THE QSAROME OF THE RECEPTOROME: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP MODELING OF MULTIPLE LIGAND SETS ACTING AT MULTIPLE RECEPTORS

Guiyu Zhao

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Division of Chemical Biology and Medicinal Chemistry at Eshelman School of Pharmacy

Chapel Hill
2011

Approved by:

Dr. Alexander Tropsha

Dr. Bryan Roth

Dr. Steve Marron

Dr. Qisheng Zhang

Dr. Shawn Gomez

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

ABSTRACT

GUIYU ZHAO: The QSARome of the Receptorome: Quantitative Structure-Activity
Relationship Modeling of Multiple Ligand Sets Acting at Multiple Receptors
(Under the direction of Alexander Tropsha)

Recent advances in High Throughput Screening (HTS) led to the rapid growth of chemical libraries of small molecules, which calls for improved computational tools and predictive models for Virtual Screening (VS). Thus this dissertation focuses on both the development and application of predictive Quantitative Structure-Activity Relationship (QSAR) models and aims to discover novel therapeutic agents for certain diseases.

First, this dissertation adopts the combinatorial QSAR framework created by our lab, including the first application of the Distance Weighted Discrimination (DWD) method that resulted in a set of robust QSAR models for the 5-HT_7 receptor. VS using these models, followed by the experimental test of identified compounds, led to the finding of five known drugs as potent 5-HT_7 binders. Eventually, droperidol (K_i = 3.5 *nM*) and perospirone (K_i = 8.6 *nM*) proved to be strong 5-HT_7 antagonists.Second, we intended to enhance VS hit rate. To that end, we developed a cost/benefit ratio as an evaluation performance metric for QSAR models. This metric was applied in the Decision Tree machine learning method in two ways: (1) as a benchmarking criterion to compare the prediction performances of different classifiers and (2) as a target function to build QSAR classification trees. This metric may be more suitable for imbalanced HTS data that include few active but many inactive compounds.

Finally, a novel QSAR strategy was developed in response to the polygenic nature of most psychotic disorders, related mainly to G-Protein-Coupled Receptors (GPCRs), one class of molecular targets of greatest interest to the pharmaceutical industry. We curated binding data for thousands of GPCR ligands, and developed predictive QSAR models to assess the GPCR binding profiles of untested compounds that could be used to identify potential drug candidates. This comprehensive study yielded a compendium of validated QSAR predictors (the GPCR QSARome), providing effective *in silico* tools to search for novel antipsychotic drugs.

The advances in results and procedures achieved in these studies will be integrated into the current computational strategies for rational drug design and discovery boosted by our lab, so that predictive QSAR modeling will become a reliable support tool for drug discovery programs.

ACKNOWLEDGEMENT

I give my truehearted thanks to my mentor, Dr. Alexander Tropsha, for his scientific guidance and education throughout my graduate study and research. His insight, courage, and determination all inspired me to dedicate myself to hard work and the courage to inquire. I thank him for nurturing me from a naïve computational scientist. I received great inheritance from Alex's lab and hope to pass it on in my future life.

I am very grateful to Dr. Denis Fourches, Dr. Eugene Muratov, Dr. Clark Jeffries, and Dr. Xiang (Simon) Wang, for their invaluable help and scientific inspirations. Without their professional skills and close mentorship of my projects, my research would never have been completed.

I also want to thank my other committee members Dr. Bryan Roth, Dr. Steve Marron, Dr. Qisheng Zhang, and Dr. Shawn Gomez. They provided me with wonderful ideas for my projects. Special thanks are given to Dr. Bryan Roth for providing the fruitful PDSP database and insightful interpretation of our computational results.

I give thanks to other members and former members of the molecular modeling lab for their friendship and support, especially Dr. Alexander Golbraikh, Dr. Alexander Sedykh, Yen Low, Dr. Jui-Hua Hsieh, Dr. Hao Tang, Dr. Liying Zhang, and Dr. Hao Zhu, for their daily ardent help with my research projects.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD	Applicability Domain
ADME	Absorption, distribution, metabolism, and excretion
CADD	Computer-aided Drug Design
CCR	Correct Classification Rate
CV	Cross Validation
GPCR	G Protein-Coupled Receptor
HTS	High Throughput Screening
HTS	High Throughput Screening
kNN	k-Nearest Neighbors
LBVS	Ligand-based Virtual Screening
LOO	Leave One Out
MLR	Multiple Linear Regression
PLS	Partial Linear Squares
QSAR	Quantitative Structure Activity Relationship
RF	Random Forest
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
SBVS	Structure-based Virtual Screening
SE	Sensitivity
SP	Specificity
SVM	Support Vector Machines
TC	Tanimoto Coefficient
VS	Virtual Screening

CHAPTER 1

INTRODUCTION

1.1. Overview

The drug discovery and development pipeline is a notoriously time-consuming and costly process. To successfully launch one New Chemical Entity (NCE) from the discovery stage to market takes about 15 years and hundreds of millions (Figure 1.1)[1]. Applying Computer-Aided Drug Design (CADD) strategies could provide both time- and cost-savings for drug research and development programs (i.e., integration of computational tools into the standardized pipeline should further raise the efficiency of drug design).

