University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2012

Structure-activity relationship model for estrogen receptor ligands.

Huihui Wu 1979-
University of Louisville

Follow this and additional works at: https://ir.library.louisville.edu/etd

Recommended Citation

Wu, Huihui 1979-, "Structure-activity relationship model for estrogen receptor ligands." (2012). *Electronic Theses and Dissertations.* Paper 1596.
https://doi.org/10.18297/etd/1596

STRUCTURE-ACTIVITY RELATIONSHIP MODEL FOR ESTROGEN

RECEPTOR LIGANDS

By

Huihui Wu

B.S., Anhui College of Traditional Chinese Medicine, 2003

M.S., Shanghai University of Traditional Chinese Medicine, 2006

A Thesis

Submitted to the Graduate Faculty of the

School of Medicine of the University of Louisville

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

Department of Pharmacology and Toxicology

University of Louisville

Louisville, Kentucky

May 2012

STRUCTURE-ACTIVITY RELATIONSHIP MODEL FOR ESTROGEN

RECEPTOR LIGANDS

By

Huihui Wu

B.S., Anhui College of Traditional Chinese Medicine, 2003

M.S., Shanghai University of Traditional Chinese Medicine, 2006

A Thesis Approved on

April 18, 2012

By the following Thesis Committee

Albert R. Cunningham, Ph.D.

John Trent, Ph.D.

Chi Li, Ph.D.

Keith R. Davis, Ph.D.

This thesis is dedicated to my beloved family and friends,

who have given me continuous love and support.

ABSTRACT

STRUCTURE-ACTIVITY RELATIONSHIP MODEL FOR ESTROGEN

RECEPTOR LIGANDS

Huihui Wu

November 29, 2011

Xenoestrogens are spread throughout the environment affecting our daily

lives and may produce potential toxic effects on human health. The purpose of

this study was to develop a mechanistically reliable model capable of identifying

xenoestrogens. Our hypothesis was that there are identifiable structural

characteristics among a diverse set of estrogen receptor ligands that differentiate

estrogenic and nonestrogenic compounds. The model's learning set was

developed by collecting compounds from the National Center for Toxicological

Research Estrogen Receptor Binding database (NCTRER) . The

categorical-SAR (cat-SAR) expert system was used to build the models and

perform leave-none-out, leave-one-out, leave-many-out and external validations

for model analysis. The values of all validations were between 0.80 and 0.97.

Based on several analyses of rational subsets of compounds included in the

NCTRER based on potency or chemical structure, it was observed that the

developed SAR models predictivity varied across sets. This indicates that

variability in the SAR models or the *in vitro* assay results themselves must be

considered when applying SAR models for prediction or mechanistic analyses of

estrogen receptor ligands. Fragment analysis was carried out to study the mechanism of estrogen receptor binding, and various important fragments were identified that demonstrate potential structural characteristics important for binding. Furthermore, this led to the discovery that the cat-SAR expert system was able to make a higher percentage of correct predictions on specific classes of xenoestrogen expressing these key functional groups. In conclusion, this estrogen receptor ligand model has good predictive performance and is based on model attributes that are mechanistically sound.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Estrogen is an essential hormone in many biological processes such as sexual development, reproduction, cardiovascular and bone health. It is a steroid hormone, and includes three naturally occurring types: estrone, 17β-estradiol and estriol. 17β-estradiol is the predominant form in non-pregnant females. For medical applications, estrogen receptor agonists and antagonists can be used as oral contraceptives, hormone replacement therapies, and breast cancer therapies.

Estrogen and its derivatives produce effects though the interaction with the estrogen receptor (ER). As of now, three types of ERs have been identified: ER-α, ER-β, and G protein-coupled receptor 30 (GPR30) [1-4]. ER-α and ER-β are classic nuclear receptors and act as ligand-activated nuclear transcription factors that bind regulatory response elements in the promoter regions of genes [5] and regulate gene expression. GPR30 is a seven transmembrane domain G protein-coupled receptor (GPCR) with low homology to existing GPCRs [6] It binds estrogen and triggers the rapid non-genomic signaling events such as mitogen-activated protein kinase (MAPK) and Akt pathway activation [7]. Therefore, estrogens can trigger both genomic and non-genomic signaling pathways. Also, the ERs are widely distributed throughout the body, and found in such systems as the cardiovascular, nervous, reproductive, and musculoskeletal

1

[8]. The various types and locations of ERs as well as the multiple effects they cause make them critical in human physiology and pathology.

Recent studies have shown that there are various types of chemicals found in the environment that can mimic the action of natural estrogen. These compounds are called endocrine disruptors (EDs), and are substances that "interfere with the synthesis, secretion, transport, binding, action, or elimination of natural hormones in the body that are responsible for development, behavior, fertility, and maintenance of homeostasis (normal cell metabolism)" [9]. In 2009 The Endocrine Society released a scientific statement outlining mechanisms and effects of endocrine disruptors on reproduction, development, breast cancer, prostate cancer, neuroendocrinology, thyroid, metabolism and obesity, and cardiovascular endocrinology. They also used results from experimental and epidemiological studies "to implicate EDs as a significant concern to public health" [10].

EDs can be found in a variety of materials, including drugs, plant constituents, pesticides, compounds used in the plastics industry, consumer products, and other industrial by-products and pollutants [41]. Some are pervasive and widely dispersed in the environment. Some are persistent organic pollutants (POP), and can be transported long distances across national boundaries [11]. Food is a major route by which people are exposed to EDs. Diet is thought to account for up to 90% of a person's polychlorinated biphenyls (PCB) and dichlorodiphenyltrichloroethane (DDT) body burden [10]. With the increase in

household products containing EDs, indoor air has also become a significant source of exposure [12].

Xenoestrogens, a subset of EDs, are a diverse group of chemicals that bind to the ER, mimic natural estrogen action, and interfere with normal endocrine system function. Xenoestrogens, which can be produced naturally (e.g. phytoestrogen) or synthetically (e. g. bisphenol A (BPA), DDT, etc.), are currently the most studied EDs due to their various potential effects on human health. Humans are exposed to xenoestrogens in their everyday life, from the food they eat to the products they use, and their potential effects are very complicated. Some xenoestrogens have been implicated in a variety of environmental health problems, in both males and females, including disrupting the normal secretion of hormones and disturbing the body's metabolism, which can have serious consequences, including damage to reproductive functions [13]. DDT has been reported to induce the feminization of gull embryos [14]. Another important potential effect of some xenoestrogens is carcinogenesis, specifically in relation to breast cancer. These xenoestrogens can mimic 17β-estradiol by binding the ER causing alterations to normal gene transcription and expression of ER and may lead to the occurrence of breast cancer [15]. There is substantial evidence in a variety of recent studies to indicate that estrogenic chemicals can increase breast cancer cell line growth in tissue culture [16]. In clinical practice, hormone replacement therapies have been related to increased cases of breast cancer [17].

In spite of the harmful effects, xenoestrogens also have been associated with benefiting human health. For example soybean products, which are rich in phytoestrogens including coumestrol, are common in the daily diets of East Asians and are believed to be the reason that the incidence of breast cancer in East Asian women is much lower than western women [18].

As far as the complicated potential effects of xenoestrogens, a better understanding of xenoestrogens, their identification, and mechanisms of action is of great significance. The United States Environmental Protection Agency (EPA) has initiated a screening and testing strategy to determine whether exogenous substances may have an effect in humans similar to those of natural hormones. Eighty seven thousand chemicals were quoted by the EPA as potentially requiring analysis for endocrine activity [19].

Traditional methods for estrogen determination are bioassays, such as the competitive receptor binding assay, E-screen assay or uterotrophic assay. The E-screen assay was developed to assess the estrogenicity of environmental chemicals using the proliferative effect of estrogens on their target cells (MCF-7) as an end point [39]. The uterotrophic assay is an *in vivo* assay for estrogenicity. It is based on the principle that the growth phase of the uterus in the natural estrous cycle is under the control of estrogen [40]. However, concern about the amount of chemicals needed for the test, prohibits the timely and costly route of bioassay analysis. Therefore, alternative approaches, such as structure activity relationship (SAR) modeling, may overcome these problems and make it possible to screen a large number of chemicals in a reasonable time.

4

SAR is a methodology to study the relationship between the chemical structure of a molecule and its biological activity. The analysis by SAR may identify the chemical groups responsible for producing a biological effect in an organism. Therefore, SAR has become a powerful tool for screening and mechanism of action analysis. SAR models can be built using chemicals with similar or diverse structures that demonstrate similar toxic effects, which can then be used as screening tools for compounds with unknown activities. Furthermore, it may be possible to modify a compound's structure to determine which substructure is associated with a specific biological activity. Medicinal chemists use the techniques of chemical synthesis to insert new chemical groups into a biomedical compound and test the biological effects of the modifications.

