KUOPION YLIOPISTON JULKAISUJA C. LUONNONTIETEET JA YMPÄRISTÖTIETEET 191
KUOPIO UNIVERSITY PUBLICATIONS C. NATURAL AND ENVIRONMENTAL SCIENCES 191

ARJA ASIKAINEN

Use of Computational Tools for Rapid Sorting and Prioritising of Organic Compounds Causing Environmental Risk with Estrogenic and Cytochrome P450 Activity

Doctoral dissertation

To be presented by permission of the Faculty of Natural and Environmental Sciences of the University of Kuopio for public examination in Auditorium L21, Snellmania building, University of Kuopio, on Saturday 28th January 2006, at 12 noon

Department of Environmental Sciences
University of Kuopio

KUOPION YLIOPISTO
KUOPIO 2006

Distributor:
Kuopio University Library
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 17 163 430
Fax +358 17 163 410
http://www.uku.fi/kirjasto/julkaisutoiminta/julkmyyn.html

Series Editors:
Professor Pertti Pasanen, Ph.D.
Department of Environmental Sciences

Professor Jari Kaipio, Ph.D.
Department of Applied Physics

Author´s address:
Department of Environmental Sciences
University of Kuopio
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel: +358 17 162 893
Fax: +358 17 163 191
E-mail: Arja.Asikainen@uku.fi

Supervisors:
Docent Kari Tuppurainen, Ph.D.
Department of Chemistry
University of Kuopio

Professor Juhani Ruuskanen, Ph.D.
Department of Environmental Sciences
University of Kuopio

Reviewers:
Professor Paola Gramatica, Ph.D.
Department of Functional and Structural Biology
Insubria University
Italy

Docent Lars-Olof Pietilä, Ph.D.
Orion Pharma
Helsinki

Opponent:
Professor Pentti Minkkinen, Ph.D.
Department of Chemical Technology
Lappeenranta University of Technology

ISBN 951-27-0349-1
ISBN 951-27-0444-7 (PDF)
ISSN 1235-0486

Kopijyvä
Kuopio 2006
Finland

Asikainen, Arja H. Use of computational tools for rapid sorting and prioritising of organic compounds causing environmental risk with estrogenic and cytochrome P450 activity. Kuopio University Publications C. Natural and Environmental Sciences 191. 2006. 51 p.
ISBN 951-27-0349-1
ISBN 951-27-0444-7 (PDF)
ISSN 1235-0486

ABSTRACT

The amount of chemicals in the environment has increased considerably as society becomes more and more industrialized. Parallel, the health risk incurred by chemicals with unwanted biological activity has also increased. Estrogenic regulation and cytochrome P450 metabolism are examples of biochemical systems, which can suffer from the action of the chemicals. The estrogenic system maintains the reproductive capability of humans and other animals, with the estrogen receptor playing a crucial role due to its ability to bind the hormones and form a ligand-receptor complex that regulates gene activation and the gene transcription processes. Disturbances in the estrogenic systems may lead to reproductive problems, development problems and provoke the growth of tumours.

The estrogenic activity of chemicals can be measured with *in vivo* and *in vitro* assays, but these tests are time consuming, laborious and expensive. Computational QSAR methods can be of assistance in this situation, providing tools for the prediction of the activity based solely on the molecular structure of the compound and this will decrease the number of animals used in the tests and conform to the EU recommendations. Moreover, QSAR models can also be employed by pharmaceutical and chemical industries in the development of new medicines and chemicals.

The most widely used QSAR-method is CoMFA, which is based on the molecular fields calculated from the 3D-structure of the molecules. One shortcoming of CoMFA is its need for alignment, i.e. the arrangement of the molecular structures with each others, which is time consuming, laborious and sometimes even impossible, if the set of compounds used for the study contains molecules with extensive structural variation. The main goal of this research was to examine the performance of several simple and fast-performing QSAR methods, such as PLS, MLR, kNN, and SOMFA, for the prediction of estrogenic activity, as alternatives to the more complicated QSAR methods. The performance of some methods was also examined with cytochrome P450 data sets. Moreover, three classification methods (DT, LVQ, and kNN) were employed for the discrimination of the estrogenic compounds into active and inactive compounds.

Models employing consensus kNN, a modified kNN method developed during this research, and PLS employing spectroscopic EEVA descriptors produced promising models. Their performance was even better than that of the previously reported models. In the classification study, models employing kNN and LVQ methods produced models with a good level of performance, and even models with an excellent level of performance with data sets comprising structurally similar compounds. The results obtained here are also in line with the previous results indicating that the performance of QSAR models is clearly dependent on the data sets employed. Overall, the results indicate that QSAR models can be employed for the prediction of estrogenic and cytochrome P450 activities, although the usefulness in terms of general applicability for prediction of new chemicals of these "small and congeneric" models is of little value. The performance of a QSAR model is at its best when the structural variation among the chemicals is small and the number of compounds to be evaluated is not very large.

Universal Decimal Classification: 504.064, 541.69, 577.175.8
National Library of Medicine Classification: WA 671, QV 627, WP 522, WH 190

Medical Subject Headings: organic chemicals; environmental pollutants; estrogens;
receptors, estrogen; cytochrome P-450 enzyme system; risk assessment;
quantitative structure-activity relationship; models, molecular; computer simulation; classification

ACKNOWLEDGEMENTS

This study was carried out at the University of Kuopio, Department of Environmental Sciences during the years 2003-2005 and I am grateful to the department for providing such excellent working facilities. Financial support was provided by the Academy of Finland, University of Kuopio, SYTYKE graduate school and Konkordia association. These are all gratefully acknowledged.

I wish to thank my main supervisor Docent Kari Tuppurainen for all his help, advice, support and encouragement during this time. Without him, this research might have been too much to handle. I also sincerely thank my other supervisor Professor Juhani Ruuskanen for all his help with this research, and also during the past years for providing an opportunity to work in the University of Kuopio. I am grateful to the reviewers Dr. Paola Gramatica and Docent Lars-Olof Pietilä for spending their time in reading this thesis and for their useful suggestions to improve it.

I express my gratitude to all my co-authors in the publications included in this thesis for their valuable contribution. I also wish to thank Teri Hiltunen for his guidance in the computing details, Juhani Tarhanen for his help in so many small but nonetheless so significant details covering all sorts of problems, and Ewen MacDonald for checking the English of this thesis.

I wish to thank all my friends for being so clueless about my work; it was a great help to me to be able to forget this thesis during my free time. I am grateful for your endless friendship, which always helps to put things to the proper perspective and help me remember that life is so much more than just work.

Finally, I wish to express my warmest thanks to my dear parents Maija and Huugo, who have never pressured me to study or have never valued people because of their status. You have provided excellent foundations for my life. I also thank my sisters Eija, Paula and Heljä for their love and support, and for their faith in me, especially in those moments when my own belief or ambition was lost.

Kuopio, December 2005

Arja Asikainen

ABBREVIATIONS

AMBER	assisted model building with energy refinement
AR	androgen receptor
CoMFA	comparative molecular field analysis
CoSA	comparative spectra analysis
CYP450	cytochrome P450
DBD	DNA binding domain
DES	diethylstilbestrol
DNA	deoxyribonucleic acid
DT	decision tree
E_2	17β-estradiol (estradiol)
ED	endocrine disruptor
EDKB	endocrine disruptor knowledge base
ER	estrogen receptor
EEVA	electronic eigenvalue
EVA	eigenvalue
kNN	k-nearest neighbour
LBD	ligand binding domain
LOO-CV	leave-one-out cross validation
LVQ	learning vector quantization
MD	molecular dynamics
MgVol	McGowan's characteristic volume
MLR	multiple linear regression
MM-PBSA	molecular mechanics–Poisson–Boltzmann surface area
MR	molar refraction
NR	nuclear receptors
NMR	nuclear magnetic resonance
PAH	polycyclic aromatic hydrocarbon
PCA	principal component analysis
PLS	partial least square
PR	progesterone receptor
PRESS	predicted residual sum of squares
Pr-R^2	predictive correlation coefficient (denoted also as Q^2_{EXT})
Q^2	cross-validated correlation coefficient
QSAR	quantitative structure activity relationship
QSPR	quantitative structure property relationship
RAR	retinoid receptor
RBA	relative binding activity
S_{press}	cross-validated standard error of prediction
SA	simulated annealing
SAR	structure activity relationship
SDEP	standard error of prediction
SOM	self-organizing map
SOMFA	self-organizing molecular field analysis
SXR	steroid xenobiotic receptor
TR	thyroid hormone receptor

ORIGINAL PUBLICATIONS

This thesis is based on the following publications, which are referred to with Roman numerals I-VI in the text.

I Asikainen, A.; Ruuskanen, J.; Tuppurainen, K. (2003). Spectroscopic QSAR methods and self-organizing molecular field analysis for relating molecular structure and estrogenic activity, Journal of Chemical Information and Computer Sciences 43, 1974-1981.

II Asikainen, A.; Ruuskanen, J.; Tuppurainen, K. (2004). Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds, SAR and QSAR in Environmental Research 15, 19-32.

III Asikainen, A.; Ruuskanen, J.; Tuppurainen, K. (2004). Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands, Environmental Science & Technology 38, 6724-6729.

IV Asikainen, A.; Kolehmainen, M.; Ruuskanen, J.; Tuppurainen, K. (2006). Structure-based classification of active and inactive estrogenic compounds with decision tree, LVQ and kNN methods, Chemosphere 62, 658-673.

V Asikainen, A.; Tarhanen, J.; Poso, A.; Pasanen, M.; Alhava, E.; Juvonen, R. (2003). Predictive value of comparative molecular field analysis modelling of naphthalene inhibition of human CYP2A6 and mouse CYP2A5 enzymes, Toxicology in Vitro 17, 449-455.

VI Asikainen, A.; Ruuskanen, J.; Tuppurainen, K. (2005). Alternative QSAR models for selected estradiol and cytochrome P450 ligands: Comparison between classical, spectroscopic and CoMFA/GOLPE methods, SAR and QSAR in Environmental Research, In Press

Some unpublished results are also cited.

CONTENTS

1 GENERAL INTRODUCTION

1.1 Background

The passage of chemicals into the environment has increased in parallel with rate of industrialization of modern society and the introduction of technology. The development of pharmaceutical and chemical industry, not only has beneficial effects, it also increases the risk of harmful effects for humans and wildlife produced by the chemicals released into the environment. One important group of these environmentally concerning chemicals are those which have biological activity for the hormonal systems of humans and animals. The hormonal system includes several receptors, including the estrogen receptor (ER), which maintains the activity of the endocrine system so it is ready for reproduction and sexual development. Chemicals having unwanted activity at ER, called endocrine disruptors (EDs), can interfere with or interrupt the normal action of ER and induce serious health related problems, such as infertility and growth of tumours.