As an integral part of CADD, Quantitative Structure-Activity Relationships (QSAR) is experiencing one of the most important periods in its history, highlighted by the availability of vast chemical databases with abundant bioactivity data, such as ChEMBL[2], PDSP[3], and dozens of others[4]. The explosive growth of such data provides a good opportunity for large-scale QSAR modeling across diverse pharmaceutically relevant targets. Resulting QSAR models could become valuable tools for identifying novel molecular probes and potential leads for drug discovery.

To develop statistically robust QSAR models, our lab built a rigorous workflow for development and validation of QSAR models. Major stages of this workflow include the division of the original datasets into training, test, and external validation sets; Y-randomization validation; model selection based on given statistical performance; and Virtual Screening (VS)[5–8]. The underlying components of this workflow, such as data curation, development of ensemble models, hit rate of VS, or Applicability Domain (AD)[9][9], are all active research areas and further improvement in overall model predictivity can be expected through their advancements.

This dissertation focuses on the target class of G Protein-coupled Receptors (GPCRs), a group of molecular targets of great interest to pharmaceutical industry[10]. As of 2003, the number of GPCRs in human genome from five main families (glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin) had been estimated at over 800[11]. However, the true number is much higher now due to the known existence of alternatively spliced variants and editing isoforms of GPCRs. In addition, GPCRs with unknown functions (i.e., lack of known natural transmitters), called "orphan" GPCRs, account for a large portion of newly identified GPCRs[12].

The impact of GPCRs on drug discovery is phenomenal. Previous studies suggest that at least one-third[13], and perhaps up to half[14] of currently marketed drugs target GPCR family members, which represent only around 3% of known molecular targets[15]. Actively ongoing studies of GPCRs such as deorphanization of orphan GPCRs provide huge opportunities for new drug discovery. For instance, the majority of drug targets related to central nervous system (CNS) disorders (e.g., depression, schizophrenia and bipolar disorder) belong to this receptor family. However, most antipsychotic drugs have complex GPCR polypharmacology, leading either to therapeutic effects or undesired adverse events. Thus, it will be beneficial to understand functioning bioprofiles of GPCR ligands to enhance their potential therapeutic effects and avoid possible adverse reactions. Our goal of focusing on this receptor class is to search for antipsychotic drugs, both selective to a specific GPCR and those that non-selectively target a combination of critical GPCRs.

1.2. Quantitative Structure-Activity Relationship (QSAR)

Previous studies (e.g., SAR analysis) have shown that structural features of small molecules have significant effect on their physicochemical and biological properties. Compared with conventional SAR analysis, the QSAR analysis intends to *quantitatively* explain the relationship between chemical structures and the corresponding activity. The QSAR analysis is based on the assumption that compounds with similar structures are expected to exhibit similar properties (the Similarity Property Principle[16]). This assumption serves as a foundation behind experimental SAR studies by medicinal chemists, as well as the basis for computational QSAR studies since the 1960s when Dr. Corwin Hansch established the very first QSAR analysis to predict chemical solubility[17]. However, the definition of similarity is not straightforward because the estimated degree of similarity depends on a number of underlying factors such as molecular descriptors, variable selection methods, and the similarity metrics.

To briefly explain the fundamental concepts, any QSAR method can be generally expressed in the following form[18]:

$$\mathrm{P_i} = \hat{k}(D_1, D_2, \mathrm{K}\ , D_n)........................(1.1)$$

Where $\mathrm{P_i}$ is the biological activity of molecule I (dependent variable), D_1, D_2, …, D_n are independent variables, which are either calculated molecular descriptors or experimentally measured properties of molecule i, and $k(D_i)$ is a function that relate the descriptors to the biological activity $\mathrm{P_i}$. $k(D_i)$ could be either linear (whose output is directly proportional to its input variables) or nonlinear (whose output is not directly proportional to its input variables) function, depending on the expected relationship between the descriptor values D (input variables) and target property P (output). In essence, all machine learning techniques aim to find such mathematical representation of $k(D_i)$ that would best reproduce the trend in biological activities.