In comparison to bioassay studies, SAR modeling may be an efficient way to screen xenoestrogens. Thousands of chemicals can be screened per day, making it a very important alternative and complimentary technique for xenoestrogen screening. By using SAR modeling, some structural alerts related to estrogenic ligand binding have been discovered. However, this information is not complete because the current structural alerts cannot explain the potency or the activity (i.e., from ER binding to high level health effects) of all the xenoestrogens. For example, the aromatic ring is considered an important structural alert for estrogenicity. However, not all aromatic compounds are estrogenic, including flavone and catechin. Furthermore, aromatic estrogenic compounds have a wide range of affinities for the ER. Some of these compounds are strong binders (17β-estradiol), while others are weak binders (BPA). Therefore, aromaticity alone cannot explain

the activity of xenoestrogens. This project will use SAR modeling to build models to screen xenoestrogens with regard to their potency and study the mechanisms for receptor binding. We will also model several common subclasses of ER mimics, such as biphenyl, steroid and phytoestrogen, to investigate the effect of classes on predictivity. By doing so, we hope to get a better understanding of the roles that the existing structural alerts play.

Some SAR models for the estrogen receptor ligands have been developed, including qualitative and quantitative ((Q) SAR) models. For example, the Multiple Computer Automated Statistical Evaluation Expert System (MultiCASE) is a semi-quantitative model, and has been recently employed for screening chemicals with ER binding potential [20]. Gilles et al. used the MultiCASE expert system to do the SAR study on a diverse set of ER ligands. In their study, substructural features associated with ER binding activity and features that prevent receptor binding were identified [20]. The fundamental assumption of MultiCASE is that the observed biological activity of a molecule is governed by substructures called biophores. However, there are certain disadvantages of using MultiCASE. For example, the false positives are not correctable, and predictions may not be defensible and may or may not reflect known mechanisms. This is because MutiCASE is a black box system, and people cannot get into the inside of a model to see the process of its predictions. For QSAR models, the Comparative Molecular Field Analysis (CoMFA) is a widely used 3D QSAR method in drug design, and has been widely used to develop models correlating structural differences in molecules with their ability to compete for binding to the

ER [21]. Wu *et al.* carried out a study on 3D QSAR of flavonoids and ER based on docking. In this study, they identified the structural features associated with estrogenic activity by providing insight into the interaction between the ligands and key amino acid residues in the binding pocket [43]. The basic assumption in CoMFA is that a suitable sampling of static (van der Waals) and electrostatic fields around a set of aligned molecules yields information necessary to explain their biological activities. The disadvantage of CoMFA is that it requires structurally similar compounds and their accurate alignment. In summary, previous xenoestrogen models lacked transparency, were not mechanistically sound, or could not be used with a diverse groups of chemicals. The goal of this study is to eliminate these deficiencies.

For this study, we used the cat-SAR expert system. The cat-SAR expert system tries to compensate for some limitations of the existing modeling systems. It is a computationally based SAR expert system that was originally developed to associate 2D chemical fragments with active and inactive compounds in a learning set. Unlike other 2D approaches including MultiCASE, cat-SAR is transparent and does not include proprietary code. The approach is sharable and allows unrestricted scrutiny, intervention, and optimization throughout the modeling process. Unlike CoMFA, cat-SAR also does not require a congeneric set of molecules, which makes it more applicable for diverse sets of compounds. The previous studies on the MCF-7 cell proliferation model [22] and rat carcinogenesis model [37] with cat-SAR produced validated results which demonstrate that

7

cat-SAR is a reliable modeling method for identifying structural attributes associated with xenoestrogens.

SAR modeling can also contribute to breast cancer prevention and therapy. Currently, effective therapies exist to treat breast cancer, but there is a lack of effective chemopreventative agents. SAR models can act as a screening tool to differentiate the beneficial and harmful effects related to xenoestrogens. Modeling of xenoestrogens may reduce breast cancer risks by allowing for quick identification and understanding of their mechanisms of action. Therefore, SAR can play a role in breast cancer prevention by either reducing the exposure of carcinogenic xenoestrogens or broadening the use of anti-cancer xenoestrogens. For breast cancer therapy, SAR may be helpful to facilitate the development of novel anti-breast cancer medications by maximizing the drug's specific action on the breast tissue but minimizing the toxic effect on other organ sites.

In this thesis, the cat-SAR expert system was used to model the NCTRER database. Cat-SAR models for ER binding will be built based on this database. The effects of ligand potency and chemical structure will be studied. Xenoestrogen will be divided into several groups according to their relative binding affinity (RBA) value and chemical classes, such as biphenyls, diphenymethane, phytoestrogen, phenols, DES, and steroids. The relationship between model accuracy and RBA value and chemical classes will be analyzed based on the cat-SAR models. The model's performance and the structural characteristics of these groups will be studied. Moreover, the fragments of representative chemicals created by the cat-SAR models will be investigated.

CHAPTER II

MATERIALS AND METHODS

The database

The NCTRER data was collected from the *in vitro* ER competitive-binding assay, which provides quantitative assessment of a chemical's ability to bind to the ER. The NCTRER database consists of 232 chemicals of which 131 are ligands, 93 non-ligands and 8 marginally binding compounds. The compounds were selected *a priori* based on structural characteristics and tested in a well validated and standardized *in vitro* rat uterine cytosol ER competitive-binding assay [23, 24]. This assay tested the IC_{50} of 17β-estradiol and each potential ligand. The relative binding affinity (RBA) values were calculated by dividing the IC_{50} of 17β-estradiol by the IC_{50} of the competitor and multiplying by 100 (RBA = (17β-estradiol IC_{50}/ Competitor IC_{50}) $\times$ 100). IC_{50} is a measure of the effectiveness of a compound in inhibiting biological or biochemical function. This quantitative measure indicates how much of a particular drug or other substance (inhibitor) is needed to inhibit a given biological process (or component of a process, i.e. an enzyme, cell, cell receptor or microorganism) by half. The chemicals were divided into ligand or non-ligand by their RBA values. If the RBA value is equal to 0, it is considered as non-ligand. If the RBA value is greater than 1×10^{-5}, it is considered as ligand. Otherwise it is a marginally binding compound. The database is a structurally diverse set of natural, synthetic, and environmental

9

estrogens covering most known estrogenic classes and spanning a wide range of biological activity. It represents the largest published ER binding database of same-assay results generated in a single laboratory [25]. NCTRER database was downloaded from the EPA website. The structures of the 232 chemicals were input into the sybyl8.1 software [42] for the purpose of fragmentation, visualization and record keeping.

Create activity file

The activity file is a crucial file in the modeling process because it is needed for matrix building and all the validations. After the database was built, an activity file was created according to the RBA value reported by the NCTRER database. In the activity file, the ligand was defined as "1", and the non-ligand "0", and the number "1" and "0" were listed in one column in the same order with the database so that it can be input into the database directly. The activity file was saved as a .txt file.

Learning set development

The cat-SAR models were built through a comparison of structural features found amongst two designated categories of compounds in the model's learning set: ligand (active) and non-ligand (inactive). The cat-SAR learning set consists of the chemical name, its structure as a MOL2 file, and its categorical designation (e.g. "1" for ligand and "0" for non-ligand). Typically, organic salts are included as the freebase, simple mixtures and technical grade preparations are included as the major or ligand component, and metals, metal organic compounds, polymers, and mixtures of unknown composition are not included. In our study, we have

10

developed three learning sets: marginal chemicals as ligands (M$^+$), marginal chemicals as non-ligands (M$^-$) and marginal chemicals excluded (non-M).

***In silico* chemical fragmentation and the compound-fragment matrix**

The Tripos Sybyl8.1 HQSAR module [26] was used to fragment the chemicals *in silico* into all possible fragments meeting user-specified criteria. In HQSAR, the attributes were selected for fragments determination such as atom counts (*i.e.*, the size of the fragments), bond types, atomic connections (*i.e.*, the arrangement of atoms in the fragment), explicit hydrogen atoms, chirality, and hydrogen bond donor and acceptor groups. Fragments can be linear, branched, or cyclic moieties. Models developed herein contained fragments between two and seven atoms in size and considered atoms, bond types, and atomic connections as well as the hydrogen atoms.

After fragmentation, a compound-fragment data matrix was produced with a Sybyl HQSAR addon as a text file. In the matrix, the rows are intact chemicals and columns are the molecular fragments. Thus for each chemical, a tabulation of all its fragments are recorded across the table rows, and for each fragment all chemicals that contain it are tabulated down the columns. The compound-fragment matrix is analyzed with the cat-SAR programs to identify structural features associated with the categorized ligand and non-ligand compounds.