The estrogenic activity of a chemical can be a wanted feature, as is the case with some drugs (e.g. tamoxifene, raloxifene and toremifene) or it can be unwanted or even unknown feature. Several *in vivo* and *in vitro* testing methods are available for the measurement of the estrogenic activity, but conducting these tests in a large group of chemicals is both time consuming and costly. The use of various computational methods can be of assistance in these situations, including the screening of the chemicals already present in the environment and which possibly possess estrogenic activity or for chemicals being developed which have the possibility of becoming EDs. The effect of structural features for the biological activity can be estimated with structure-activity relationship (SAR) and quantitative structure-activity relationship (QSAR) models providing a possibility to identify which chemicals should be prioritized for testing with more precise experimental assays.

Thus, the purpose of this research was to find easy-to-use and rapid computational methods to be used with diverse sets of compounds for the prediction of their estrogenic activity. The selection of methods was based on their simplicity, computational speed and the promising results obtained with those methods in previous studies. The methods tested with various sets of estrogenic compounds were partial least squares (PLS) employing spectroscopic electronic eigenvalue (EEVA) and eigenvalue (EVA), comparative spectra analysis (CoSA), *k*-nearest neighbours (kNN), consensus kNN (using averages of the predictions of several individual kNN models, a method developed during this research), and self-organizing molecular field analysis (SOMFA). In some situations it is not necessary to know the exact binding affinity of the chemical, simply the knowledge of whether a chemical is either active or inactive will suffice. To this end, the performance of three classification methods (learning vector quantization (LVQ), decision tree (DT) and kNN) was also tested with several estrogenic data sets.

Since cytochrome P450 enzymes play a major role in the biochemical processes involved in the metabolism of various compounds, it is possible that compounds having no estrogenic activity themselves are oxidized by cytochrome P450 enzymes to compounds possessing estrogenic activity. As a representative example, the oxidation of naphthalene to 1-naphthol and 2-naphthol by cytochrome P450 enzymes was also tested. Additional QSAR models were derived for a data set comprising of the inhibitory activities of several compounds for coumarin 7-hydroxylase employing comparative

molecular field analysis (CoMFA), Hansch-type QSAR, and PLS with EEVA and EVA descriptors.

1.2 Estrogen receptor

1.2.1 Function and structure

The estrogen receptor (ER) is a member of the hormone nuclear receptor (NR) superfamily, which includes also thyroid hormone receptor (TR), retinoid receptor (RAR), steroid xenobiotic receptor (SXR), progesterone receptor (PR), and androgen receptor (AR), which regulate the function of a large group of hormones (McLachlan 2001). Each receptor is responsible for the function of a group of specialized hormones (ligands) and can be classified according to the similarity of the ligands which they are binding (Weatherman et al. 1999). The receptor-hormone complex influences some cell or organ function, such as reproduction, maintenance of normal levels of glucose or ions in the blood, blood pressure, general metabolism, and other muscle or nervous system functions (Lintelmann et al. 2003). ER modulates the activity of hormonal steroids, such as estrogen and testosterone, resulting in effects on the reproductive, central nervous, immune, and cardiovascular systems and skin and bone. These tissue specific estrogenic responses occur when an estrogenic compound binds to the estrogen receptor activating the receptor-ligand complex to bind to specific sites of the DNA and in this way to change the expression of estrogen-responsive genes (Gaido et al. 1999).

The ER, like all steroid hormone nuclear receptors, consists of six functional domains labelled as A to F, of which the C domain is the DNA binding domain (DBD) and the E domain is the ligand binding domain (LBD) (Weatherman et al. 1999). Two estrogen receptors subtypes have been identified (ERα and ERβ) having the same structural features, but one of the two ER subtypes is dominant depending on the tissue in question and thus both ER subtypes regulate different hormonal processes (Gustafsson 1999). The domain structure, percentages of the domain homologies and the numbering of the amino acids are presented in Figure 1 for human ERα and ERβ.

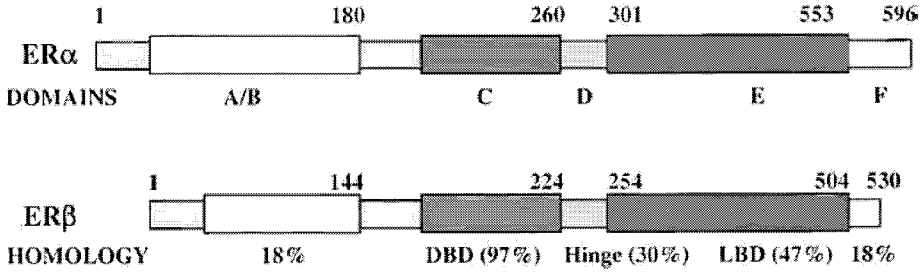


Figure 1. Domain structure, percentages of the domain homologies and the numbering of the amino acids for human ERα and ERβ (modified from Figure 1 of Kong et al. 2003).

The amino acid sequences of the ERα and ERβ have been identified for several species, such as human, mouse, rat, horse, pig and several fishes, and for the human ERα and ERβ sequence alignments are presented in Appendix 1. Also the crystallographic

structures for ER of several species are available from the RCSB Protein Data Bank (http://www.rcsb.org/pdb/).

1.2.2 Ligands

The structural variability of the ligands with estrogenic activity is very broad, and this feature makes ER unique compared to the other nuclear receptors, which tend to bind ligands with very constricted structural features. 17β-estradiol (E_2, Figure 2) is the most active of the endogenous estrogenic compounds, and structure activity relationship studies have revealed several factors essential for activity i.e. hydrophobicity, a ring structure, H-bond donor mimicking the 17β-OH, precise steric hydrophobic centers mimicking steric 7α- and 11β-substituent and phenolic ring mimicking the 3-OH for possession of H-bonding ability, which is the most important feature (Fang et al. 2001). Also the presence of small hydrophobic substituents at positions 3, 12β, 14 and 16α enhance the binding activity, and larger hydrophobic substituents are tolerated at positions 7α, 11β and 17α (Anstead et al. 1997). Some examples of the structures of steroidal and non-steroidal ligands are presented in Figure 2.

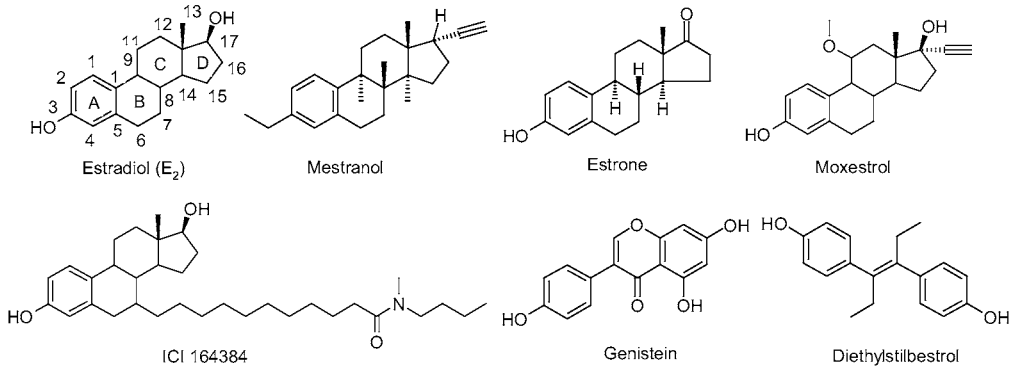


Figure 2. The structures of some steroidal and non-steroidal ER ligands.

1.2.3 Ligand binding domain

The ability of ER to bind compounds having wide variations in structures is due to the flexibility of the LBD. The residues forming the binding cavity of the ER are positioned differently depending on the ligand binding to the ER and this is schematically shown for ER with E_2 and raloxifene in Figure 3. There are also differences in the LBD between species and this means that some ligands may have estrogenic activity in one species, but not for the other species or the binding affinity is much smaller for one species than for others (see Table A2 in Appendix 2). Although ERα and ERβ are homologous, the homology of the LBD is only about 50% depending on the species. This may provide an explanation why some ligands possess different activity towards ERα and ERβ, and are involved in different transcriptional responses (Kuiper et al. 1997; Barkhem et al. 1998; Gustafsson 1999).

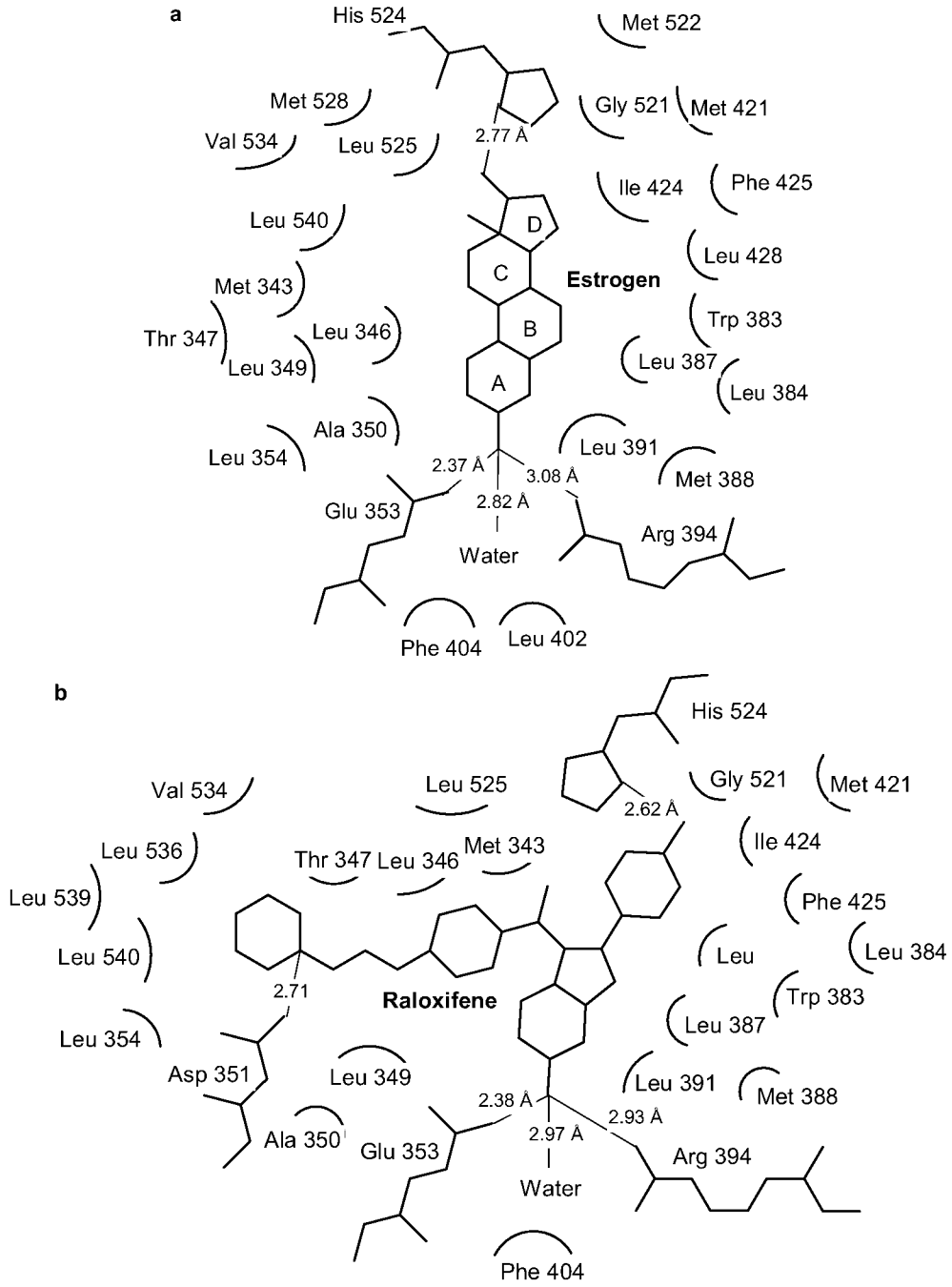


Figure 3. Schematic representation of the interactions made by E_2 (a) and raloxifene (b) within the binding cavity (modified from figure 2 by Brzozowski et al. 1997).