The recent explosive growth of experimental data due to the technological advances in High Throughput Screening (HTS)[19–22] calls for the use of fast QSAR methods to establish QSAR models of large and complex data sets. During the past few decades of development, the field of QSAR has grown rapidly in terms of novel molecular descriptors, nonlinear regression methods, QSAR for toxicity and ADME (Absorption, Distribution, Metabolism, and Excretion), and 3D QSAR[23–28]. The differences among various QSAR approaches mainly depend on the descriptors used to characterize the molecules and the machine learning methods used to establish relationships between input descriptor values and biological activities. To list a few popular methods, nonlinear approaches of multivariate analysis include the Decision Trees[29], Random Forest (RF)[30], Artificial Neural Networks (ANN)[31], *k* Nearest Neighbors (*k*NN)[32], and Support Vector Machines (SVM)[33]. However, the most serious issue faced by these methods is the High-Dimension Low-Sample Size (HDLSS) problem, which means that the number of descriptors (usually from hundreds to thousands) is much greater than the number of samples in the studied dataset (less than a hundred compounds is common). To overcome this problem, we have applied recent developed Distance Weighted Discrimination (DWD) method[34] that was developed as a more robust alternative to SVM and is capable of handling HDLSS problem common for small modeling datasets.

1.3. Validation Criteria for Virtual Screening

Aside from interpretation of found relationships, important practical application of validated QSAR models is to screen large untested databases to assist the discovery of novel bioactive chemical entities [6,7]. At this point, two important aspects should be clarified: the classification of QSAR approaches based on the nature of the modeled response variable (**target property**), and the importance of rigorous model validation[9].

Generally speaking, QSAR approaches can be grouped in to three classes according to the target properties (referred to as dependent variables or response variables in statistical data modeling sense): classification, category, and continuous QSAR[35]. To explain in more detail, classes of target properties are different from categories in terms of whether or not they can be ordered in some scientifically meaningful way. The former, also regarded as categorical unrelated, cannot be rank ordered, i.e., classes do not relate to each other in any continuum. For example, compounds belonging to different pharmacological classes (interacting with different receptors) or classified as drugs vs. non-drugs cannot be rank ordered. On the other hand, the categorical related, can be rank ordered as the classes of target properties that cover certain ranges of values, e.g., very active, active, moderately active, and inactive. For the purpose of subsequent analysis, such classes are often encoded numerically (for example, one for active or zero for inactive). Continuous QSAR is based on the real values covering certain range, e.g., pK_i (log-transformed binding constant values), IC_{50}, ED_{50}, *etc*. Understanding this classification is very important when considering the nature of target property, its data quality, the choice of molecular descriptors and associated modeling techniques. Often, continuous activity data can be categorized and modeled as such to avoid fitting models to the experimental noise.

The choice of validation procedures and criteria is often dictated by the type of target properties which defines the classes of QSAR practices. For validation of QSAR models, Y-randomization test (randomization of the response variable)[36] is often used to check for the possibility of chance correlation[37]. Y-randomization procedure is discussed in detail in Chapter 2. In addition, the most critical way to ensure the predictive power of a QSAR model is estimating its performance on a validation (test) set which was not used in model development[38]. The model must demonstrate a significant correlation between predicted and observed target activities of compounds in such an external dataset. The practical way to achieve this is to divide experimental data into the training and test sets[39]. The criteria to select models from the training set, however, can be subject to a series of filtering rules. Many authors apply the leave-one-out (LOO) or leave-some-out (LSO) cross-validation procedure to the entire modeling dataset, which is now considered as insufficient for rigorous model validation[40]. For continuous QSAR models, the outcome of this procedure is a cross-validated correlation coefficient q^2 for the training set of compounds, and R^2 for the test set, which are calculated respectively by the formulas[38]:

$$q^2 = 1 - \frac{\sum (y_i - \tilde{y}_i)^2}{\sum (y_i - \bar{y})^2} \text{.................................}(1.2)$$

$$R^2 = \frac{\sum (y_i - \bar{y})^2 (\tilde{y}_i - \bar{\tilde{y}})^2}{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2} \text{...................}(1.3)$$

where y_i, $\tilde{y}_i$, and $\bar{y}$ (or $\bar{\tilde{y}}$) are the actual, predicted, and the average actual (or predicted) activities, respectively. We emphasize highly on the ability of the models to predict the activity of compounds in an external validation set, instead of only considering high q^2 as an indicator or even the ultimate proof of the predictive power of a QSAR model, which is often misleading and cannot guarantee the extrapolation power of respective models.