Identifying important fragments

A measure of each fragment's association with biological activity was next determined. To ascertain an association between each fragment and biological

activity, a set of rules are parameterized to choose important ligand and non-ligand fragments. The first selection rule is the "number rule" which is the number of chemicals in the learning set that contains a fragment. The second rule is the "proportion rule" which considers the percentage of ligand or non-ligand compounds that possess each fragment. For example 4/0.9/0.85, the number 4 reflects the "number rule" which means the fragment must be found in at least 4 compounds. When the number rule is too small, it may risk inclusion of fragments that do not relate to certain biological activity because a large amount of fragments will be produced. On the other hand, the large value of this number would increase the chance of missing important features based on the diverse nature of the learning set. The numbers 0.9 and 0.85 are the "proportion rule", and this means the fragment should be found in at least 90% of ligands or 85% of non-ligands. The values of the "number rule" and "proportion rule" were estimated by the cat-SAR Rule Optimization routine. The Rule Optimization routine in our study allowed the "number rule" to range between 1 and 8 with increasing intervals of 1 and the "proportion rule" to range between 0.50 and 0.95 with increasing intervals of 0.05 [27]. We reasoned that even if a particular fragment is associated with a ligand, there may yet be other reasons for the compound from which it is derived to be classified as a ligand (e.g., other fragments or chemicophysical properties), thus it would not be expected to be found in 100% of the ligand. Likewise is true for fragments associated with non-ligand. Thus, if we considered only those fragments found exclusively in ligand or non-ligand we would rarify the fragments pool to an unreasonable level and risk losing valuable

information. On the other hand, we expected that fragments found to be presented approximately equal in the ligand and non-ligand fragment sets would not be associated with biological activity. Such fragments may serve as structural scaffolds holding the biologically features and are not directly related to activity or inactivity. It should be noted that the cat-SAR program uses a weight-of-evidence approach to select important fragments, rather than statistical analysis.

Predicting Activity

The resulting list of important fragments can be used to predict the activity of an unknown compound (compound that we do not know its activity). The approach compares the important fragments between learning sets and those in the unknown compound. If they have no common important fragments, no prediction of activity is made If there are common important fragments, cat-SAR can make a prediction for the compound with uncertain activity according to the percentage of activity of the common important fragments. The probability of activity or inactivity is then calculated based on the total number of ligand and non-ligand compounds containing the fragments.

To classify an unknown compound back to a ligand or non-ligand category, rather than a probability of activity, the program identifies an optimal cut-off point that is able to separate ligand from non-ligand based on the model validation analysis [28]. The compound predicted with a value larger than the cut-off value is considered to be a ligand; otherwise it is considered a non-ligand. The cut-off point can be adjusted according to the best overall concordance or the balance of sensitivity and specificity.

Model validation

Both internal and external validations were conducted for each model. For the internal validation, the leave-none-out (LNO), leave-one-out (LOO), and leave-many-out (LMO) validations were used. For the LNO, a model was developed from the complete learning set of 232 compounds and the model was used to predict the activity of each compound in the learning set. For the LOO validation, each chemical, one at a time, was removed from the model's learning set. The remaining n-1 compounds were used to build a n-1 model. The activity of the removed compound was then predicted by the n-1 model. Predicted *vs.* experimental values for each chemical were then compared and concordance, sensitivity, and specificity values were calculated. For the LMO validation, randomly selected 10% of the chemicals were removed from the learning set. Then the remaining compounds were used to develop the model. The activity of each removed chemical was predicted by the n-10% model. Predicted *vs.* experimental values for the removed chemicals were then compared, and the n-10% model's concordance, sensitivity, and specificity were determined. This was repeated 10,000 times and the average concordance, sensitivity, and specificity values were calculated.

The results of the LNO, LOO and LMO were expressed by three values; they are sensitivity, specificity and concordance. The equations for them are: Sensitivity = Correct positive predictions/Total positive predictions; Specificity = Correct negative predictions/Total negative perditions; Concordance = Correct predictions/Total predictions.

For the external validation, ten random sets of 10% of the chemicals in the learning set were removed, and the remaining 90% of the compounds were used to develop a model. The model was then used to predict the activity of those left out, and the average sensitivity, specificity, and concordance values of the 10 random sets were calculated. In contrast to the LMO, the external validation is more independent because it does not use any information from the testing set. However, the LMO validation uses information of the testing set to decide the cutoff point and the percentage of activity.

CHAPTER III

RESULTS AND DISCUSSION

Model summary

Together, three sets of cat-SAR models were derived from the NCTRER dataset: M^+, M^- and non-M. The Rule Optimization process was used to seek the best model based on the LOO validation. For M^+, M^- and non-M models, the best model's parameters were 4/0.9/0.85, 3/0.85/0.85 and 3/0.9/0.9 (Table 1), respectively. For the best models, they are not the models with the highest concordance value. Actually, the highest concordance value for M^+, M^- and non-M were 0.93, 0.92 and 0.89 respectively, and the parameters for them were 2/0.8/0.95, 8/0.9/0.85 and 5/0.9/0.95 respectively (Table 1). The reason was that the model with the highest concordance value did not satisfy other requirements for a best model such as the coverage, the balance between sensitivity and specificity or the validation results of LNO or LMO.

For the M^+ best model, 1,849 "important" fragments were created, among which 909 were fragments associate with ligands, and 940 were fragments associated with non-ligands. For M^- model, 2,386 important fragments were used, and 1,122 were fragments associated with ligand and 1,264 were non-ligand related. For this best model, M^+ has 535 fewer important fragments than M^-. This may be due to the parameter difference or the classification of the marginal chemicals. To further study this question, the important fragment number for

different parameters and classification of marginal chemicals were extracted (Table 2). As shown in Table 2, the classification of marginal chemicals has very little effect on the fragment number. When comparing M^+ and M^- models with the same parameters there was a difference of only 65 fragments for 3/0.85/0.85 and 28 fragments for 4/0.9/0.85. But if the classification of the marginal chemicals is the same, the change of parameters produces a bigger difference on fragment number. For instance, 4/0.9/0.85 has 612 fewer fragments than 3/0.85/0.85 for M^+ and 565 fewer for M^-. This shows that the difference in parameters is the main reason for the difference of important fragments. When the number rule increases from 3 to 4, more fragments will be ruled out. It is the same for the proportion rule. When it changes from 0.85/0.85 to 0.9/0.85, the number of qualified fragments will decrease. Our experimental data reflect this rule.

In order to investigate the performance of the M^+ and M^- models on marginal chemicals, the predictions of the eight marginal chemicals by the two sets were analyzed. Table 3 shows the prediction of the eight marginal chemicals. For the M^+ model, three of them were correctly predicted as ligand, three were incorrectly predicted as non-ligand, and two were unpredictable. For M^- model, five of them were correctly predicted as non-ligands, two of them were incorrectly predicted as ligands, and one was unpredicted (Table 3). The M^- model made more correct predictions on marginal chemicals than M^+ model. However, we can not conclude which model is better because neither this difference because neither the difference nor the sample size was big enough to make the conclusion.

LOO validation

LOO validation was performed before LNO and LMO validation because the best models were selected based on LOO validation. For the best model of M^+, sensitivity, specificity and concordance values were 0.91, 0.74 and 0.85, respectively (Table 4). For M^-, the sensitivity, specificity and concordance values were 0.90, 0.81 and 0.86, respectively. For non-M learning set, the sensitivity, specificity and concordance values were 0.92, 0.76 and 0.86, respectively. Comparing the LOO validation results of the best model of the three sets (M^+, M^- and non-M), the sensitivity and concordance values were very similar. This shows that the marginal chemicals did not greatly affect the performance of the model. Therefore, for the remainder of this study, we concentrated on the models from the M^+ and M^-.

LNO validation

After the best models were selected based on LOO, LNO validations were carried out using the parameters from the best models. Both M^+ and M^- models produced the same concordance value of 0.92 (Table 4). The sensitivity and specificity values for the M^+ were 0.96 and 0.84, while for the M^- they were 0.97 and 0.86. For the LNO, the performances of the best model from the two sets were very close to each other. LNO validation is also called self-fit validation. The characteristic of this kind of validation is that the model is developed from the whole learning set, and that model is used to predict the activity of each compound in the learning set.

LMO validation

The LMO validation yields a concordance value of 0.84 for both M^+ and M^-

models. The sensitivity values were 0.90 and 0.89, and specificity 0.75 and 0.77 for M^+ and M^-, respectively (Table 4). The validation results of M^+ and M^- were very close to each other. LMO is also a cross-validation. The vital aspect of this validation is that more than one chemical was taken out every time from the testing set. The selection of testing set was random. In this study, the process was repeated for 10,000 times.

In summary the LNO, LOO and LMO are all internal validations. Comparing the three values (sensitivity, specificity and concordance) of LNO, LOO and LMO, all the values are in the same order: LNO > LOO > LMO. This trend is reasonable because an increasing number of chemicals were removed from the learning set from LNO to LMO validation. LNO did not have any chemical removed, LOO had one removed each time, and LMO has 10% of the total chemicals removed each time. With more chemicals removed, less information is available to develop a model, and the difficulty of making a correct prediction increases. Therefore, it is reasonable that the value of the LMO validation was lower than LNO and LOO. If the results did not show this trend, for example, the LMO had better value than LNO or LOO, this means that a systemic mistake may exist.