1.3 Endocrine disruptors

1.3.1 Compounds

Endocrine disruptors (EDs), i.e. the chemicals that can interfere with the action of ER as well as other nuclear receptors, have been defined by the U.S. Environmental Protection Agencys (EPAs) Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) as follows (EPA 1998):

"The EDSTAC describes an endocrine disruptor as an exogenous chemical substance or mixture that alters the structure or function(s) of the endocrine system and causes adverse effects at the level of the organism, its progeny, the populations, or subpopulations of organisms, based on scientific principles, data, weight-of-evidence, and the precautionary principle."

These compounds may be naturally occurring (such as herbal steroids called phytoestrogens) in the environment or man-made chemicals released intentionally (such as pesticides) or unintentionally by chemical accidents or as industrial pollutants to the environment. EDs include chemicals from the following chemical groups: polychlorinated biphenyls (PCB), polycyclic aromatic hydrocarbons (PAH), phenols, bisphenols, flavonoids, alkylphenol ethoxylates, synthetic nonsteroids (such as diethylstilbestrol (DES)), and pesticides (such as DDT and DDE) with their metabolites though many other chemicals may be active as well. The structures of these chemicals and natural estrogens can be found in Figure A2 in appendix 2.

1.3.2 Function

Endocrine disrupting chemicals can act as agonists or antagonists at the estrogen receptor. Agonists disturb the action of the ER system by binding to the ER in the same way as the natural ligand and activate the ligand-receptor complex producing the same effects as seen with the natural ligand. An antagonist is a compound which binds to the ER and inhibits the binding and action of the endogenous ligands of the ER. ED chemicals can also alter the synthesis and metabolism of natural hormones and in that way modify the hormone receptor levels (Sonnenschein and Soto 1998). The agonist and antagonist effects of the different ligands may be a consequence of the flexibility of LBD, which adapts and forms diverse bonds with the ligand, depending on its structure. These differences can be seen in Figure 3, where the interactions of ER with E_2 and raloxifene are presented.

The harmful effects attributable to ED chemicals include infertility, development defects, cancer, and problems with reproduction, which all have been detected with many species in laboratory experiments and in several case studies from natural wildlife (Danzo 1998; Tyler and Routledge 1998; McLachlan 2001). However, there are also critical comments about whether the results obtained in the laboratory can be extrapolated to the effects occurring in the natural environment, because it is a complex task to prove the cause and effect relationships in wildlife. This makes it almost impossible to determine conclusively which problems occurring in nature are really attributable to the endocrine disruptors (Cooper and Kavlock 1997; van der Kraak 1998; Li and Li 1998).

1.3.3 Activity assays

The bioactivity of a chemical can be measured with several *in vivo* and *in vitro* binding activity assays. With *in vivo* assays it is possible to examine the effects of estrogenic compounds on the total hormonal mechanisms, such as reproductivity, by using long-term animal tests. In *in vitro* tests only some tissue or a model of the tissue is used as research material and these tests are suitable for the situation in which the actual binding activity of a chemical is under investigation. For a comprehensive perspective of the bioactivity of a chemical both *in vivo* and *in vitro* tests need to be conducted, because the results of *in vitro* tests are not fully comparable with the real situation, where the whole organism with all its metabolising systems is involved (Reel et al. 1996; Shelby et al. 1996; Jobling 1998).

Competitive binding assays are one of the most widely used *in vitro* assays. In these tests the uterine cytosol of some species, such as human, rat, mouse or lamb, or the cytosol of human breast adenocarcinoma cell line MCF-7 or semi-purified human ER protein is used as the material containing the ER. Usually the binding of E_2 is used as the reference and the relative binding affinity (RBA) is the ratio of the molar concentration of E_2 to that of the competing chemical required to decrease the binding of E_2 by 50% multiplied by 100 (eq. 1).

$$RBA = (IC_{50} \text{ of } E_2 / IC_{50} \text{ of test chemical}) \times 100 \quad (1)$$

The main problems associated with *in vivo* and *in vitro* tests are that the tests are time-consuming, costly and very laborious, and the use of several approaches to conduct the same tests in different conditions leads to criticism about the comparability and reliability of the results (Zacharewski 1998).

1.3.4 Metabolism through cytochrome P450

Many molecular families do not exhibit estrogenic activity *per se*, but metabolic oxidation can transform inactive compounds into estrogens. In biological systems, the oxidation processes are often mediated by the family of cytochrome P450 enzymes so that the reaction chain is schematically:

"Pro-estrogen" (ox., P450) --> "estrogen" (ox., P450) --> quinone, epoxide etc.

The connection between cytochromes P450 and the estrogenic activity of molecules is not simply of academic interest – it is the topic of intensive research (Charles et al. 2000; Sugihara et al. 2000; Fertuck et al. 2001; Sanoh et al. 2002; Fujimoto et al. 2003; Kitamura et al. 2003; Mikamo et al. 2003; van Lipzig et al. 2005a; van Lipzig 2005b). It seems likely that the oxidative metabolism can increase considerably the number of endocrine disruptors. For example, naphthalene is a compound which itself does not possess any estrogenic activity, but it may be oxidized by cytochrome P450 to compounds having the structural properties leading to estrogenic activity.

1.4 QSAR

The idea behind QSAR is to identify some relationships between the structural descriptors of molecules and their biological activity. Computational QSAR modelling has become an important tool in the fields of chemistry, biochemistry, and toxicology. In recent decades, the development of computer technology has stimulated the use and development of various QSAR methods, and today QSAR is regarded as a scientifically credible tool in environmental toxicology and drug discovery. The relevance of the QSAR modelling for the manufacturing of new chemicals has been noted by the U.S. Environmental Protection Agency's (EPA) chemical assessment and control program (pre-manufacturing notice, PMN) and by the new EU Policy for Chemicals (REACH). The use of QSAR in toxicology has been reviewed recently by Schultz et al. (2003) and there is a review with the emphasis on estrogenic data sets by Schmieder et al. (2003).

1.4.1 Classical Hansch-type QSAR

In classical QSAR, physicochemical parameters, steric properties or some structural features are used as descriptors. Thus these models rely on the assumption that the biological activity of a chemical is dependent on its structure. The relationship between lipophilicity and some biological properties, such as narcotic and toxic properties, have been known over a century (Lipnick 1989), but it was not until the 1960's (Hansch et al. 1963, Hansch and Fujita 1964) when Hansch and his co-workers formulated a general QSAR equation (eq 2) combining hydrophobic, electronic and steric properties of molecules:

$$\log(1/C) = a(\log P) + b(\log P^2) + c(\sigma) + d(E_s) + e \qquad (2)$$

in which logP is the term describing hydrophobicity (measured often as octanol-water partition coefficient, $\log P_{ow}$ or $\pi = \log P - \log P_0$), σ is electronic Hammett substituent constant, and the steric term E_s is typically Taft's substituent constant (Hansch and Leo 1995). Since then, Hansch-type QSAR equations have been presented for hundreds of biochemical systems, including estrogenic (Gao et al. 1999) and cytochrome P450 (Hansch et al. 2004) data sets. One weakness of the Hansch method is that it is not suitable for the data sets containing compounds with wide structural variability (i.e. the congenericity principle).

An example of the regression equation obtained with Hansch method for the estrogenic activity of a data set with lamb uterine ER is presented in equation 3 (Gao et al. 1999) and for the inhibition activity of data set of mouse cytochrome P450 CYP2A5 enzyme in equation 4 (Hansch et al. 2004), respectively.

$$\log RBA = 1.38(\pm 0.32)\pi,11 - 1.08(\pm 0.39)MR,11 - 0.71(\pm 0.23)I + 1.92(\pm 0.24) \qquad (3)$$
$$n = 30,\ R^2 = 0.82,\ s = 0.30$$

$$\log 1/C = 10.1(\pm 2.35)MgVol - 24.12(\pm 8.81)MgVol^2 - 4.14(\pm 1.57) \qquad (4)$$
$$n = 28,\ R^2 = 0.90,\ s = 0.44,\ Q^2 = 0.88$$

in which π is hydrophobicity, MR is molar refraction, 11 refers to the numbering of the estrogenic skeleton (Figure 2), and I is an indicator variable (1 for the compounds

containing a 16α-OH group and 0 for other compounds), MgVol is McGowan's characteristic volume, n is the number of compounds, R^2 is the conventional correlation coefficient, s is the standard error of estimation and Q^2 is the cross-validated correlation coefficient

1.4.2 Modern QSAR methods

The Hansch method represented the beginning of the modern QSAR, which has now advanced toward more complex modelling as the processing power of the computers continues to increase and several new methods have been introduced into the field of QSAR. These methods can be roughly divided into 2-dimensional (2D) and 3-dimensional (3D) modelling depending on whether the descriptors are calculated using 2D or 3D structure of the compounds. However, nowadays this kind of subdivision is not so straightforward, as there are several methods, such as k-nearest-neighbours (kNN) and partial least squares (PLS) that employ both 2D and 3D theoretical descriptors.

3D modelling methods together with molecular dynamics (MD) simulations are the most advanced techniques, and they should be, at least in principle, able to take into account the interactions between the ligand and the receptor and to predict the behaviour of the whole ligand-receptor complex producing more precise predictions, comparable with the situation prevailing in real life. The shortcomings of these methods are their complexity and that they are not suitable for structurally diverse sets of compounds.

The methods for real life screening applications as well as being reliable should be easy to use and rapid to perform. Therefore, the applicability of several modern QSAR methods for the prediction of estrogenic activity and coumarin 7-hydroxylase inhibition for cytochrome P450 enzymes with some comparable examinations using the Hansch method was evaluated in this research. Additionally, the suitability of three classification methods for discriminating estrogenic compounds into inactive and active compounds was also tested. The methods used are described in more detail in the Material and Methods section.