Correct classification rate (*CCR*) is often used to evaluate the predictivity of a binary classification model (i.e., for a two categories of activity that are usually called “active” and “inactive”). *CCR* is the average of sensitivity (SE) and specificity (SP), which are calculated by below formulas[41]:

$$\mathit{Sensitivity}(SE)=\frac{TP}{TP+FN}\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.4)$$

$$\mathit{Specificity}(SP)=\frac{TN}{TN+FP}\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.5)$$

$$CCR=\frac{SE+SP}{2}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1.6)$$

$$\mathit{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}\ldots\ldots\ldots\ldots\ldots(1.7)$$

where *TP, TN, FP, FN* are true positives (accurately predicted actives), true negatives (accurately predicted inactives), false positives (inactives predicted as actives), and false negatives (actives predicted as inactives), respectively. Together they compose a confusion matrix (Figure 1.2) which is the common resort to evaluate a classifier. *CCR* is preferred as a performance measure of a classifier since it is not biased to the major class in the case of imbalanced data in which the minority class is often more important (e.g., active compounds are often fewer than inactive ones). However, predictive *accuracy* (*Equation* 1.7) simply favors performance of the majority class.

It should be noted that data balancing is an important issue to consider before running a classification modeling. Most machine learning algorithms assume that their training sets are well balanced, and demonstrate poor performance when they deal with imbalanced data sets[42]. However, in many cases we cannot control the influence of this imbalance issue by simply re-sampling the data sets (e.g., removal of certain cases in the major class). As a result, we need to resort to other algorithms resistant to the issue.

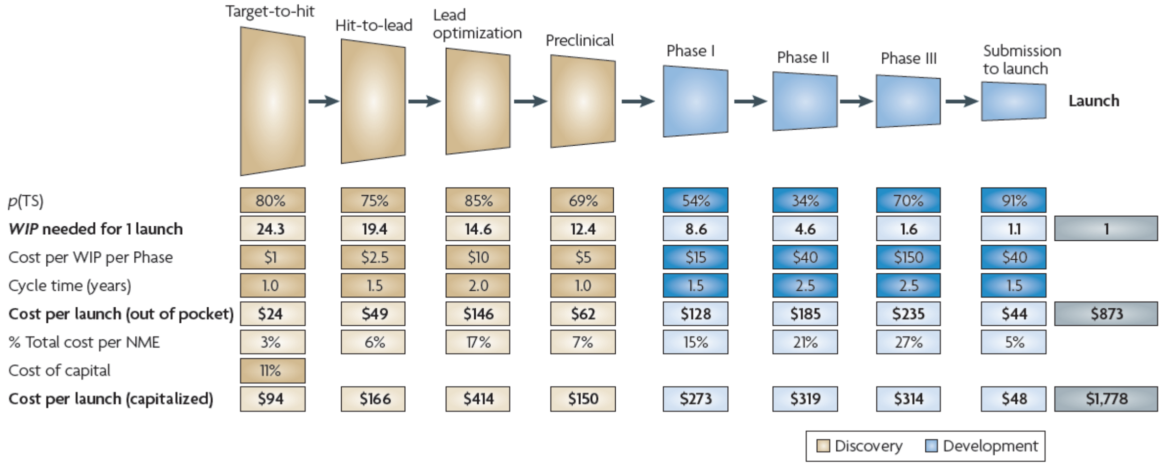
1.4. Thesis Outline

This dissertation concentrates on the application of QSAR approaches to discover novel antipsychotic drugs. Extensive efforts have been made in terms of data collection and curation, QSAR modeling, and the quest for suitable evaluation criteria.

Chapter 2 illustrates the successful practice of identifying FDA-approved drugs with newfound 5-HT$_7$ binding affinity by applying rigorous QSAR modeling workflow. The 5-HT$_7$ receptor, a member of the GPCR family, is postulated to be a potential drug target for psychotic disorders, especially for schizophrenia. A combi-QSAR approach established by our lab was used to develop predictive continuous models using *k* Nearest Neighbor (*k*NN) and classification models using Distance Weighted Discrimination (DWD). Models were rigorously validated by Y-randomization and demonstrated high accuracy in predicting external datasets. VS of the publically available compound database World Drug Index (WDI) followed by experimental testing successfully identified five known drugs with first identified 5-HT$_7$ binding affinity. Two of these drugs have been confirmed as 5-HT$_7$ receptor antagonists, which could be repositioned to treat schizophrenia.

