and an accuracy score for the training model by five-fold cross-validation.

A script was written for each isozyme that performed these three fit methods in series. Code and results are documented in Ipython notebooks that

are currently hosted on github.com and freely accessible and downloadable. Presented in this way, they are easily verifyable and extendable. Experimenting with other classification algorithms in scikit-learn simply requires adding new method.fit calls to the model building loop, because of the simple and consistent design of the scikit-learn library.

Chapter 4

Results: Evaluation

Assessment of the Quality of the Models

The success of any QSAR model depends on accuracy of the input data, selection of appropriate descriptors and statistical tools, and most importantly validation of the developed model. Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose. For QSAR models validation must be mainly for robustness, predicition performances and applicability domain (AD) of the models. [Lapins et al., 2013]

We used two statistical methods: the overall prediction accuracy and the Area Under the Receiver Operating Characteristic (AUROC) curve. We assessed the predictive ability of the models by performing cross-validation and external predictions.

Accuracy

Accuracy is simply the percentage of correctly classified instances and is calculated as

$$ACC = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives or overpredictions, and FN is the number of false negatives or missed predictions.[Lapins et al., 2013]

Accuracy is not an optimal measure of model performance if the data set is unbalanced (i.e. sizes of the classes are unequal) or if certain errors are to be considered more serious than others (e.g. false negatives compared to false positives).[Lapins et al., 2013]

AUROC

In contrast to accuracy, the AUROC is a measure of discriminatory power that is insensitive to changes in class distribution and the costs of making certain errors. A ROC curve is obtained by calculating sensitivity and specificity at various discrimination threshold levels.

Sensitivity is the fraction of true positives among all positively classified instances (the true positive rate) and is calculated as :

$$sensitivity = \frac{TP}{(TP + FN)}$$

Specificity is the true negative rate and is calculated as:

$$specificity = \frac{TN}{(TN + FP)}$$

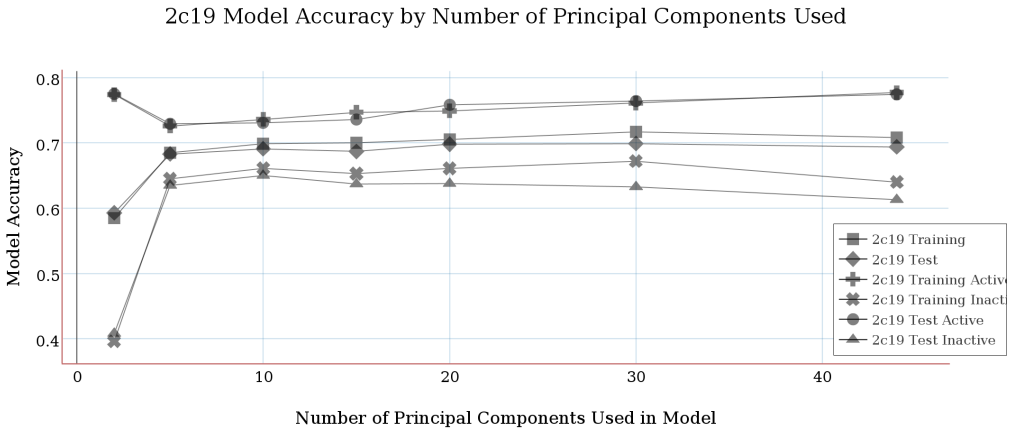
An increased sensitivity is always accompanied by decrease specificity. A ROC curve is plotted as $sensitivity$ versus $1 - specificity$, at varied discrimination cut-offs. An area under the ROC curve (AUROC) close to 1 means that the classifier can perfectly separate the two classes, whereas an area 0.5 indicated that the classifier performs no better than random guessing.[Lapins et al., 2013]

Here's a scratch pad for all my results tables.