2 AIMS OF THE PRESENT STUDY

Broadly, the purpose of this study was to investigate the suitability of several computational approaches for the prediction of the estrogenic activity of structurally diverse sets of compounds simply on the basis of their structure. The main objective was to find an easy-to-use and rapid method to provide an additional tool for the preliminary screening of compounds, and for distinguishing which of the compounds should have highest priority for evaluation in the more elaborate and expensive testing procedures.

The detailed objectives were as follows:

- To investigate the performance and suitability of several QSAR methods for the prediction of the estrogenic activity of a large, diverse sets of compounds (**I-III, VI**).
- To test the performance of LVQ method and compare it with the performance of decision tree and kNN methods for the classification of estrogenic compounds into active and inactive groups (**IV**).
- To examine the oxidation of naphthalene to 1- and 2-naphthol by CYP2A5 and CYP2A6 enzymes (**V**).
- To investigate the performance of CoMFA and PLS with spectroscopic EEVA and EVA descriptors for QSAR modelling of coumarin 7-hydroxylase inhibitory activity with CYP2A5 and CYP2A6 data sets (**V, VI**).
- To compare the effect of different data sets for the performance of the QSAR method employed (**III, IV, VI**).
- To evaluate the suitability of the tested QSAR methods for real applications.

3 MATERIAL AND METHODS

3.1 Biological data

3.1.1 Activities

The information about the data sets used for QSAR modelling, comprising of the binding activities for ER (**I-IV**, **VI**) of several species and of the coumarin 7-hydroxylation inhibition activities for two cytochrome P450 enzymes (**V**, **VI**), is presented in Table 1.

Table 1. Summary of the biological data sets and the modelling methods used in the studies **I-VI**.

Study	Reference	n[1]	Receptor/ enzyme	Structures/ activities	Methods
I	Sadler et al. (1998) and Sippl (2000)	36	Mouse ER	Figure 1 and Table 1 in publication **I**	PLS/EEVA PLS/EVA CoSA SOMFA
II	EDKB[2]	142	Rat ER	Appendix 2	kNN consensus kNN
III	EDKB[2]	245	Calf ER Human ERα Human ERβ Mouse ER Rat ER	Appendix 2	consensus kNN
IV	EDKB[2]	339	Calf ER Human ERα Human ERβ Mouse ER Rat ER	Appendix 2	DT LVQ kNN
	Saliner et al. (2003)	117	Human ERα	Saliner et al. (2003)	
V	Juvonen et al. (2000)	24	CYP2A5 CYP2A6	Poso et al. (2001)	CoMFA
VI	Juvonen et al. (2000)	28	CYP2A5 CYP2A6	Figure 1 and Table 1 in publication **VI**	Classical QSAR PLS/EEVA PLS/EVA MLR/EEVA MLR/EVA
	Napolitano et al. (1995)	30	Lamb ER	Figure 2 and Table 2 in publication **VI**	

[1] Number of the compounds in the data set.
[2] EDKB = endocrine disruptor knowledge base (http://edkb.fda.gov/databasedoor.html)

3.1.2. Oxidation of naphthalene

Mouse CYP2A5 and human CYP2A6 enzymes were obtained as liver samples, which were used for the determination of the inhibitory activity of naphthalene for coumarin 7-hydroxylation and for the measurement of the naphthalene oxidation to 1- and 2-naphthol.

Experiments were performed with the incubation procedures which are described in detail in publication **V**.

3.2 Molecular modelling

In the studies **I-IV** and **VI,** the three-dimensional structures of the molecules were modelled with the HYPERCHEM program package (Hypercube, Inc) and minimized with MM+ force field. The structure of the lowest energy conformer was then fully minimized with AM1 (Dewar et al. 1985) calculations as implemented either in HYPERCHEM (**IV**) or in AMPAC program package (QCPE No. 506, version 2.11) (**I-III, VI**). In study **V,** the 3D-structures were modelled by using the sketch option in the SYBYL 6.6 program package (Tripos Inc.), and the structures were further minimized with MMFF94 force field (Halgren 1992).

3.3 Structural descriptors and variable selection

3.3.1 Spectroscopic descriptors

The computation of spectroscopic QSAR descriptors by means of Gaussian smoothing involved the following steps: (i) eigenvalues of molecular orbital energies and vibrational frequencies for EEVA and EVA descriptors were calculated employing the HYPERCHEM program package and the ^{13}C and ^{1}H NMR chemical shifts for CoSA were calculated employing the gauge-invariant atomic orbitals (GIAO) method (Wolinski et al. 1990) (see publication **I** for detailed information about the calculations), (ii) eigenvalues were transformed to a bounded scale, (iii) a Gaussian kernel of fixed standard deviation σ was placed over each eigenvalue, and (iv) the descriptors were calculated by summing the overlaid kernels at intervals of L (usually set at σ/2, eq 5):

$$Descriptor(x) = \sum_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-E_i)^2/2\sigma^2} \qquad (5)$$

where E_i is the ith eigenvalue of the molecule in question and N is the number of eigenvalues.

The adjustable parameter σ has to be optimized separately for each data set and this was performed employing leave-one-out cross-validation (LOO-CV) tests for a large number of reasonable σ values, and choosing the value producing the best predictive performance. Performing LOO-CV and the validation of the best performing model are described in detail in section 3.5. The selected σ value was then used for the external models employing randomized training and test sets. The effect of σ value for the performance of the models is described in Figure 2 in publication **I**.

3.3.2 Molecular descriptors

With the other methods tested in this research, in addition to the Hansch type QSAR, the used molecular descriptors were calculated with DRAGON program package (version 3.0, Todeschini et al.; Todeschini and Consonni 2000) (**II-IV**) employing either simulated annealing (SA) (**II, III**), which is described more detailed by Kirkpatrick et al. (1983) and

by Zheng and Tropsha (2000), or principal component analysis (PCA) (**IV**) (described in the next section) as the variable selection method. The typologies of the molecular descriptors calculated by DRAGON are listed in Table I in publication **I**. With Hansch-type QSAR (**VI**), the used descriptors were McGowan's characteristic volume and its square (MgVol and $MgVol^2$) in the study employing the CYP2A5 and CYP2A6 data sets, and molar refraction (MR), hydrophobicity (π) and an indicator variable in the study employing the lamb ER data set.

3.3.3 Principal component analysis

Principal component analysis (PCA) (Wold et al. 1987) is a linear method for the reduction of the dimensionality of the set of descriptors by finding uncorrelated linear combinations of the original descriptors (principal components, PC), which explain as much of the variance in the original descriptors as possible. PCA is especially suitable for situations where the number of descriptors is larger than the number of compounds.

Since the same number of PCs can be calculated as the number of original descriptors, the number of significant PCs has to be determined in some way. Three methods are mainly used for carrying out this procedure. The simplest is Kaiser's criterion, in which PCs with eigenvalue greater or equal to one are selected. The second alternative is the Cattell's scree test, in which the eigenvalues are plotted and the number of PCs is determined by visually inspecting the inflection point of the eigenvalue curve. The third method is called Humphrey-Ilgen parallel analysis, in which data with random numbers of the same size as the original data is analysed with PCA, the eigenvalues of both data sets are plotted and the intersection point of these curves determines the number of PCs to be selected. Each method is suitable for different situations, depending of the composition of the original data.

In this research, PCA was employed in the classification study (**IV**) as an easy to use descriptor reduction method employing Humphrey-Ilgen parallel analysis as the selection criteria (see Figure 1 in publication **IV**). PCA calculations were performed with the MATLAB program package (version 5.3). The number of PCs was further reduced with the feature selection option of the TOOLDIAG program package (version 2.1) provided by Rauber.

3.4 Computational methods

The modelling methods used in each study are listed in Table 1 on page 22 and described briefly below. More detailed information is available from the references mentioned in the text. The selection of the tested methods was based on their simplicity and for the promising results obtained with these methods in previous studies conducted by other research groups.

3.4.1 Multiple linear regression

Multiple linear regression (MLR) is a simple method to find relations between several independent descriptors and a dependent variable (such as biological activity) by fitting a linear equation to the data by the means of the least squares. The limitation of MLR is that the independent descriptors should be uncorrelated with each other, but this can be

controlled by using the appropriate methodologies. The result of the MLR is a regression equation (such as equations 3 and 4 on page 19), which can be employed for the prediction, and it was employed in this research to construct a few simple regression equations to be compared with the performance of PLS employing spectroscopic EEVA and EVA descriptors, and with GRID/GOLPE and CoMFA methods (**VI**).

3.4.2 Partial least squares

Partial least squares (PLS) (Geladi and Kowalski 1986) is a modelling method, which is suitable for the situations where the number of descriptors is much larger than the number of compounds and the effect of individual descriptors and the underlying effect of the combinations of the original descriptors on dependent variable are difficult to verify. It has the advantage of being able to handle multicollinearity among the independent descriptors (i.e. the descriptors can be correlated with each other). With PLS, PCs are constituted as mutually independent linear combinations of original descriptors, which are chosen in such a way that they provide maximum correlation with the dependent variable (in this case, biological activity).

With PLS, as with PCA, the optimum number of the principal components included to the models has to be determined. This is empirically carried out with the cross-validation by constructing PLS models using an increasing number of components and the number of PCs producing a model with smallest standard error of prediction is chosen (S_{press} value, see Table 2).

PLS was employed in this research with all the QSAR methods tested (**I-VI**). Most of the PLS analyses were performed using in-house MATLAB scripts written by the authors. The scripts employ an efficient modification of the PLS algorithm, SVDPLS (singular value de-composition PLS), which facilitates very rapid cross-validation runs.

3.4.3 Spectroscopic methods

Spectroscopic QSAR methods are a group of simple performing tools employing either experimental or theoretically calculated spectra as descriptors, which can be based on various molecular features such as orbital energies (EEVA) (Tuppurainen 1999), vibrational frequencies (EVA) (Ferguson et al. 1997), ^{13}C and ^{1}H nuclear magnetic resonance (NMR) chemical shifts, infrared absorption (IR) or electron ionization mass spectra (EI MS). Modelling methods commonly used with spectroscopic descriptors are PLS, Comparative Spectra Analysis (CoSA) (Bursi et al. 1999) and Artificial Neural Networks (ANN) (Vračko 1997). All these methods have produced comparable results as have been obtained with more complicated QSAR methods employing various data sets. With the estrogenic data set, CoSA employing ^{13}C NMR data have yielded models with good predictive abilities (Beger et al. 2000; Beger et al. 2001a). This encouraged us to test the performance of PLS employing EEVA and EVA descriptors and the performance of CoSA employing ^{13}C and ^{1}H nuclear magnetic resonance (NMR) chemical shifts also in this research (**I, VI**).

3.4.4 3D methods

The methods based on the 3D structure of the compounds employing some molecular field, such as CoMFA (Cramer et al. 1998), CoMSIA (comparative molecular similarity indices analysis) (Klebe et al. 1994) and SOMFA (Robinson 1999), are perhaps the most reliable QSAR methods available. The shortcomings of these methods are the need for structural alignment, which is a time consuming procedure and has a major effect on the predictive power of the models. The alignment also becomes more difficult as the structural variability of the compounds increases, so these methods are not suitable for use with complicated data sets.