External validation

Table 4 also shows the results of the external validation analysis. The average concordance values for M^+ and M^- are 0.82 and 0.80; the sensitivity values were 0.81 and 0.90, and the specificity values were 0.83 and 0.69. The concordance values were lower than the LMO validation concordance values. According to the external validation, the M^+ model is better than M^- because it is

19

more balanced and has a higher concordance value. For the external validation, the testing set is more independent from the training set compared to LMO. Therefore the results are more reliable. The values of sensitivity, specificity and concordance of the external validation indicate that the NCTRER cat-SAR models can identify the ER ligands.

The relationship of RBA and the prediction

To study the relationship of the model's LOO predicted values and the RBA values, 232 compounds were divided into five groups according to their RBA values. The range of the five groups was defined by the NCTRER database [26] (Table 5). The slight binders (what we called marginal compounds) group (0<RBA<1E-5) had the lowest percentage of correct predictions with three out of eight predictions being correct, which was followed by the non-ligand group in which 80% had been predicted correctly. This was a relatively lower percentage of correct predictions, especially when compared to the ligand compounds, which is over 90%. The weak, medium and strong ligand groups had similar percentages of correct predictions with values of 94, 92, and 93%, respectively. To explore the cause of this discrepancy, the structure of the compounds incorrectly predicted in the lower percentage groups were investigated. There were 19 incorrectly predicted compounds in the non-ligand group, and all of them had an aromatic ring and 11 (or 57%) of them contained a phenolic ring. For the slight binder, or marginal group, all four compounds predicted incorrectly had aromatic rings and none of them had a phenolic ring. As shown previously, the aromatic or phenolic ring is an important biophore for estrogenicity and makes them very similar to

ligand compounds that actively bind to the ER [23]. If a compound has an aromatic or phenolic ring but is not estrogenic, this makes it more likely to be incorrectly predicted as a ligand compound. Based on this analysis, it can also be concluded that aromatic ring or phenolic ring are not sufficient to determine estrogenicity because the non-ligands can also have them. There should be other structures that contribute to the binding activity. On the other hand, this also shows the predictions of our cat-SAR expert system were based on the structure of the chemicals. This should be studied further in the future.

The relationship of different chemical classes and model prediction

To further study the relationship between prediction accuracy and the chemical categories, the chemicals were sorted by their chemical categories according to the NCTRER database classification. The uneven distribution of correct predictions was found among different chemical categories for M^+ and M^-. Table 6 shows the result for M^+ and M^-, where M^- had similar results as M^+. The order for percentage of correct predictions is: biphenyls < diphenylmethane < phytoestrogen < phenols < DES < steroids. According to the results, the biphenyls group had the lowest percentage of correct predictions, and the steroid group had the highest. One reason for the difference in correct predictions may be that the sample size for biphenyls is too small. The group only has 12 chemicals in total, and one is not predicted. Among the 11 predicted chemicals, four of them were incorrectly predicted, and seven of them were correctly predicted. The small sample size makes the results easier to be skewed. For other groups, the sample size is much larger than the biphenyls. Another possible explanation for this

group's low percentage of correct predictions is that the non-ligand chemicals in this group all have aromatic ring. For biphenyls, there are three aromatic non-ligands out of 12, and all of them are incorrectly predicted. But the steroids, which have a high percentage of correct predictions, do not have any aromatic non-ligands. Actually the non-ligand aromatic chemicals decreased the performance of the model because they have structural alerts similar to ligands for the ER. This may lead the model to predict them as ligands. This agrees with what was discovered by RBA classification on aromatic non-ligand chemicals. Furthermore, according to the structure of biphenyls, they do not have a hydrophobic center. This may also contribute to its low performance because hydrophobic center is one of the important structures related to ER binding. According to our investigation, all the good ER binders have an ideal hydrophobic center. The ideal hydrophobic center means the proper size with enough hydrophobicity. Therefore, the hydrophobicity can also be used to make a prediction on binding activity. Comparison of compounds in the biphenyl group to those in the diphenylmethane group demonstrates that they all have two aromatic rings. The difference is that the diphenylmethane have a hydrophobic center, but biphenyls do not. Diphenylmethanes also have non-ligand chemicals with aromatic rings. There were nine such chemicals out of 28 chemicals compared to biphenyls, which had three out of 12, and six of them were incorrectly predicted. Although diphenylmethane compounds also have non-ligand chemicals with aromatic ring, it has a higher percentage of correct prediction than biphenyls. This is due to its hydrophobic center which helped the cat-SAR expert system make a

22

correct prediction on the ER ligands. The uneven performance on different chemical classes shows that its prediction is based on the structure analysis, and important fragments play a vital role in deciding a chemical's activity. It also reminds us that having a bigger sample size and similar number of chemicals of different chemical classes may improve the performance of the models. Furthermore, making the application domain more specific to a chemical class (i.e. biphenyls or steroids) may be a better way to improve the model's performance. Moreover when these models are used to assess the ligand binding potential for a new or untested chemical, it is likely that chemical class and the compound's true potency will affect the reliability of the prediction.

Examples of cat-SAR predictions

In order to investigate model prediction for compounds, four chemicals of different activity were chosen to demonstrate the process. They were 17β-estradiol, coumestrol, BPA and progesterone whose RBA values were 100, 0.9, 0.008 and 0, respectively. They are representatives for the strong, medium and weak ligands, and the non-ligand categories.

As demonstrated by the activity and fragment information of 17β-estradiol, it was correctly predicted as a ligand compound by M^+ and M^- (Table 7). Figure 1 shows the structure of 17β-estradiol and some of the fragments created by two sets. In order to perform the analysis, the fragments were divided into three sections: section 1 specifically covered the 3-OH group and the affiliated aromatic ring A, and section 2 covered the interior B and C rings, and section 3 specifically covered the 17-OH group and the affiliated ring D.

Fragments from section 1 identify the aromatic ring A and the affiliated 3-OH group, which have been identified as two of the most important structural alerts for 17β-estradiol [29]. It contributes about 1.5kcal/mol to the total binding energy, which is 12kcal/mol [30]. If the aromatic ring is replaced with other rings, the binding affinity will decrease dramatically. For example, if 17β-estradiol's aromatic ring is replaced with cyclohexane, it becomes 3α-androstanediol (Table 8), and its RBA value decreases to 0.002 [31]. In fact, 3α-androstanediol cannot be predicted by either of our models. Due to its lack of an aromatic ring, its fragments could not be found in the important fragments list of either model. In our study, all the fragments of 17β-estradiol contain part of the aromatic ring. All of this information indicates that the model identified this structure and used it to make a prediction on ligands or non-ligands.

To further study the effects of the aromatic ring on model predictions, the non-aromatic chemicals were taken out to study independently. Among 232 compounds, there are 28 compounds that do not have an aromatic ring. For 28 non-aromatic compounds, five of them are ligands, and 23 are non-ligands. Both models correctly predicted all 23 non-ligands, but for the five non-aromatic ligands, three were incorrectly predicted, and two were not predicted in both models. Thus, there were no correct predictions for the five non-aromatic ligands. This again suggests that the aromatic ring is an important structure for ER binding, and the model expert system uses it to make a correct prediction for ER ligands.

Section 1 also includes the structure 3-OH. The 3-OH group contributes approximately 1.9kcal/mol of binding energy as a hydrogen bond donor [30]. It

forms a hydrogen bond with the Glu 353, Arg 394, and a water molecule [31, 32]. When the 3-OH group is removed, the RBA value drops significantly when comparing similar compounds (e.g. 3-deoxy-estradiol). In the fragment list of 17β-estradiol, we found four fragments that contained the 3-OH group, and all of them were mostly found in the fragment list of ligands. This shows that our models can identify the 3-OH group as a structural alert for ER ligands.

The fragments from section 2 are derived from the hydrophobic center for 17β-estradiol [30]. The ligand binding domain (LBD) of ER is a hydrophobic pocket, which creates a favorable environment for binding ligands that possess a hydrophobic center [33]. Another important aspect of section 2 in relation to 17β-estrodiol is that it creates a favorable distance between the 3-OH and 17-OH group. The distance between these two hydroxyl groups is d_{o-o}. It is a factor that affects the binding affinity of a ligand [31]. Either too large or too small d_{o-o} is not favorable for ER binding, and a certain range of d_{o-o} can make the binding more stable. For example, the d_{o-o} for 17β-estrodiol is 11.0 Å, which allows the 3-OH and 17-OH to appropriately align with the ER binding pocket and form hydrogen bonds, which creates a much stronger interaction between the receptor and the ligand, thus increasing the binding affinity. The compounds that do not have a steroidal backbone usually have very low RBA value, even though they have an aromatic ring or a –OH group, such as the phenols, and biphenyls. The reason for this is that the steroidal backbone offers the favorable conformation for ligand binding [31]. This explains why the average RBA value of steroid compounds is higher than biphenyls because the steroids have the hydrophobic center and the

d_{o-o} value is similar to the natural estrogen.