2c19 Training Set Accuracy			
PC	Total	Active	Inactive
2	0.585	0.774	0.396
5	0.685	0.726	0.645
10	0.699	0.736	0.661
15	0.700	0.747	0.653
20	0.705	0.749	0.661
30	0.717	0.761	0.672
44	0.708	0.777	0.640

2c19 Test Set Accuracy			
PC	Total	Active	Inactive
2	0.593	0.775	0.408
5	0.683	0.729	0.635
10	0.691	0.731	0.650
15	0.687	0.736	0.637
20	0.698	0.758	0.638
30	0.699	0.764	0.633
44	0.694	0.774	0.613

Table 4.1: 2c19 MOE Model Accuracy



1a2 Training Set Accuracy			
PC	Total	Active	Inactive
2	0.638	0.752	0.522
5	0.704	0.746	0.662
10	0.734	0.755	0.712
15	0.737	0.770	0.704
20	0.741	0.781	0.701
30	0.735	0.782	0.686
44	0.725	0.777	0.672

1a2 Test Set Accuracy			
PC	Total	Active	Inactive
2	0.632	0.758	0.517
5	0.705	0.773	0.643
10	0.746	0.749	0.743
15	0.752	0.775	0.731
20	0.748	0.779	0.721
30	0.739	0.783	0.698
44	0.720	0.773	0.672

Table 4.2: 1a2 MOE model results

2c9 Training Set Accuracy			
PC	Total	Active	Inactive
2	0.620	0.695	0.547
5	0.683	0.722	0.645
10	0.699	0.744	0.655
15	0.702	0.752	0.653
20	0.710	0.756	0.665
30	0.711	0.754	0.668
44	0.712	0.770	0.655

2c9 Test Set Accuracy			
PC	Total	Active	Inactive
2	0.609	0.702	0.510
5	0.675	0.713	0.634
10	0.682	0.730	0.632
15	0.686	0.739	0.630
20	0.685	0.731	0.636
30	0.690	0.730	0.646
44	0.688	0.750	0.621

Table 4.3: 2c9 MOE model results

2d6 Training Set Accuracy			
PC	Total	Active	Inactive
2	0.589	0.753	0.426
5	0.670	0.706	0.634
10	0.685	0.721	0.648
15	0.686	0.709	0.662
20	0.705	0.721	0.689
30	0.703	0.732	0.673
44	0.690	0.725	0.656

2d6 Test Set Accuracy			
PC	Total	Active	Inactive
2	0.590	0.738	0.443
5	0.667	0.685	0.650
10	0.662	0.688	0.636
15	0.665	0.692	0.639
20	0.683	0.699	0.668
30	0.686	0.696	0.677
44	0.669	0.707	0.632

Table 4.4: 2d6 MOE model results

3a4 Training Set Accuracy			
PC	Total	Active	Inactive
2	0.644	0.699	0.589
5	0.675	0.727	0.623
10	0.698	0.741	0.655
15	0.701	0.746	0.656
20	0.705	0.749	0.660
30	0.709	0.761	0.657
44	0.698	0.773	0.623

3a4 Test Set Accuracy			
PC	Total	Active	Inactive
2	0.627	0.683	0.573
5	0.656	0.719	0.595
10	0.677	0.719	0.637
15	0.680	0.733	0.628
20	0.686	0.736	0.638
30	0.686	0.742	0.631
44	0.676	0.747	0.607

Table 4.5: 3a4 MOE model results

Training Set Methods Comparison					
Isozyme	1a2	2c9	2c19	2d6	3a4
MOE 20pc	0.741	0.710	0.705	0.705	0.705
kNN	0.761	0.721	0.730	0.717	0.716
Random Forest	0.770	0.717	0.736	0.729	0.719
SVD	0.800	0.749	0.767	0.755	0.755
Ensemble					

Table 4.6: Comparison of methods for Training Set

Test Set Methods Comparison					
Isozyme	1a2	2c9	2c19	2d6	3a4
MOE 20pc	0.748	0.685	0.698	0.683	0.686
kNN	0.754	0.692	0.720	0.678	0.674
Random Forest	0.769	0.687	0.721	0.707	0.683
SVD	0.804	0.720	0.756	0.725	0.714
Ensemble					