CoMFA is the most popular of the field based methods, which has also been employed extensively with estrogenic data sets (Gantchev et al. 1994; Waller et al. 1995; Waller et al. 1996; Tong et al. 1997; Tong and Perkins 1997; Wiese et al. 1997; Sadler et al. 1998; Tong et al. 1998; Xing et al. 1999; Shi et al. 2001; Coleman et al. 2003; Waller 2004). The descriptors for the CoMFA models are calculated from the grid-based molecular fields, usually steric (representing the shape of the compound) and electrostatic fields, based on the aligned 3D structures of the compounds. The alignment of the compounds is usually performed as a ligand-based alignment, although the receptor based alignments have been claimed to produce models with better predictive abilities (Sippl 2000; Sippl 2002). PLS is usually employed for the calculation of the correlations between biological activities and CoMFA descriptors. The previous models employing CoMFA for estrogenic data sets were used as comparison in several studies in this thesis and as a primary modelling method with the cytochrome P450 data set for the inhibitory activity of the compounds against coumarin 7-hydroxylase (**V**).

SOMFA is another 3D grid-based method employing molecular fields with no need to use any additional statistical tools such as PLS. The grid can be calculated with any molecular field, also with the steric and electrostatic fields usually employed with CoMFA. Robison et al. introduced SOMFA in 1999 and since then many SOMFA models with predictive ability comparable with CoMFA have been published (Li et al. 2003; Smith et al. 2003; Martinek et al. 2005). Therefore SOMFA was tested with the estrogenic data set employing steric (shape), electrostatic and polarizabilty fields as the descriptors (**I**).

3.4.5 Molecular dynamics simulations

One step forward from the 3D methods are methods employing molecular dynamics simulations used for the calculation of the interactions of receptor-ligands complexes. These methods are able to create the situation closest to the situation prevailing in the real life systems. Since Oostenbrink et al. (2000) and van Lipzig et al. (2004) have proved that the free energy of the estrogenic compounds can be calculated accurately with molecular dynamics simulations of ligand-receptor complex and furthermore, the relationship between the binding energies and the binding activities of the estrogenic compounds has been verified by Hanson et al. (2003), we decided to study the relationship between binding activities and the binding energies calculated from MD simulations.

The method used for the calculation of the binding energies was based on the molecular mechanics–Poisson–Boltzmann surface area (MM-PBSA) method (Vorobjev 1998; Kollman 2000), and its capability to calculate free energies was first tested with the data set introduced by Oostenbrink et al. (2000). The data set used for the modelling

consisted of 39 compounds with the binding affinities for human ERα obtained from the EDKB. The crystallographic structure of the LBD for the human ERα was obtained from the RCSB Protein Data Bank (http://www.rcsb.org/pdb/). The 3D-structures of the molecules were constructed with HYPERCHEM optimized with Gaussian 98. Fitting of the compounds to LBD was performed against the crystallographic structure of DES (diethylstilbestrol), which was obtained with the structure of the LBD, using the SYBYL program package. MD simulations were performed with the Sander program code, which is a part of the AMBER program package (Pearlman et al. 1995, Ponder and Case 2003). These calculations will remain as unpublished data, for the reasons discussed in detail in the section 4.1.3.

3.4.6 kNN and consensus kNN

k-nearest-neighbour is perhaps the simplest pattern recognition method ever presented, and it can be used for classification with a categorical dependent variable or for regression problems with continuous dependent variable. Zheng and Tropsha (2000) introduced the kNN method in the field of QSAR, employing simulated annealing (SA) as an automated variable selection method for evaluating an estrogenic data set and these workers obtained promising results. Moreover, kNN is a simple, rapid and highly automated method and it provides a useful alternative to the more complicated 3D modelling methods

Each data point is described with independent descriptors and a dependent variable, i.e. biological activity in this thesis, and the similarity of the descriptor vectors is calculated using Euclidian distances. For predictive purposes, the weighted average of the dependent variable of the k nearest neighbours is used as the outcome. The optimum number of neighbours (k_{opt}) used for prediction is the most important factor affecting the results of the kNN models and this needs to be set to a value large enough to minimize the probability of misclassification but small enough so that the used k nearest neighbours are close enough to the molecule to be predicted. The number of k is usually defined with LOO-CV method, choosing the value producing the model with the best predictive power (the use of LOO-CV is described in section 3.5).

The consensus or ensemble models are a combination of several models using the average of the predictions for the final prediction. This approach has been used with many methods, especially with neural networks (Devillers et al. 1998; Manallack et al. 2002; Tetko 2002; Tetko and Tanchuk 2002), producing better predictive performances than can be achieved with the individual models. With estrogenic data sets, the decision forest (i.e. a combination of decision tree models) has been employed by Tong et al. (2004). As a consequence, the use of the consensus kNN method, in which the average of the predictions of 50 individual kNN models was used as the final result, was decided to be tested in our research. The differences between the individual models were the descriptor pools selected with SA and the number of the k nearest neighbours used for prediction. In addition to our group, the consensus kNN method has been employed by Votano et al. (2004) in their QSAR study employing a data set of 3363 compounds with Ames genotoxicity as the biological activity and its performance was found to be better than that of the individual models.

In this research, kNN was tested for both regression (**II**) and classification (**IV**) problems, and the consensus kNN approach was tested only with the regression problem

(**II, III**). The calculations were performed with the MATLAB program package (MathWorks, Inc., version 5.3) employing the scripts written by our research group. More details of the calculations can be found in publications **II**, **III** and **IV**.

3.4.7 Learning vector quantization

Learning vector quantization (LVQ) is a supervised neural network method, similar to self-organizing maps (SOM) (Kohonen 1998; Kohonen 2001). Both are capable of converting relationships of high-dimensional data into a simple low-dimensional form, with the difference that SOM is an unsupervised method. The LVQ network is trained by using model vectors, which are adapted according to the LVQ updating rule, such that the coordinates of the model vectors will eventually be characteristic of the original descriptor vectors in each class. The difference between the training of the LVQ and SOM is that no self-organizing between model vectors takes place in the LVQ as it does with SOM. The difference between LVQ and kNN is that the original vectors are used for the prediction with kNN whereas these vectors used for the tuning of the models vectors in LVQ. More information about the equations used for the training of the LVQ models is provided by Kangas and Kohonen (1996), for example.

The main parameters impacting on the predictive ability of the LVQ model are the number of training epochs used during the model development and the number of model vectors. Both parameters are dependent on the data sets and have to be adjusted prior to the modelling. The effect of the number of model vectors to the performance of the model is described in Figure 4 in publication **IV**. As far as I am aware, LVQ has not been employed with estrogenic data sets earlier but Baurin et al. (2004) did describe excellent performing models employing LVQ in their screening test with a COX-2 inhibition data set and therefore the performance of LVQ was tested in this research and compared with the other classification methods employed (**IV**). The calculations were performed with MATLAB (version 6.5), using the codes included in the Neural Network Toolbox (version 4.0).

3.4.8 Decision tree

Decision tree (DT) is a supervised rule based method, which is typically used for classification problems, although it can also be employed for regression problems. The earliest applications of the decision trees were introduced in the 1960s (Morgan and Sonquist 1963), but the wider use of DT originates from the work of Breinman et al. (1984) who introduced the classification and regression tree (CART) method, which has been in widespread use since that time. The basis of DT is to find some features from the descriptor pool typical for compounds in each class using a training set and based on these features some rules are created, called the nodes or leafs of the tree. After the tree is trained with a teaching set, it is possible to use it for predictive purposes. The advantages of the DT are its ability to handle noisy data, its rapidity and its capability to model nonlinear problems.

Decision trees and decision forest, i.e. combinations of several decision trees, have been employed earlier with estrogenic data, producing good performing models (Hong et al. 2002; Tong et al. 2003). The results of these earlier studies were the main reason for choosing DT as one of the classification methods to be compared with the performance of

LVQ (**IV**). The models reported here were constructed with the C++ program written by Borgelt (freely available on the Internet at http://fuzzy.cs.uni-magdeburg.de/~borgelt/doc/dtree/dtree.html, accessed in May 2004), with the default parameters. More details of the DT algorithm can be found in Borgelt (1998).

3.5 Validation of the models

The validation of a QSAR model, i.e. the determination of the predictive power of the model, is one of the most important steps to produce reliable and useful models (Eriksson et al. 2003; Tropsha et al. 2003). Cross-validation, specifically leave-one-out (LOO) and leave-*n*-out methods are commonly used. LOO-CV, in which a compound is left out from the input data, then a QSAR model is constructed and the activity of the omitted compound is predicted, after which the cycle continues until all the activities have been predicted once, is an unambiguous and simple method, which provides comparable results between models obtained with different methods (Wold 1991). Leave-*n*-out cross-validation is a similar procedure leaving *n* compounds out instead of single one compound, leading to two sets of compounds, of which one is used for training of the model and the other as the test set. This process is repeated with different subsets until each compound is used once for the test set (Shao 1993). The predictive power of the cross-validated models is evaluated with S_{press} (cross-validated standard error of prediction) and Q^2 (cross-validated correlation coefficient) values (Table 2).

LOO-CV is a measure of internal predictability of a QSAR model and provides a simple way to compare different models and methods with each other, but for the real life situation, the external predictive power, i.e. the performance of the model for the prediction of unknown biological activities, is much more significant. Since good internal performance achieved with cross-validation does not necessarily guarantee good external predictive power, some external validation tests need to be done with all of the present QSAR models (Golbraikh and Tropsha 2002a). This is usually carried out by using external test sets, which are not used during the modelling until the final model has been constructed and then the dependent responses (i.e. biological activities) of the external test set are predicted (Tropsha et al. 2003).

The main problem associated with external validation is how to choose the compounds to be used in the test set. One solution is to divide the set of compounds randomly into training and test sets with some ratio (for example 2/3), calculate a model, and then repeat this procedure several hundreds of times, employing different training and test sets. This method gives perhaps the most realistic impression of the external predictability of the model. Another approach is to group the compounds according to their activity and structure and then to choose the compounds covering most extensively the variations in the activities and structures by means of various clustering methods (Golbraikh and Tropsha 2002b; Gramatica et al. 2004).

The LOO-CV was used as the internal validation method in all of the studies included in this thesis. The external validation was performed with a random selection of the training and test set with all other methods except in the first study with the kNN (**II**), in which the compounds for the test sets were selected with the second method described above. The results of the validation and the goodness of a model are expressed with the calculated parameters, which describe the predictive power of the model and the

parameters with their abbreviations and equations used in this research are presented in Table 2.

In addition to the internal and external validation, the possibility of chance correlations, i.e. fortuitous correlations without any predictive ability, of the models should be always checked. This can be performed with y-randomization tests, i.e. the activities of the compounds are scrambled randomly between the descriptors and the modelling is then repeated as with the correct data set. These tests were performed with all models constructed within this research and no chance correlations were found.