In the fragment list, there are 253 fragments that cover this section. Among the three sections, this section has the largest number of fragments because this section also includes parts of the aromatic ring. There is no important fragment that does not have any part of the aromatic ring. Therefore, we can conclude that only the hydrophobic fragments are not enough for receptor binding. It has to be coupled with an aromatic ring to form a ligand. This may explain why fragments of section 2 all have some part of the aromatic ring.

The fragments from section 3 identify the 17-OH group, which contributes about 0.6kcal/mol as a hydrogen bond acceptor [29]. The 17-OH group can form an H-bond with His 524 [31]. If the 17-OH group was removed, the RBA would decrease. For example, the 17-deoxy-estradiol has a RBA value of 14.1 which is seven times lower than 17β-estradiol. Although 17-deoxy-estradiol does not have the 17-OH group, it was correctly predicted as a ligand in our model (Table 8). This suggests that 17-OH is important, but it is not imperative for activity. The M^+ and M^- models have the same fragment for section 3, which is a 17-OH group affiliated with a four ring skeleton, and it is the only fragment found in this section. These findings demonstrate that 17-OH is a structural alert for ER ligands, but by itself is not enough to greatly affect ligand's binding ability to the ER. In fact, it seems that 17-OH needs to be combined with hydrophobic fragments and aromatic fragments to construct a complete ligand. Therefore, the 17-OH is not a necessary structure for the ER ligand, but it is a structure that can increase the binding affinity of a compound to the ER.

26

Coumestrol is a phytoestrogen and another ER ligand in the learning set. Coumestrol has three aromatic rings, two at the ends of the compound and one in the middle. The chemical shape of coumestrol orients its two hydroxyl groups in the same position as the two hydroxyl groups in 17β-estradiol, allowing it to mimic the structure confirmation of 17β-estradiol. However, The RBA value for coumestrol is 0.90, which is about 100 times lower than 17β-estradiol. It was predicted as a ligand compound by M^+ and M^-. The M^+ and M^- models created eight and ten fragments in total, respectively, and all of them described the phenol ring. Five representative fragments from both sets are shown in Figure 2. From the fragments of coumestrol, we found that all of the important fragments describe the aromatic rings, either ring A or D. Comparison of coumestrol to 17β-estradiol suggests that they have many similarities. They both have four rings and two –OH groups at the A ring and D ring, which may explain why coumestrol binds to the ER. However, coumestrol's RBA value is much lower than 17β-estradiol, which means coumestrol is a relatively weaker binder compared to 17β-estradiol. This may due to the fact that coumestrol's hydrophobic center is weaker than 17β-estradiol because it has three oxygen atoms, which makes it more hydrophilic [34]. For the important fragments of coumestrol, none can represent the hydrophobic center. This study shows the important fragments created by cat-SAR were related to the ER binding activity.

Bisphenol A is used to make polycarbonate plastic and epoxy resins, along with other applications. There is concern that the wide daily use of it may be related to some potential negative health effects [35-36]. It was predicted as a

27

ligand by both M^+ and M^- models. The important fragments of BPA are shown in Figure 3. These fragments describe three parts: the –OH group, the aromatic ring and the bridge hydrocarbons. All of the fragments are related to the activity of BPA. The bridge hydrocarbons act in a similar manner as the fragments from section 2 for 17β-estrodial by forming a hydrophobic center. However, this center in BPA is much smaller than in 17β-estradiol. Therefore, the d_{o-o} of BPA is shorter than 17β-estradiol, and may explain why BPA's RBA value is much smaller than 17β-estradiol.

Progesterone, a vital hormone for pregnancy, is a non-ligand in both M^+ and M^- learning sets and was correctly predicted by both models. In the M^+ model, it had 285 fragments in total, and 52 of them had an oxygen atom. All of the oxygen atoms in those fragments were linked by a double bond, which prevents the compound from forming a hydrogen bond with the ER. Also, this compound has no aromatic rings. The lack of these structural features contributes to progesterone's inactivity. Figure 4 shows some of the representative fragments of progesterone and represents the differences in important fragments associated with ligands and non-ligands.

From the above examples, we get a better understanding of what type of fragments are most likely to construct a ligand or non-ligand compound. Therefore, the cat-SAR expert system can not only make predictions for unknown compounds, but also assist in the analysis of identifying the binding mechanism and, therefore, potentially help in designing new compounds with specific characteristics.

CHAPTER IV

SUMMARY AND CONCLUSION

Overall, the present study demonstrated the utility of SAR modeling for xenoestrogens screening and the ER binding mechanism study. The cat-SAR expert system is a qualitative SAR, and its predictions are based on the activity of the fragments derived from each chemical, and the predicted activity of each compound is calculated by the frequency with which each fragment is found in either ligands or non-ligands. Therefore, the learning set always consists of a certain number of ligand and non-ligand chemicals as the NCTRER database used in this study.

NCTRER is a unique database for analyzing xenoestrogens and ER binding, since it contains a diverse set of chemicals with a wide range of binding affinities. The database includes 232 compounds, of which 131 are ligands, 93 are non-ligands and 8 are marginally binding chemicals. In total, 37 descriptors were listed for each chemical, including their RBA value, their chemical class, and the number of aromatic rings each chemical possesses, among others. In this study these descriptors were used to understand how they affect ligand binding to the ER. For example, the relationship of why different chemical classes bind to the ER with varying degrees of RBA was analyzed. Although the role of just a few descriptors was investigated in this study, it is believed that the cat-SAR expert system will be very useful in investigating the remainder of the descriptors

found in the NCTRER database. Four different validation methods were used in this investigation: LNO, LOO, LMO and external validation. The sensitivity, specificity and concordance value were calculated for each of the validations. The concordance values were all above 80%, meaning the cat-SAR methodology is capable of making correct predictions for ER ligands.

The overall performance for the model was very good, but a discrepancy was identified in the overall correct predictive rate among six different chemical classes, including biphenyls, diphenylmethane, phytoestrogen, phenols, DES and steroids. For example, the percentage of correct predictions for biphenyls was only 63%, but it was 93% for steroids, according to the M^+ model. This indicates that the closer a chemical's structure to the natural estrogen the higer possibility of it being predicted correctly by the model. This also suggests that the models have different predictive performances on different chemical classes (i.e., chemical structure or potency). Therefore, when these models are used to assess the ligand binding potential for a new or untested chemical, it is likely that chemical class and the compounds true potency will affect the reliability of the prediction.

Important fragments were identified by the cat-SAR expert system for each chemical. The important fragments for the ligands covered most of the existing biophores for ER binding such as the aromatic ring, 3-OH group, 17-OH group and the hydrophobic center. But most of non-ligands do not contain the fragments that were known as structures alerts related to ER binding. This shows the cat-SAR expert system is mechanistically sound and can be used to carry out

mechanism analysis in the future. Meanwhile, there are some fragments that can not be explained by the exiting structural alerts; this offers the possibility of discovering new biophores.

This study identified some important fragments for ER binding. However, it is far from completed because we found some chemicals that do not have structure alerts, but are ligands. For instance, we found that there are 25 chemicals with the phenolic ring but were non-ligands, and 14 of them were correctly predicted by the M^+ model. This demonstrates that there should be other structures that define an estrogenic chemical besides the phenolic ring. Even when considering the phenolic ring, the location of the –OH group critically affects the activity of the chemicals. These are examples of chemical structural analysis that could be part of our future studies.

In conclusion, the NCTRER cat-SAR ER binding model is a reliable model for xenoestrogen identification. The cat-SAR expert system identified important fragments for ER binding that help explain why certain xenoestrogens bind to the ER better than others. Therefore, this model can be used to do xenoestrogen screening and potentially identify how well compounds will bind to the ER, strictly based on their chemical structure. Furthermore, future studies may lead to the discovery of other possible structural alerts for ER binding or estrogenicity. Understanding these structural characteristics may lead to a better mechanism for dealing with the carcinogenic and anti-carcinogenic properties associated with these xenoestrogens. Making use of this model, or combining it with other models, could explore this question in a meaningful direction.

Table 1. M$^+$, M$^-$ and non-M Model comparison

		Parameters	Sen[1]	Spe[2]	Ocp[3]	Cutoff	Coverage
M$^+$	Best Model	4/0.9/0.85	0.91	0.74	0.85	0.90	0.86
	Highest Ocp	2/0.8/0.95	0.93	0.86	0.91	0.83	0.97
M$^-$	Best Model	3/0.85/0.85	0.90	0.81	0.86	0.86	0.91
	Highest Ocp	8/0.9/0.85	0.92	0.91	0.92	0.92	0.61
Non-M	Best Model	3/0.9/0.9	0.89	0.76	0.86	0.89	0.90
	Highest Ocp	5/0.9/0.95	0.96	0.77	0.89	0.90	0.84

Note: 1. Sen is the abbreviation for Sensitivity.

2. Spe is the abbreviation for Specificity.

3. Ocp is the abbreviation for observed correction prediction, and equals to concordance value.

The Sensitivity, Specificity and Concordance value is based on the LOO validation.