Table 4.7: Comparison of Methods for Test Sets

Chapter 5

Discussion

5.0.1 Cheng's results compared to mine

It is highly desirable to develop computational models that can predict the inhibitive effect of a compound against a specific CYP isoform. In this Cheng's study inhibitor predicting models were developed for five major CYP isoforms, namely 1A2, 2C9, 2C19, 2D6, and 3A4, using a combined classifier algorithm on a large data set containing more than 24,700 unique compounds, extracted from PubChem. The combined classifiers algorithm is an ensemble of different independent machine learning classifiers including support vector machine, C 4.5 decision tree, /*kappa*-nearest neighbor, and naive Bayes, joined by a back-propagation artificial neural network (BP-ANN). Those models were validated by 5-fold cross-validation with a diverse validation set composed of about 9000 diverse unique compounds. The range of the AUROC for the validation set was 0.764 to 0.886. The overall per-

formance of combined classifiers fused by BP-ANN was superior to that of three classic fusion techniques (Mean, Maximum, and Multiply). Cheng et al. claim these classification models are applicable for virtual screening of the five major CYP isoforms inhibitors or can be used as simple filters of potential chemicals in drug discovery.[Cheng et al., 2011]

5.1 Machine Learning

Developing successful machine learning applications still requires a substantial amount of black art that is hard to find in textbooks.[Domingos, 2012]

Easily the most important factor is the features used. If you have many independent features that correlate well with the class, learning is easy. On the other hand, if the class is a very complex function of the features, you may not be able to learn it.[Domingos, 2012]

Often raw data is not in a form that is amenable to learning, but you can construct features from it that are. This is typically where most of the effort in a machine learning project goes. It is often one of the most interesting parts, where intuition, creativity and black art are as important as the technical stuff.[Domingos, 2012]

...consider how time-consuming it is to gather data, integrate it, clean it, and preprocess it, and how much trial and error can go into feature design.[Domingos, 2012]

Also machine learning is not a one-shot process of building a data set and

running a learner, but rather an iterative process of running the learner, analyzing results, modifying the data and/or learner, and repeating. Learning is often the quickest part of this, but thats because weve already mastered it pretty well! Feature engineering is more difficult because it is domain-specific, while learners can be largely general-purpose.[Domingos, 2012]

There is ultimately no replacement for the smarts you put into feature engineering. On the other hand, running a learner with a very large number of features to find out which ones are useful in combination may be too time-consuming, or cause overfitting.[Domingos, 2012]

As a rule, it pays to try the simplest learners first(e.g. naive Bayes before logistic regression, κ-nearest neighbor before support vector machines). More sophisticated learners may be seductive, but they are usually harder to use, because there are more tuning parameters and their internals are more opaque.[Domingos, 2012]

Learners can be divided into two major types: those whose representation has a fixed size, like linear classifiers, and those whose representation can grow with the data, like decision trees. (The latter are sometimes called non-parametric, but this is somewhat ufortuante, since they usually wind up learning many more parameters than parametric ones.) Fixed-size learners can only take advantage of so much data. Variable-sized learners can in principle learn any function given sufficient data, but in practice they may not, because of limitations of the algorithm or computational cost. Also, because of the curse of dimensionality, no existing amount of data may be

enough.[Domingos, 2012]

Creating model ensembles is now standard in machine learning. In the simplest technique, called bagging, we simply generate random variations of the training set by resampling, learn a classifier on each, and combine the results by voting. This works because it greatly reduces variance while only slightly increasing bias. In boosting, training examples have weights, and these are varied so that each new classifier focuses on the examples the previous ones tended to get wrong. In stacking, the outputs of individual classifiers become the inputs of a higher-level learner that figures out how best to combine them.[Domingos, 2012]

Representable does not imply learnable.[Domingos, 2012]

Correlation does not imply causation.[Domingos, 2012]

but I have fallen short of this goal due to my own technical shortcomings and the necessary inclusion of proprietary software for crucial descriptor generation.

Chapter 6

Conclusion

In summary, I have demonstrated several methods for building models for binary classification for QSAR of cytochrome P450 isozymes. They all do pretty well and are significantly better than randomly guessing. There are many more machine learning methods to try, of particular interest are neural networks and more Bayesian methods. The data and code for model comparison for this study is available online, so there are no artificial barriers to benchmarking further results.

Ensembles of models frequently outperform any single model, so before building a tool for decision making in drug development, it is recommended to combine results from all models developed.