In Chapter 3, we propose a new evaluation metric, the Economic Ratio (ER), not only as a performance parameter for the developed models, but also as a target function during model training. After applying this metric with the Decision Tree (DT) machine learning method to various datasets, we found that some trees generated using ER differ in structure and performance from those generated using traditional metrics for branch selection. The cost/benefit economic ratio, ER, can thus be used in two different but complementary ways: (1) as a benchmarking criterion to compare the prediction performances of various classifiers and (2) as a target function to build QSAR classification trees.

In Chapter 4 and Chapter 5, we extend our modeling strategy to develop multiple robust predictors for a set of GPCR targets. The resulting models can be used to predict the binding bioprofiles of untested chemicals. In summary, we curated and integrated binding data for thousands of GPCR ligands extracted from both ChEMBL and PDSP databases. First, we used 5-HT$_{1A}$ as a scheme to decide what properties should be applied for data collection and modeling processes, which resulted in rigorous standards for both chemical and biological data curation. We then developed robust classification QSAR models based on the ligands of the large set of GPCRs; that is, 34 GPCRs in total including 5-HT$_{1A}$. The validated models were applied to assess the GPCR binding profiles of 13 drugs not present in the modeling sets, and we found the accuracy was as high as 70.5%. This extensive study yielded a compendium of validated QSAR potency predictors, the GPCR QSARome, providing an effective *in silico* means to search for novel antipsychotic drugs and to unveil their complex polypharmacological nature.



	Target-to-hit	Hit-to-lead	Lead optimization	Preclinical	Phase I	Phase II	Phase III	Submission to launch	**Launch**
p(TS)	80%	75%	85%	69%	54%	34%	70%	91%	
***WIP* needed for 1 launch**	**24.3**	**19.4**	**14.6**	**12.4**	**8.6**	**4.6**	**1.6**	**1.1**	**1**
Cost per WIP per Phase	$1	$2.5	$10	$5	$15	$40	$150	$40	
Cycle time (years)	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5	
Cost per launch (out of pocket)	**$24**	**$49**	**$146**	**$62**	**$128**	**$185**	**$235**	**$44**	**$873**
% Total cost per NME	3%	6%	17%	7%	15%	21%	27%	5%	
Cost of capital	11%								
Cost per launch (capitalized)	**$94**	**$166**	**$414**	**$150**	**$273**	**$319**	**$314**	**$48**	**$1,778**

Figure 0.1. R&D model yielding costs to successfully discover and develop a single new molecular entity (NME).

Money unit: million. Work in process, WIP. Probability of successful transition from one stage to the next, p(TS).

(Modified from *Steven M. Paul, et al. Nature Reviews Drug Discovery, 2010, 9: 203-214*.)

	Predicted Negative	Predicted Positive
Actual negative	TN	FP
Actual Positive	FN	TP

Figure 0.2. The confusion matrix used to evaluate a classifier.

Note: The columns are the predicted classes and the rows are the actual classes. TN is the number of negative cases corrected predicted (True Negatives), FP is the number of negative cases incorrectly predicted as positive (False Positives), FN is the number of positive cases incorrectly predicted as negative (FN), and TP is the number of positive cases correctly predicted as positive (True Positives).

CHAPTER 2

APPLICATION OF CURRENT CHEMINFORMATIC TECHNIQUES TO HUMAN 5-HT$_7$ DATASETS TO BUILD VALIDATED AND PREDICTIVE QSAR MODELS FOR DRUG REPURPOSING

2.1. Introduction

5-hydroxytryptamine (5-HT) receptors are involved in a large number of physiological and behavioral functions[43–46]. Many antipsychotic drugs act through multiple molecular targets including 5-HT receptors. Although it received little attention when first cloned in 1993, the 5-HT$_7$ receptor has become the most studied member of the 5-HT receptor family now[47]. Several distribution studies indicated that the 5-HT$_7$ receptors are located mainly in thalamus, hippocampus, and hypothalamus with relatively lower concentrations in the amygdala and cerebral cortex[48,49]. Additionally, 5-HT$_7$ receptor subtypes are found in smooth muscle cells and other peripheral tissues[50]. Scientific research on the 5-HT$_7$ receptor has mainly focused on its therapeutic effects for psychiatric disorders, especially for major depression[51] and schizophrenia[52]. Previous studies show that 5-HT$_7$ antagonists modulate the level of 5-HT and thus increase neurogenesis, indicating 5-HT$_7$ receptor is a promising molecular target for antidepressants[53]. A series of studies identify 5-HT$_7$ receptors as critical in hippocampus-dependent functions including learning and memory[54–56]. In addition, the presence of 5-HT$_7$ receptor subtypes in smooth muscle cells suggests