TABLES

Table 2. Fragments table

Model	Parameter	Total Fragments	Ligand Fragments	Non-ligand Fragments
M^+	3/0.85/0.85	2461	1214	1247
M^-	3/0.85/0.85	2386	1122	1264
M^+	4/0.90/.85	1849	909	940
M^-	4/0.9/0.85	1821	863	958

This table compares the number of fragments of the different classification of marginal chemicals and parameters. Both the classification of marginal chemicals and the parameter affect the number of fragments. M^+ model has more fragments than M^- with the same parameter. For the same classification of marginal chemicals, the fragments increase with the values of parameter increases. The changing of parameter has more effects on fragments number than the classification of the marginal chemicals.

Table 3. Marginal binding chemicals prediction.

CAS	M^+	M^-
117-81-7	incorrect	correct
2132-70-9	correct	correct
32598-13-3	unpredictable	incorrect
3424-82-6	correct	unpredictable
53-19-0	incorrect	correct
6554-98-9	correct	incorrect
85-68-7	incorrect	correct
90-00-6	unpredictable	correct

Eight marginal chemicals and their predicted results from M^+ and M^- were listed.

Table 4. Validation summary for cat-SAR ER binding models

Validation	Model	Sensitivity[1]	Specificity[2]	Concordance[3]
LNO	**M⁺**	0.96(121/126)	0.84(58/69)	0.92(179/195)
	M⁻	0.97(121/125)[1]	0.86(73/85)	0.92(194/210)
LOO	**M⁺**	0.91(115/127)	0.74(54/73)	0.85(169/200)
	M⁻	0.90(111/124)	0.81(70/87)	0.86(181/211)
	non-M	0.92(113/123)	0.76(59/78)	0.86(172/201)
LMO	**M⁺**	0.90(10.0/11.1)	0.75(5.0/6.7)	0.84(15.0/17.8)
	M⁻	0.89(10.5/11.8)	0.77(6.3/8.2)	0.84(16.8/20.0)
External	**M⁺**	0.81(107/132)	0.83(63/76)	0.82(170/208)
validation	**M⁻**	0.90(106/120)	0.69(64/92)	0.80(170/212)

Notes: 1. Number of correct positive predictions / total number of positives;
2. Number of correct negative predictions / total number of negatives;
3. Observed Correct Predictions: number of correct predictions / total number of predictions

The table shows the validation results for LNO, LOO, LMO and external validation for M⁺ and M ⁻ models, and also the LOO for the Non-M model. Sensitivity, specificity and concordance values were calculated and listed in the table. For LMO, the number of chemicals in parenthesis is the average of 10% removed chemicals.

TABLES

Table 5. Distribution of incorrect predictions for M^+ model

Compounds group	Total compounds number	Unpredictable compounds number	M^+	
			Number of correct predictions	Percentage of correct predictions
Non-ligand (RBA=0)	93	0	74	80%
Slight binder (0<RBA≤1E-5)	8	2	3	50%
weak Ligand (1E-5<RBA≤0.01)	61	9	49	94%
medium Ligand (0.01<RBA≤1)	41	1	37	92%
strong Ligand (RBA>1)	29	0	27	93%

The compounds were divided into five groups according to their RBA value. The number of correct predictions and percentage of correct perditions were calculated.

Table 6. Prediction on different chemical classes

Ligands Categories	Number of Compound[1]	Number of Correct Prediction	Percentage of correct Predictions
Biphenyls	11	7	63%
Diphenymethane	28	21	75%
Phytoestrogen	44	35	80%
Phenols	27	24	89%
DES	22	20	91%
Steroids	29	27	93%

Notes: 1. Number of compound does not include the unpredicted compounds.

This table shows the performance of cat-SAR ER binding model of M^+ on different chemical classes. The number and percentage of correct predictions were calculated.
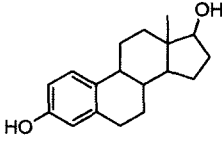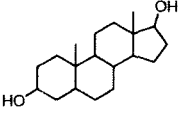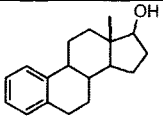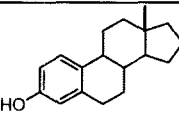
Table 7. Activity and fragment information for 17β-estradiol

Model	Experimental activity	Predicted percentage of activity	Fragments number
M^+ (4/0.9/0.85)	ligand	97%	258
M^- (3/0.85/0.85)	ligand	95%	281

The activity and fragment information for 17β-estradiol in both models.

TABLES

Table 8. Structure and activity comparison for steroids

	Structure	RBA	Predicted activity
17β–estradiol RBA=100	3α -androstanediol	0.002	Unpredictable
	3-Deoxy-estradiol	0.5	Ligand
	17-Deoxy -estradiol	14.1	Ligand

The effects of 3-OH, 17-OH group and aromatic ring on RBA value were shown. The predictions of these three chemicals were also listed. The aromatic ring has the biggest effect on activity then is the 3-OH group, then the 17-OH group.