Not Related - Systems Biology

The ultimate goal of systems biology is an understanding of physiology and disease across the multiple hierarchical levels of organization, from chemical and molecular interactions to pathways and pathway networks, at the cell-cell and issue level, organs and organ systems and, ultimately, to the functioning of the whole organism.(Ellen Berg 2014)

Systems biology research encompasses the generation of highthroughput datasets of system components (omics data), experimental methods of analysis and data integration, as well as the development and application of network approaches and computationally derived models(Berg 2014)

In pharmaceutical research, systems biology efforts are directed towards the identification of drug targets, the development of novel therapeutics and new indications for existing drugs.

Studies tend to be compound-centric, concerned with the identification and characterization of small molecules or biologics that selectively inhibit (or activate) specific molecular targets or pathway mechanisms. Thus, studies related t drug mechanisms of action and those that support drug development goals, such as clinical indication selection and patient stratification, are of particular interest. (Berg 2014)

Omics tools, developed over the past several decades, can provide global information on the levels and dynamic changes in cellular and tissue components at specific time points insamples from cellbased assays, precinical animal models or human studies. Omics data sets derived from transcrip-

tomics, proteomics, and metabolomics are being used and integrated with each other as well as genomics information and other data types to construct models of cell signaling, pathway and disease networks to identify new targets as well as to help better understand and predict drug action in vivo.

Bibliography

[Berg, 2014] Berg, E. L. (2014). Systems biology in drug discovery and development. *Drug discovery today*, 19(2):113–25.

[Boulesteix and Schmid, 2014] Boulesteix, A. L. and Schmid, M. (2014). Machine learning versus statistical modeling. *Biometrical Journal*, 4(56):588–593.

[Breiman, 2001] Breiman, L. (2001). Statistical Modeling : The Two Cultures. 16(3):199–231.

[Cheng et al., 2011] Cheng, F., Yu, Y., Shen, J., Yang, L., Li, W., Liu, G., Lee, P. W., and Tang, Y. (2011). Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers. *Journal of chemical information and modeling*, 51(5):996–1011.

[Danielson, 2002] Danielson, P. B. (2002). The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. *Current drug metabolism*, 3(6):561–97.

[Domingos, 2012] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78.

[Hansch and Fujita, 1964] Hansch, C. and Fujita, T. (1964). Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of American Chemical Society*, 86:1616–1626.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

[Kruhlak et al., 2012] Kruhlak, N. L., Benz, R. D., Zhou, H., and Colatsky, T. J. (2012). (Q)SAR modeling and safety assessment in regulatory review. *Clinical pharmacology and therapeutics*, 91:529–34.

[Lapins et al., 2013] Lapins, M., Worachartcheewan, A., Spjuth, O., Georgiev, V., Prachayasittikul, V., Nantasenamat, C., and Wikberg, J. E. S. (2013). A unified proteochemometric model for prediction of inhibition of cytochrome p450 isoforms. *PloS one*, 8(6):e66566.

[Nantasenamat, 2009] Nantasenamat, C. (2009). A practical overview of quantitative structure-activity relationship. *EXCLI J*, 8:74–88.

[Prli and Lapp, 2012] Prli, A. and Lapp, H. (2012). The plos computational biology software section. *PLoS Comput Biol*, 8(11):e1002799.

[Singh et al., 2011] Singh, D., Kashyap, A., Pandey, R. V., and Saini, K. S. (2011). Novel advances in cytochrome P450 research. *Drug discovery today*, 16(17-18):793–9.

[Veith et al., 2009] Veith, H., Southall, N., Huang, R., and James, T. (2009). Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nature ...*, 27(11):1050–1055.

[Visser et al., 2014] Visser, S. a. G., de Alwis, D. P., Kerbusch, T., Stone, J. a., and Allerheiligen, S. R. B. (2014). Implementation of quantitative and systems pharmacology in large pharma. *CPT: pharmacometrics & systems pharmacology*, 3(October):e142.

[Zlokarnik et al., 2005] Zlokarnik, G., Grootenhuis, P. D. J., and Watson, J. B. (2005). High throughput P450 inhibition screens in early drug discovery Case study REVIEWS. 10(21):1443–1450.