Table 2. The names, abbreviations, equations, and definitions of the statistical parameters used for the validation of the models.

Name	Abbreviation	Equation	Definition
Predicted residual sum of squares	PRESS	$\sum(y_{obs} - y_{pred})^2$	y_{obs} = observed activity, y_{pred} = predicted activity. Optimal number of PCs is determined by minimizing PRESS value.
Cross-validated standard error of prediction	S_{press}	$\sqrt{(PRESS/(n-c-1))}$	c = number of PCs, n = number of compounds. Estimates the standard error of LOO validated model. The best performing internal model can be determined by minimizing S_{press} value.
Cross-validated correlation coefficient	Q^2	$1 - \frac{PRESS}{\sum(y_{obs} - y_{mean})^2}$	y_{mean} = mean value of the observed activities; Estimates the predictive ability of LOO validated model, and can be used instead of S_{press} value for validating internal models. $Q^2 > 0.5$ indicates the model as being statistically significant.
Standard error of prediction	SDEP	$\sqrt{PRESS/n}$	PRESS is calculated using only the test set compounds, n = number of test set compounds. Estimates the standard error of external prediction.
Predictive correlation coefficient	Pr-R^2	$1 - \frac{\sum(y_{obs} - y_{pred})^2}{\sum(y_{obs} - y_{mean})^2}$	y_{obs} and y_{pred} refers to the test set compounds. y_{mean} is the mean of the training set activities. Estimates the external predictive ability of the model. Denoted also as Q^2_{EXT}.

4 RESULTS AND DISCUSSION

4.1 Modelling of estrogenic data

4.1.1 Spectroscopic PLS/EEVA, PLS/EVA and CoSA methods

PLS employing spectroscopic EEVA and EVA descriptors, and CoSA with ^{13}C and ^{1}H chemical shifts were tested with the mouse ER data set (**I**), and PLS/EEVA and PLS/EVA also with the lamb data set (**VI**). With the mouse ER data, the internal predictability only for CoSA ^{13}C model was satisfactory, producing almost comparable results to those obtained with the CoMFA model as reported by Sadler et al. (1998) employing the same data set. These results are, however, clearly inferior to that achieved with the same data set using a highly sophisticated CoMFA model that employed a receptor-based alignment of ligands and smart region definition (SRD) for variable selection (Sippl 2002). The S_{press} and Q^2 values for our models and for the CoMFA models by Sadler et al. (1998) and by Sippl (2002) are presented in Figure 4. Detailed numerical parameters for our models can be found in Table 2 in publication **I.**

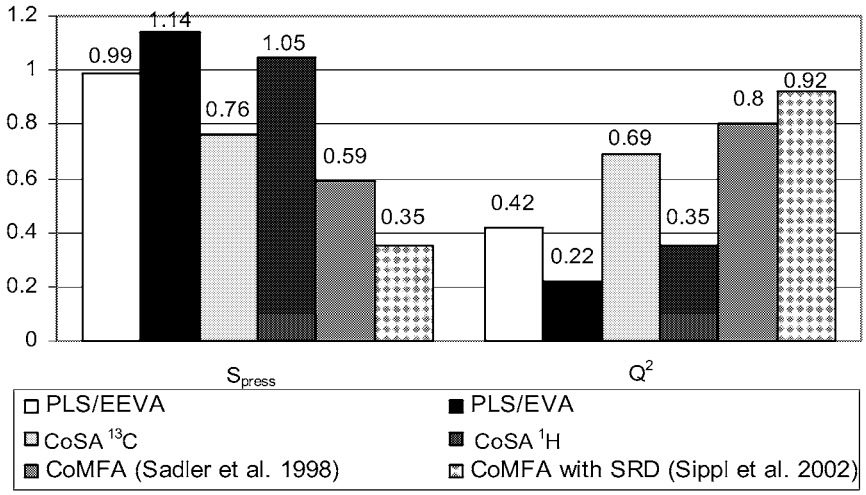


Figure 4. S_{press} and Q^2 values for the LOO-CV PLS/EEVA, PLS/EVA and CoSA models (**I**), together with CoMFA model by Sadler et al. (1998) and CoMFA model employing smart region definition (SDR) by Sippl (2002) with the same mouse ER data set.

As all other spectroscopic methods tested failed in internal prediction, the external predictability was investigated only with CoSA ^{13}C method with 500 randomized test runs. The results were not so encouraging since the average of SDEP was 0.86 and Pr-R^2 was 0.49. The weak performance of the spectroscopic methods tested was surprising, since previous studies with different receptor binding data sets employing both CoSA and PLS/EEVA have produced models with good levels of performance with LOO-CV Q^2 values being 0.71 (Beger et al. 2001b), 0.71 (Beger and Wilkes 2001a), and 0.78 (Beger and Wilkes 2001b) in CoSA studies, and 0.82 (Tuppurainen and Ruuskanen 2000) and 0.84 (Tuppurainen et al. 2002) in PLS/EEVA studies.

Examination of the data set used here (see Figure 1 in publication **I** for the structures) revealed at least two features in the structures of the compounds which accounted for the decline in the performance of the models. Both enantiomers and symmetric molecules were present in the data set and this can cause problems for all spectroscopic QSAR methods, since all of the physical properties of enantiomers (with the exception of for their interaction with polarized light), including MO energies, IR frequencies and NMR chemical shifts, are completely identical, although their biological activities may vary considerably. As a consequence, all spectroscopic descriptors lack the information required to distinguish the different activities of the enantiomer pair. Further, the increasing symmetry of molecules may result in a loss of information, resulting in weaker predictive ability of the spectroscopic methods for symmetric molecules. The influence of enantiomers and molecular symmetry on the spectroscopic descriptors is further discussed in publication **I**.

The performance of PLS/EEVA and PLS/EVA described above encouraged us to perform further tests employing these methods. A data set with the activities for lamb ER was tested with 500 randomized external test runs with PLS/EEVA and PLS/EVA, with an additional Hansch- type QSAR model providing some kind of comparison to the results obtained (**VI**). The averages of the SDEP and Pr-R^2 values for the 500 external test runs are presented in Figure 5.

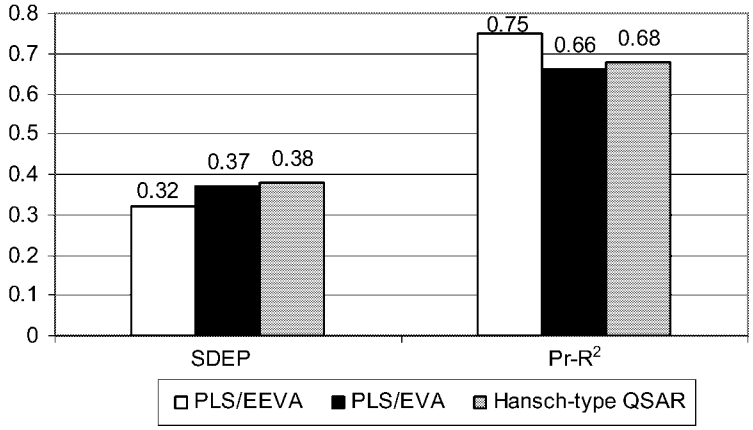


Figure 5. Averages of the SDEP and Pr-R^2 values for the 500 randomized external test runs for PLS/EEVA, PLS/EVA and Hansch-type QSAR models with lamb ER data set (**VI**).

These results support the assumption made above that the weak performance of PLS/EEVA and PLS/EVA methods may be a result of the fact that the mouse ER data set was not suitable for use with spectroscopic methods. The simple Hansch-type model successfully related the estrogenic activity with molar refraction (MR), hydrophobicity (π) and an indicator variable (1 for the compounds containing a 16α-OH group and 0 for the other compounds). This supports the fact that Hansch-type QSARs are useful with data sets comprising compounds with less extensive structural variations (see Figure 2 in publication **VI** for the detailed structures of the data set used). In the comparison between PLS/EEVA and PLS/EVA models, PLS/EEVA outperformed PLS/EVA.

4.1.2 SOMFA

The performance of SOMFA was tested with the same mouse ER data set as employed with PLS/EEVA, PLS/EVA and CoSA reported above. SOMFA models were tested with the fields based on the shape (SOMFA 1), electrostatic potential (SOMFA 2) and, as a theoretical novelty, polarizability (SOMFA 3) of the compounds producing internal predictabilities with S_{press} values of 0.63, 0.85 and 0.73, and Q^2 values of 0.76, 0.55 and 0.67, respectively (**I**). The internal performance of all SOMFA models clearly outperformed the PLS/EEVA and EVA models described above, and the SOMFA 1 model with the field based on the shape of the compound was better than that of the CoSA ^{13}C model described above and almost comparable with the CoMFA model presented by Sadler et al. (1998) (Figure 4).

The external performance of SOMFA models was tested with 500 randomized runs producing results either as good as (SOMFA 2 model) or clearly better (SOMFA 1 and SOMFA 3 models) than that of obtained with the CoSA ^{13}C model described earlier. SDEP and Pr-R^2 values for the SOMFA models with the values for CoSA ^{13}C model for comparison are presented in Figure 6, and detailed statistical parameters can be found in Table 3 in publication **I**.

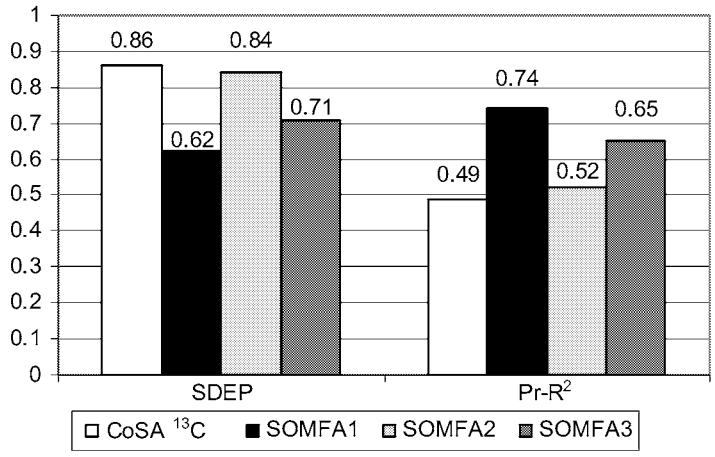


Figure 6. Averages of the SDEP and Pr-R^2 values for 500 randomized external runs for CoSA ^{13}C and SOMFA models employing mouse ER data set (**I**).

The enatiomers and symmetric compounds included in the data set did not cause any problems with the SOMFA method, and the structures of the compounds in the data set were sufficiently homogenous to permit alignment. The better performance of the CoMFA models by Sadler et al. (1998) and by Sippl (2002) (Figure 4) compared to the SOMFA models are due to the superiority of the receptor-based alignment which is utilized in the CoMFA models over the field-fit alignment used here.