FIGURES
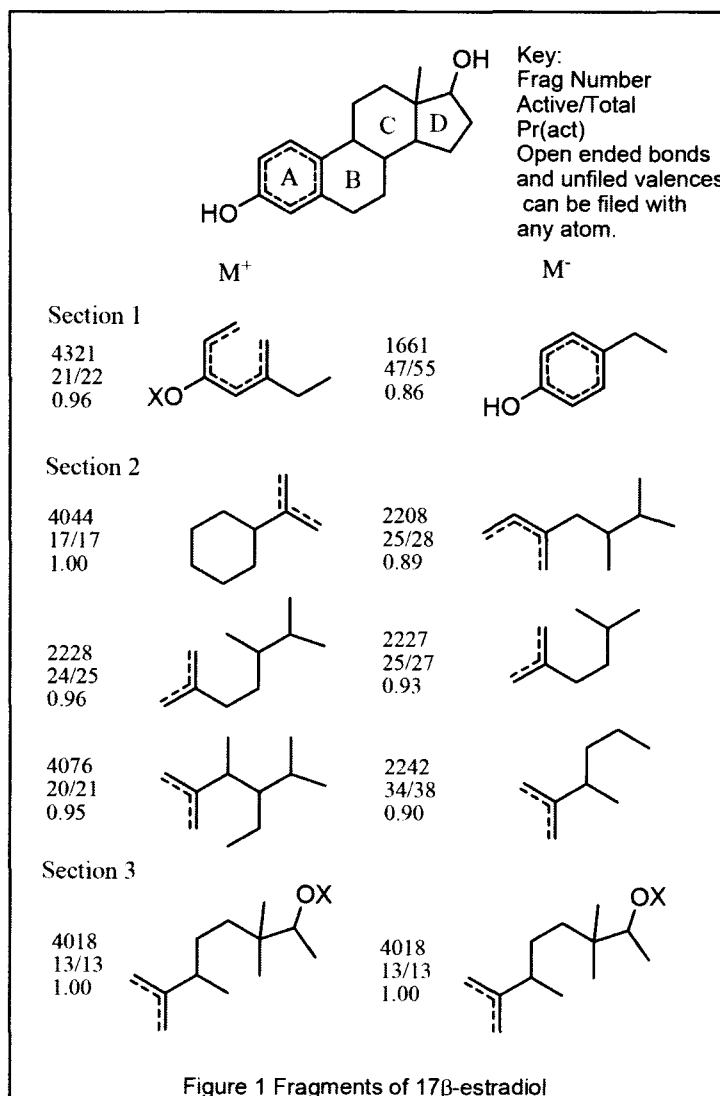

Figure 1 Fragments of 17β-estradiol

Figure 1. Important fragments of 17β-estradiol. All the important fragments can be divided into three sections. Each section represents certain part of the chemical.
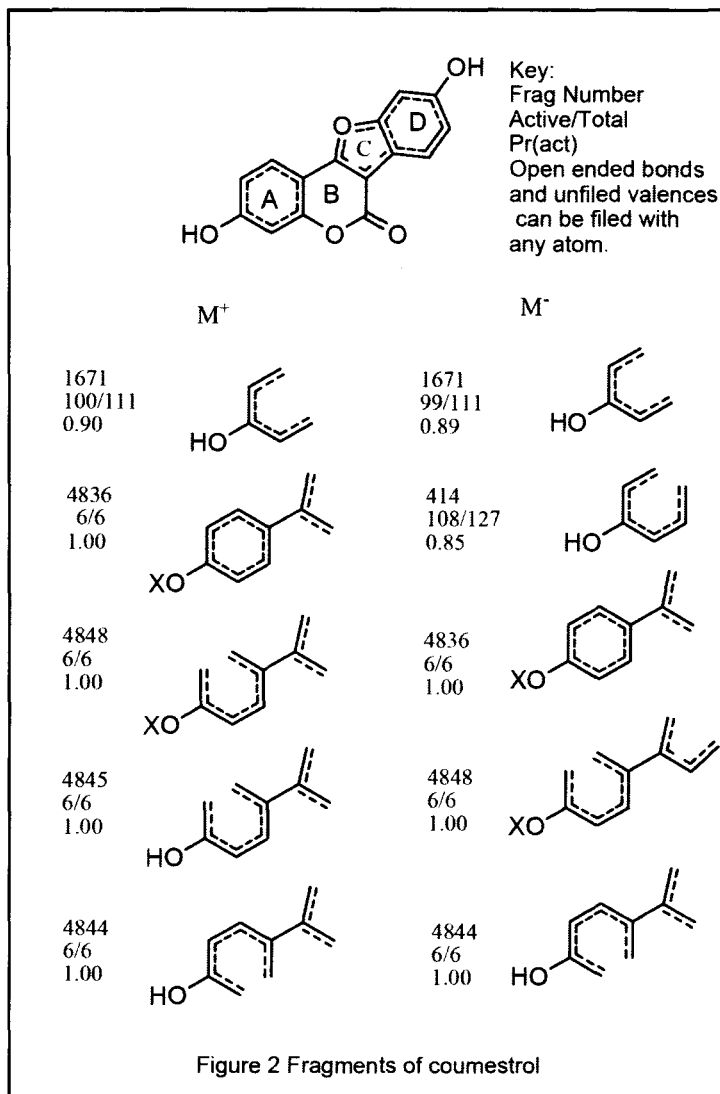
40

Figure 2 Fragments of coumestrol

Figure 2. Important fragments of coumestrol. All the important fragments contains part of an aromatic ring and a –OH group.
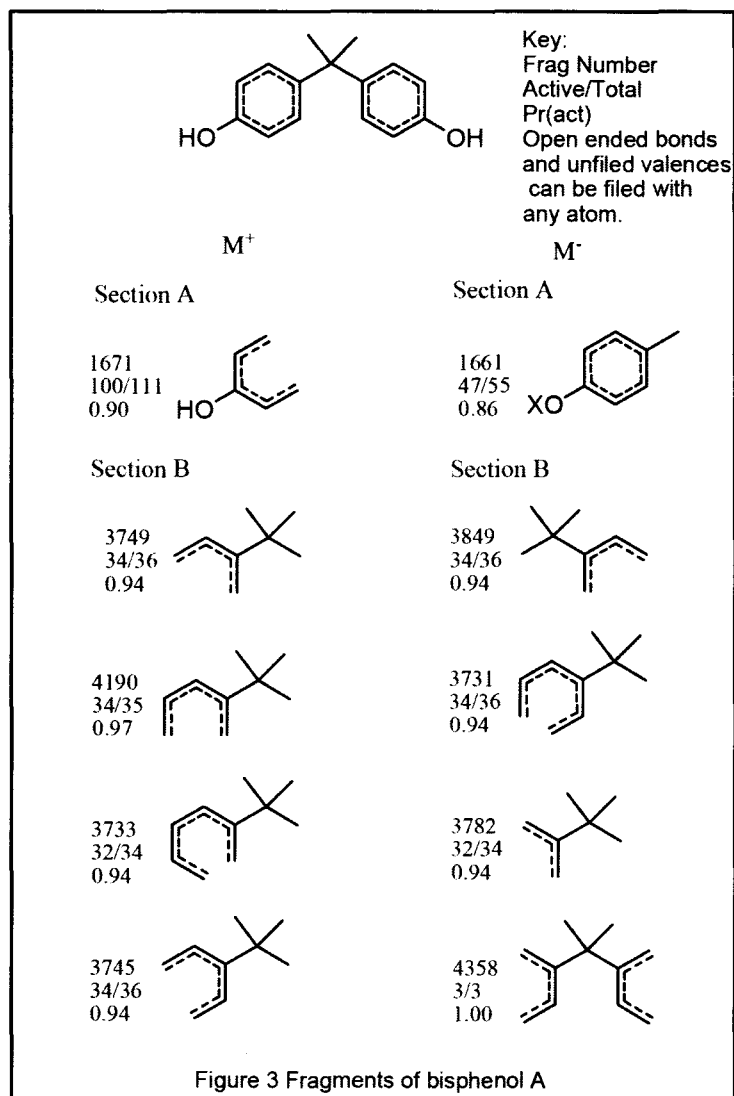
Figure 3 Fragments of bisphenol A

Figure 3. Important fragments of bisphenol A. The fragments were divided into two parts: section A and section B. Section A is the phenolic ring, and section B is the aromatic ring and the hydrophobic center.
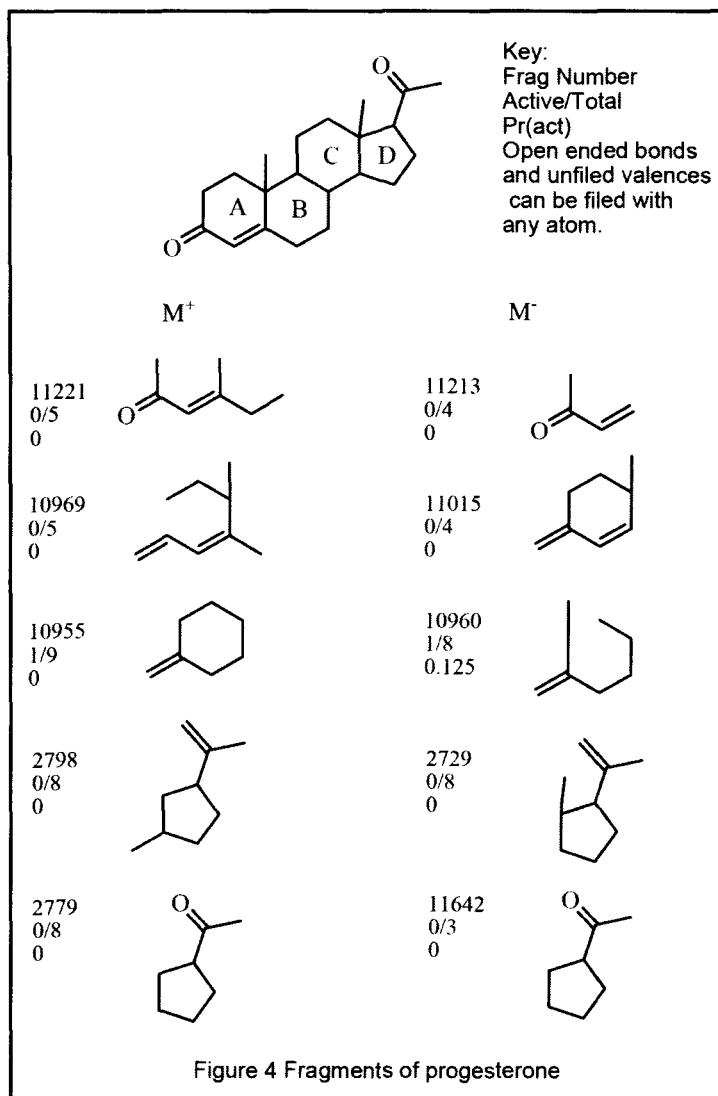
Figure 4 Fragments of progesterone

Figure 4. Important fragments of progesterone. The percentage of activity of the fragments was "0". No fragment that related to any biophores for ER binding was found in the fragment list.

REFERENCE

1. Revankar CM, Cimino DF, Sklar LA, Arterburn JB, Prossnitz ER (2005). A transmembrane intracellular estrogen receptor mediates rapid cell signaling. Science 307 (5715): 1625–30.

2. Filardo EJ, Thomas P (2005). GPR30: a seven-transmembrane-spanning estrogen receptor that triggers EGF release. Trends Endocrinol. Metab. 16 (8): 362–7.

3. Manavathi B, Kumar R (2006). Steering estrogen signals from the plasma membrane to the nucleus: two sides of the coin. J. Cell. Physiol. 207 (3): 594–604.

4. Prossnitz ER, Arterburn JB, Sklar LA (2007). GPR30: A G protein-coupled receptor for estrogen. Mol. Cell. Endocrinol. 265-266: 138–42.

5. MacGregor JI, Jordan VC (1998). Basic Guide to the Mechanisms of Antiestrogen Action. Pharmacological Reviews, 50(2): 151-196.

6. Carmeci C, Thompson DA, Ring HZ, Francke U, Weigel RJ (1997). Identification of a gene (GPR30) with homology to the G-protein-coupled receptor superfamily associated with estrogen receptor expression in breast cancer. Genomics. 45:607-17.

7. http://atlasgeneticsoncology.org/Genes/GPERID44344ch7p22.html

8. K. Moriarty, K.H. Kim, J.R. Bender (2006). Estrogen receptor-Mediated Rapid Signaling. Endocrinology. 147(12):5557-5563.

9. Crisp TM, Clegg ED, Cooper RL, Wood WP, Anderson DG, Baetcke KP, Hoffmann JL (1998). Environmental endocrine disruption: An effects assessment and analysis. Environ. Health Perspect.106 (Supplement 1): 11-56.

10. Diamanti-Kandarakis E, Bourguignon JP, Giudice LC, Hauser R, Prins GS, Soto AM, Zoeller RT, Gore AC (2009). Endocrine-disrupting chemicals: an Endocrine Society Scientific Statement. Endocr. Rev. 30 (4): 293–342.

11. http://en.wikipedia.org/wiki/Endocrine_disruptor

12. Weschler CJ (2009). Changes in indoor pollutants since the 1950s.

Atmospheric Environment. 43 (1): 153–169.

13. Laws SC, Carey SA, Ferrell JM, et al (2000). Estrogenic activity of octyphenol nonlphenol, nonlphenol, bisphenol A and methoxychlor in rats. Toxicological Sciences. 54 (1): 154-167

14. FRY, DM; Toone, CK (1981). DDT-induced feminization of gull embryos. Science. 213(4510): 922-924.

15. Singleton, DW; Feng, YX; Chen, YD; Busch, SJ; Lee, AV; Puga, A; Khan, AS (2004). Bisphenol-A and estradiol exert novel gene regulation in human MCF-7 derived breast cancer cells. Molecular and Cellular Endocrinology. 221(1-2): 47-55.

16. Steinmetz, R; Young, PCM (1996). Novel estrogenic action of the pesticide residue beta-hexachlorocyclohexane in human mammary cancer cells. Cancer Research. 56(23): 5403-5409.

17. Rossouw, JE; Anderson, GL, Prentice, RL (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women- principal results from the women's health initiative randomized controlled trial. JAMA-Journal of the American Medical Association. 288(3): 321-333.

18. Knight, DC, Eden, JA (1996). A review of the clinical effects of phytoestrogens. Obstetrics and Gynecology. 87(5): 897-904.

19. EPA, Priority-Setting in the Endocrine Disruptor Screening Program (EDSP)-Background. Washington, DC: Environmental Protection Agency, (2002), Available at http://www.epa.gov/endo/pubs/prioritysetting/index.htm.

20. Gilles Klopman, Suman K. Chakravarti (2003). Structure-activity relationship study of a diverse set of estrogen receptor ligands (I) using MultiCASE expert system. Chemosphere. 51(6): 445-459

21. Waller, CL; Oprea, TI; Chae, K (1996). Ligand-based identification of environmental estrogens. Chemical ResearchIn Toxicology. 9(8): 1240-1248.

22. Qamar Shahid, Carrasquer C. Alex (2011). Anticancer SAR models for MCF-7 and MDA-MB-231 breast cell lines. Anticancer Research. 31(10): 3247-3252.

23. Blair, R.M., H. Fang, W.S. Branham, B.S. Hass, S.L. Dial, C.L (2000). The estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands. Toxicol. Sci. 54(1):138-153.

24. W.S. Branham, S.L. Dial, C.L. Moland, B.S. Hass, R.M. Blair, H. Fang, L. Shi, W. Tong, R.G. Perkins, and D.M (2002). Sheehan. Binding of phytoestrogens and mycoestrogens to the rat uterine estrogen receptor. J. Nutr. 132(4):658-664.

25. http://www.epa.gov/ncct/dsstox/sdf_nctrer.html#DownloadTable

26. D. R. Lowis (1997), HQSAR: A new, highly predictive QSAR technique. Available at http://www.tripos.com/data/SYBYL/HQSAR Application Note 072605.pdf.

27. Cunningham AR, Qamar S, Carrasquer CA, et al (2010). Mammary carcinogen-protein binding potentials: novel and biologically relevant structure-activity relationship model descriptors. SAR and QSAR in Environmental Research. 21(5-6): 463-479.

28. A. R. Cunningham, H. S. Rosenkranz, and G. Klopman (1998). Identification of structural features and associated mechanisms of action for carcinogens in rats. Mutat. Res. 405(1): 9-28.

29. J. Devillers, N. Marchand-Geneste (2006). SAR and QSAR modeling of endocrine disruptors. SAR and QSAR in Environmental Research. 17(4): 393-412.

30. Gregory M. Anstead, Kathryn E. Carlson, John A. Katzenellenbogen (1997). The estradiol pharmacophore ligand structure-estrogen receptor binding affinity relationship and a binding affinity relationships and a model for the receptor binding site. Steroids. 62(3): 268-303.

31. Hong Fang, weida Tong, Leming M. Shi, Robert Blair, Roger Perkins, et al (2001). Structure-Activity Relationships for a large diverse set of natural, synthetic, and environmental estrogens. Chem. Res. Toxicol. 14(3): 280-294.

32. Debashis Ghosh, Vladimir Z Pletnev, Dao-Wei Zhu, et al (1995). Structure of human estrogenic 17-beta-hydroxysteroid dehydrogenase at 2.20 angstrom resolution. Structure. 3(5): 503-513.

33. Bohl M (1992). Molecular structure and biological activity of steroids. CRC Press, Boca Raton, Florida, P 144.

34. Albert R. Cunningham, Gilles Klopman, Herbert S. Rosenkranz (1997). A dichotomy in lipophilicity of natural estrogens, xenoestrogens, and phytoestrogens. Environmental Health Perspectives. 105(supplement 3): 665-668.

35. Okada H, Tokunaga T, Liu X, Takayanagi S, Matsushima A, Shimohigashi Y (2008). Direct evidence revealing structural elements essential for the high

binding ability of bisphenol A to human estrogen-related receptor-gamma. Environ. Health Perspect. 116 (1): 32–38.

36. Vom Saal FS, Myers JP (2008). Bisphenol A and Risk of Metabolic Disorders. JAMA. 300 (11): 1353-1355.

37. Thorsten Naumann; David Lowis. A new, highly predictive QSAR technique. First International Electronic Conference on Synthetic Organic Chemistry (ECSOC-1), www.mdpi.org/ecsoc/, September 1-30, 1997.

38. Soto AM, Sonnenschein C, Chung KL, Fernadez MF, Olea N, Serrano FO (1995). The E-SCREEN assay as a tool to identify estrogens: an update on estrogenic environmental pollutants. Environ Health Perspect. 103 Suppl 7:113-22.

39. http://www.epa.gov/endo/pubs/edmvac/uterotrophic_story_4_1_05.pdf

41. http://en.wikipedia.org/wiki/Xenoestrogen

42. http://tripos.com/index.php?family=modules,SimplePage,,,&page=SYBYL-X

43 Wu, Y; Wang, Y; Zhang, AQ; Yu, HX ; Wang, LS (2010). Three-Dimensional Quantitative Structure-Activity Relationships of flavonoids and estrogen receptors based on docking. Chinese Science Bulletin. 55(15): 1488-1494.

APPENDIX

List of Abbreviations

NCTRER   National Center for Toxicological Research Estrogen Receptor

RBA   Relative Binding Affinity

ER   Estrogen Receptor

MultiCASE   Multiple Computer Automated Statistical Evaluation Expert
            System

QSAR   Quantitative Structure Activity Relationship

SAR   Structure Activity Relationship

Cat-SAR   Categorical- Structure Activity Relationship

CoMFA   Comparative Molecular Field Analysis

GPCR   G Protein-Coupled Receptor

ED   Endocrine Disruptor

PCB   Polychlorinated Biphenyls

DDT   Dichlorodiphenyltrichloroethane

BPA   Bisphenol A

LNO   Leave-none-out

LOO   Leave-one-out

LMO   Leave-many-out

CURRICULUM VITAE

NAME: Huihui Wu

ADDRESS: 10003 John Silver Ct. Louisville, Ky 40229

DOB: May 29, 1979

EDUCATION AND TRANING:

Graduate student   University of Louisville, 2009-present

M.S.   Shanghai University of Traditional Chinese Medicine,   2003 - 2006

B.S.   Anhui University of Traditional Chinese Medicine,   1998 - 2003

AWARDS:

1. Third Prize of Excellent Thesis, Shanghai University of Traditional Chinese

    Medicine, 2006

PUBLICATIONS:

1. Mao, Hui-Juan; Wu, Hui-Hui *et. al.* Relationship between electroacupunture's adjustment on leukocyte and the change of spleen ultrastructure. Journal of Shanghai University of Traditional Chinese Medicine. 2006, 20:67-69.

2. Wu, Hui-Hui; Mao, Hui-Juan *et. al.* Relationship between electroacupunture's anti-inflammation effect and spleen bloodflow. Shanghai Journal of Acupuncture and Moxibustion. 2006, *25*: 43-44.

3. Sun, Ping-Long ;   Zhou, Yu-Bao ;   Mao, Hui-Juan;   Wu, Hui-Hui *et.   al.* Relationship between electroacupuncture and spleen function on leukocyte. Journal of Acupuncture and Tuina Science. 2007, *5*:336-340.

4. Wu, Hui-Hui. Social Ability Evaluation, Chapter 16th in Rehabilitation Evaluation (Edited by Zhu, Yi-Hui), 2007, Shanghai Scientific and Technological Education Press.