4.1.3 Molecular dynamics calculations

The results of the test with MD calculation have not been published, because the predicted relation between the binding energies and ER binding activities was not found, and the model failed to predict the binding activities of the used compounds. The

calculated free energies were comparable to the experimental values (for those compounds for which they were available), and it was demonstrated that the MM-PBSA method did perform properly. Thus it seems evident that the reason for the failure of the models was not due to the selected calculation method. As an alternative to QSAR modelling, MD calculations can be used to study the effect of structural variations in the binding affinities, as demonstrated by Wurtz et al. (1998).

One possible explanation for the inability of the MD modelling was that there was excessively large structural variation of the compounds in the data set, although those compounds having the most complex structures were omitted, but obviously there still remained too many complicated compounds. In fact, this was detected during the fitting of the compounds to the LBD, which proved to be a very difficult task with the compounds having a clearly different structure than that of DES, employed for the fitting. Overall, even if the MD model had been successful for prediction, the construction of the MD models was laborious and time consuming compared to the other modelling evaluated during this project, making MD calculations unsuitable for the goals of this thesis.

4.1.4 kNN and consensus kNN

In the first experiment performed with kNN, the use of simulated annealing (SA) as the variable selection method was tested and the effect of the size of the variable pool for predictive ability was studied with a data set containing binding activities for the rat ER (**II**). The size of the variable pool did not have any significant effect on the predictive ability of the models (see Table II in publication **II**), and so 250 variables were selected to be used with all of the kNN models built. The suitability of the SA as the variable selection method was tested by constructing a model employing all the descriptors along with the models using a SA validated descriptor pool, and the performance of the model using all descriptors was clearly poorer than the performance of the other models. The selection of training and the test sets was performed so that the structural variation of the compounds and the variance of the activities were covered adequately. The effect of the random selection of the training and test sets is further discussed in section 4.4.

The results of the first kNN models indicated that the predictive ability of the kNN with the used rat ER data set was very promising. Altogether 50 models validated with LOO-CV methods were produced with the average values of S_{press} and Q^2 being 0.95 and 0.70, and 50 external models validated with separate training and test sets were produced with the average values of SDEP and Pr-R^2 being 1.12 and 0.57, respectively (see Tables II and IV in publication **II**). In addition to the basic kNN, the rat ER data set was used for preliminary testing of the consensus kNN method by calculating the averages of the predictions of 50 models as the final results (50 models were confirmed to be a sufficient amount of individual models by also calculating the average of 500 models too, but this did not alter the final results), producing clearly better predictions (SDEP = 1.01 and Pr-R^2 = 0.65) than the corresponding conventional kNN models. Comparison of the statistical parameters of the individual external models and the external consensus kNN model revealed that only 6 of the 50 individual models were statistically better than the consensus model.

The consensus kNN approach was further tested with a large set of compounds divided into five subsets depending of the ER for which the binding assays were

performed (calf ER; 53 compounds, human ERα; 61 compounds, human ERβ; 61 compounds, mouse ER; 68 compounds and rat ER; 130 compounds) (**III**). The internal predictability of the consensus models was impressive, producing models with Q^2 values between 0.69 (human ERβ) to 0.79 (human ERα). Statistical parameters for all internal models can be found from Table 1 in publication **III**. The data sets of calf ER, rat ER and mouse ER have been used in several previous studies employing various QSAR methods and the LOO-CV validated results of internal consensus kNN models were compared with the results of those models (see Figure 7 for Q^2 values and Table 2 in publication **III** for detailed statistical parameters).

Consensus kNN outperformed all previous models, including the CoMFA models presented. With the calf and mouse ER data sets the difference between the performance of CoMFA and consensus kNN models was surprisingly large. Furthermore, the consensus kNN model employing mouse ER data set performed better than the original kNN model presented by Zheng and Tropsha (2000). The external performance of the consensus kNN models was also impressive with Pr-R^2 values ranging from 0.62 for human ERβ up to 0.77 for calf ER and mouse ER (see Table 3 and Figure 1 in publication **III** for the results of all external models). The effect of the ER data set of different species on the predictive ability of the consensus kNN models is discussed further in section 4.5.

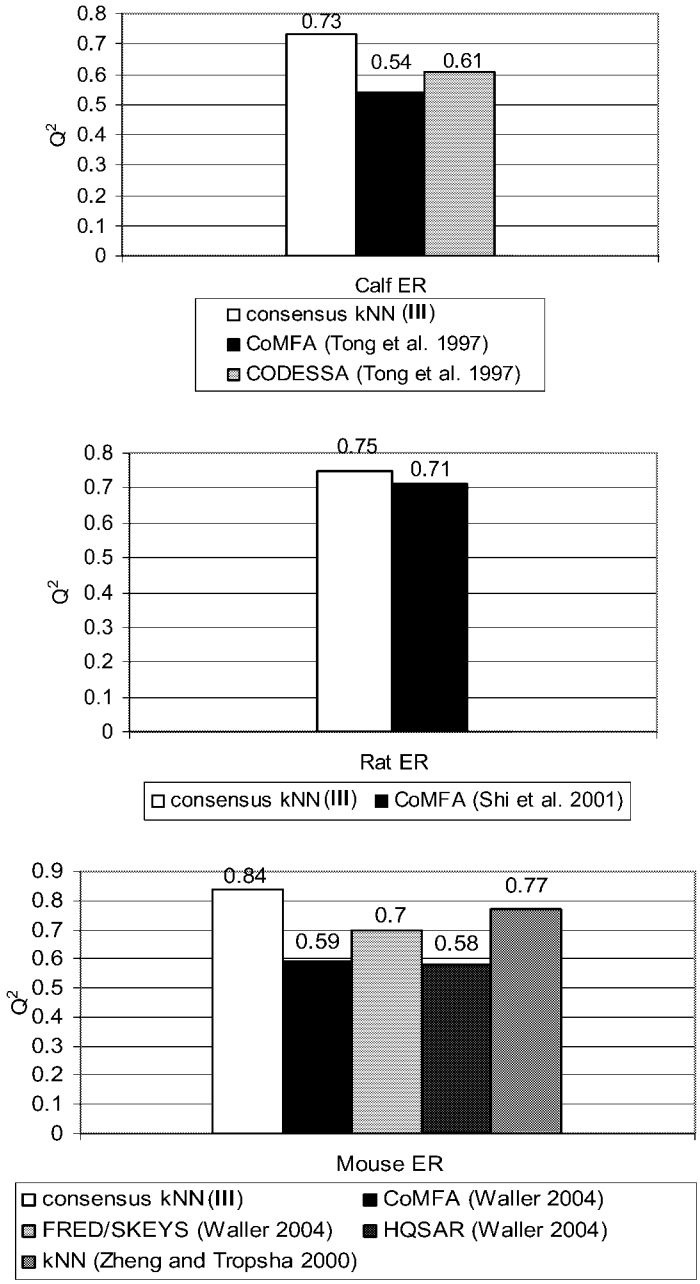


Figure 7. The Q^2 values of the consensus kNN models with calf ER, rat ER and mouse ER data sets (**III**) compared with the results obtained in previous studies employing CoMFA and CODESSA with calf ER (Tong et al. 1997), CoMFA with rat ER data (Shi et al. 2001), CoMFA, FRED/SKEYS, HQSAR (Waller 2004) and kNN (Zheng and Tropsha 2000) with mouse ER data.

4.1.5 Classification tests

The performance of the three classification methods (DT, LVQ, and kNN) was tested with the same data sets of calf ER, mouse ER, rat ER, human ERα and human ERβ (**IV**), which were employed with the above described consensus kNN models, along with an additional data set including 311 compounds regardless of the species, by classifying the compounds into inactive and active groups. The classification power of the PCA/TOOLDIAG selected variables was tested with LOO classification performed with TOOLDIAG. The percentages of the correctly classified compounds varied from 80.8 (rat ER) to 99.3 (calf ER) confirming the classification ability of the selected variables (see Table 2 in publication **IV**). Result obtained with the calf ER data was exceptionally good and it differed markedly from the results obtained with the other data sets.

The external performance of DT, LVQ and kNN was tested with 30 randomized test runs, and the average of the correctly classified compounds in the 30 test sets is presented in Figure 8. The detailed results can be found in Table 3 in publication **IV**.

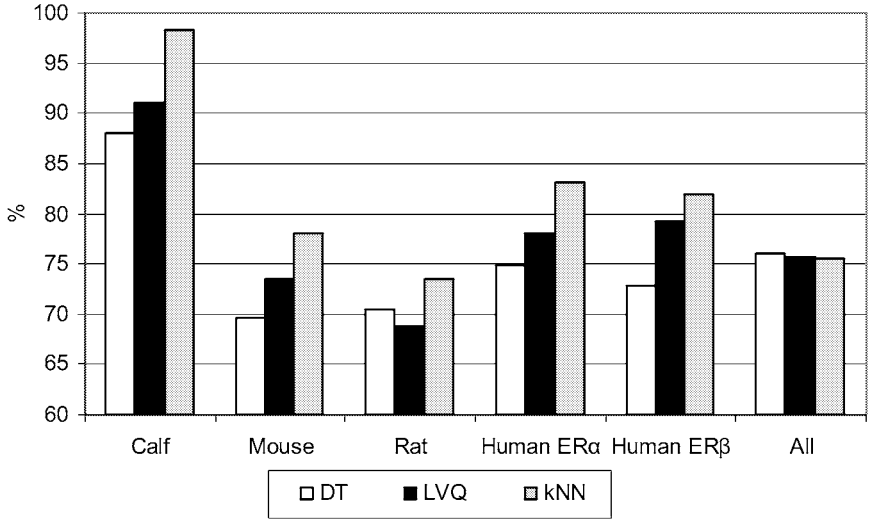


Figure 8. The averages of the percentages of correctly classified compounds for the 30 randomized external test runs with DT, LVQ and kNN methods (**IV**).

The trend of the classification power between species for the external test runs follows the trend obtained with LOO classification performed with TOOLDIAG. The performance of all three methods with the calf ER data set is clearly better than the performance with the other data sets. This can be explained by the homogeneity of the structures of the compounds in the calf data set, which makes the classification simpler. Furthermore, the active compounds were mixed with the inactive ones in the other data sets, whereas they were more clearly divided into two groups in the calf data, which was demonstrated with Sammon's mappings (Figure 2 in publication **IV**).

Comparison of the results between the employed classification methods reveals that kNN performs best for all other data sets except with that including all the compounds, in that case virtually the same performance was achieved with all methods but overall, DT had the weakest performance. Roggo et al. (2003) have noted the same trend between the performances of the DT, LVQ and kNN in their classification study with IR data. The fact

that all the methods tested, including DT, performed equally well with the data set that included all the compounds indicates that the classificatory power of the LVQ and kNN methods suffers if a broad diversity of structures is to be tested.

Tong et al. (2003) tested DT and decision forest (i.e. model ensemble DT) for their classification ability with estrogenic compounds, achieving models with good classificatory powers. Therefore the weak performance of the decision tree in our tests was unexpected. Even pruning of the trees, i.e. deletion of improper decision links in order to simplify the structure of the trees, did not improve the performance of the DT models in our case.

To find some point of comparison for the performance of our models, a data set employed by Saliner et al. (2003) in their multidimensional discriminant analysis (MDA) study, was used with DT, LVQ and kNN models. The data set of 117 compounds was randomly divided into training and test sets of equal size, and the calculations were repeated 30 times. Although our results are not fully comparable with the results obtained by Saliner et al. (2003) (as they reported the percentages of correctly classified compounds for two randomly selected test sets only and without naming the compounds selected for the test sets), the conclusion can be drawn that the performance of DT, LVQ and kNN is as good as that of MDA (see Table 4 in publication **IV** for detailed numerical values of these models). In addition, this comparison revealed that LVQ (not kNN as with our data sets) had the best classification power with this data set.

4.2 Modelling of CYP data

4.2.1 Oxidation of naphthalene

The oxidation tests of naphthalene to 1- and 2-naphthol by mouse CYP2A5 and human CYP2A6 cytochrome P450 enzymes revealed that this metabolic route really does take place, and the oxidation probably proceeds via 1R,2S- and 1S,2R-naphthalene oxides. The results also indicated that 1-naphthol was the dominant oxidation product over 2-naphthol, which supports the results obtained earlier by Jerina et al. (1970), and that the mouse CYP2A5 enzyme is more active than the human CYP2A6 enzyme in the oxidation of naphthalene to 1- and 2-naphthol (**V**).

Admittedly, the oxidation tests of naphthalene conducted within this research were very limited, and thus provided only weak support for the earlier results relating estrogenic activity and metabolism via cytochrome P450 enzyme (Charles et al. 2000; Sugihara et al. 2000; Fertuck et al. 2001; Sanoh et al. 2002; Fujimoto et al. 2003; Kitamura et al. 2003; Mikamo et al. 2003; van Lipzig et al. 2005). In reality, the study of the oxidation of the naphthalene was a minor portion of the study, e.g. we did not evaluate the estrogenic activity of 1- and 2-naphthol, which would have provided further information about the subject.

4.2.2 Modelling of the inhibitory activity of coumarin 7-hydroxylation

The inhibitory activity of various compounds (see Figure 1 in Publication **VI** for the structures) against coumarin 7-hydroxylation was modelled with PLS employing EEVA and EVA descriptors, and Hansch-type QSAR methods using data sets of mouse CYP2A5 and human CYP2A6 enzymes (**VI**), which were used also by Poso et al. (2001)

with CoMFA and GRID/GOLPE methods. Additional MLR models were constructed using one PLS component derived from EEVA and EVA descriptors with McGowan's characteristic volume (MgVol) and its square, which were found to be good predictors by Hansch et al. (2004) in their QSAR study employing the same CYP2A5 data set, as the supplement descriptor. The performances of all models were tested with an external test set including five compounds (compounds 23-28), and additionally with 500 randomized external tests with PLS/EEVA, PLS/EVA, and Hansch-type methods. The Pr-R^2 values for the models with the five compound test set are presented in Figure 9. The detailed statistical parameters for all models with five compound test set can be found in Table 3, and the corresponding parameters for 500 randomized external tests with EEVA, EVA and Hansch-type models in Table 4 in publication **VI**.

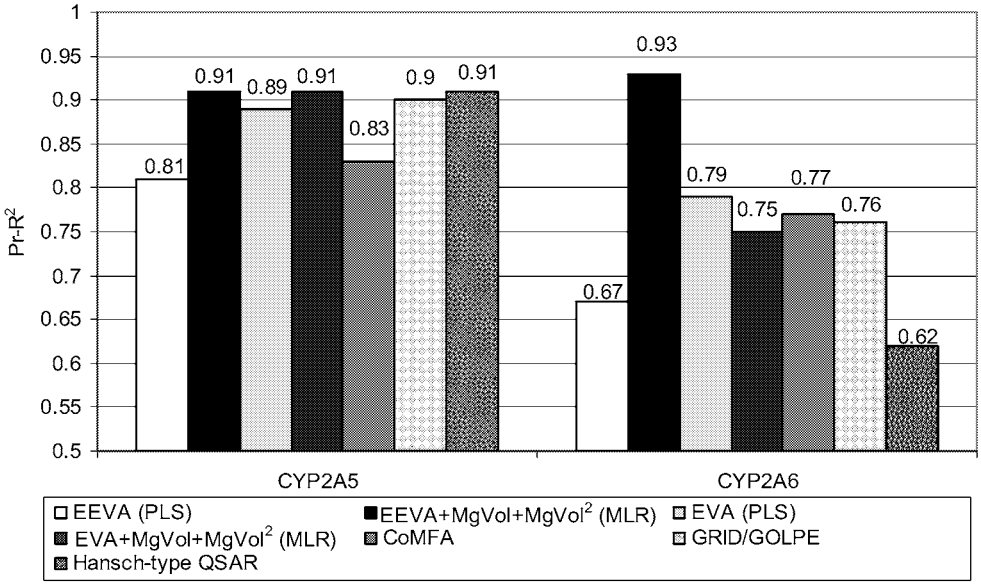


Figure 9. The Pr-R^2 values for the models predicting the inhibitory activity of coumarin 7-hydroxylation for CYP2A5 and CYP2A6 enzymes with an external test set of five compounds (**VI**). CoMFA and GRID/GOLPE results obtained from Poso et al. (2001).

As can be seen with the CYP2A5 data set, all the methods have a good level of performance, although PLS/EEVA and CoMFA do exhibit somewhat weaker performances than the other methods. Furthermore, almost all methods showed clearly better performance with the CYP2A5 data set compared to the CYP2A6 data set, which can be seen especially from the results of Hansch-type QSAR, although MLR/EEVA with MgVol descriptors seemed to manage both data sets equally well. The different kind of relationship between the MgVol descriptor and the activity of the inhibitors for the CYP2A5 and CYP2A5 enzymes (Figure 3 in publication **VI**) seems to provide one possible explanation for this difference between the models employing CYP2A5 or CYP2A6 data sets. Further, the results suggest that the flexibility of the CYP2A5 binding site is larger than that of CYP2A6, i.e. the former tolerates somewhat larger ligands. As a close relationship between octanol-water partition coefficient (logP) and biological

activity is a common phenomenon, logP was tested as an additional descriptor with the EEVA and EVA models, but this did not improve the predictive power of the models.

As a case study, the CoMFA model (**V**) originally presented by Poso et al. (2001) and spectroscopic EEVA and EVA models (**VI**) were used for the prediction of the inhibitory activity of naphthalene against coumarin 7-hydroxylation. The results of these tests are presented in detail in Table 5 of publication **VI**. It appeared that the CoMFA model failed to find any difference between the inhibitory activity of naphthalene for CYP2A5 and CYP2A6 data sets with predicted values of 3.74 and 3.72 (experimental values 4.3 and 3.22, respectively), whereas EEVA and EVA models did detect this difference, although not to the same extent as revealed in the experimental tests. One explanation for the failure of the CoMFA may be the difference between the structures of the compounds used for building the model and the structure of naphthalene, which may cause problems for the alignment leading to unsatisfactory predictions. The performance of the models with CYP2A5 data was better than with CYP2A6 data, as was to be expected based on the results obtained with the models described above. In the comparison between EEVA and EVA models, EEVA outperformed EVA, especially with CYP2A6 data set.

4.3 Variable selection

The selection of variables is a common problem in all predictive modelling, including SAR/QSAR. With the spectroscopic methods tested in this research and with the methods based on the calculation of some molecular fields, such as CoMFA and SOMFA, separate variable selection is not commonly used, although Sippl (2002) obtained clearly better performing CoMFA models employing smart region definition (SRD) as the variable selection. With the method employing a large number of separately calculated physico-chemical features or structural descriptors as the variables, some kind of variable selection is an essential part of the modelling process. DRAGON descriptors were used as variables in our kNN and consensus kNN models employing SA as the variable selection method and in the classification models employing PCA and TOOLDIAGs feature selection for variable selection.

With the kNN method the effect of the variable selection for the predictive power of the models can be illustrated by examining Table II in publication **II**, where the performance of the model including all the variables (Int_{1247}) is clearly poorer than the performance of the other models employing SA. The effect of the selected variable pool can also be seen by viewing the fluctuation of the predictive power example inside the internal model Int_{250} for which the minimum of the Q^2 value is 0.60 and the maximum is 0.75 (Table II in publication **II**) or inside the external model Ext_{250} for which the minimum of the Pr-R^2 value is 0.33 and the maximum is 0.71 (Table IV in publication **II**). These differences are based solely on the different variable pools. The above described effect of the variable pool with kNN employing SA could be diminished by running the SA procedure more thoroughly, i.e. testing more combinations of the original variables, so that the best performing subset is more likely to be found. With consensus kNN, however, this was not needed, since consensus kNN diminishes the effect of the selected variable pool and performs properly without running SA so thoroughly. The shortcoming of the SA is that the procedure is very time consuming, especially with large data sets.

PCA/TOOLDIAG employed with classification methods also proved to be a feasible strategy for the discrimination between active and inactive compounds; at least their use produced better performing models than that obtained by using all of the calculated DRAGON descriptors. Of course, the SA method could have been employed also with classification models, which would have provided an opportunity to use consensus modelling with these models, but this procedure would have been computationally demanding and time-consuming, and therefore the use of PCA was justified.

4.4 Selection of the training and test set

The random selection (usually 500 runs) of the training and test set was used for the external validation of almost all of the models constructed here, except for the kNN and consensus kNN QSAR models. The effect of the composition of the training and the test set can be seen from the results obtained with CoSA ^{13}C and SOMFA methods for external prediction producing Pr-R^2 values with the standard deviations from 0.12 to 0.22 (Table 3 in publication **I**) and from the results obtained with classification methods tested producing as much as 30% difference between the performances of the worst and best models obtained with the same method (see, for example, the performance of the LVQ with the human ERα data set in Table 3 in publication **IV**). Saliner et al. (2003) used the random selection of the training and test set for two runs and selection of the training and test set on the basis of the distribution of the most significant descriptors for two runs. As expected, they obtained poorer results with the randomly selected training and test set, indicating random selection to be the most revealing way of testing the true performance of a model. This was also observed with the kNN model employing the rat ER data set by producing 500 randomized external models with the average performance of Pr-R^2 value being 0.40 (Figure 2 in publication **II**) compared to the 50 models employing the same training and test set (i.e. not random selection) with different descriptor pools producing models with the average of Pr-R^2 value of 0.57. In view of the above, it seems reasonable to propose that validation tests should be conducted in each QSAR study, as emphasized recently by Tropsha et al. (2003).

However, with kNN and consensus kNN methods, the use of randomly selected training and test sets is not so straightforward. The performance of the external model fluctuates considerably even with the same training and test set due to the different variable pools, as described in the previous chapter. For that reason, it would be impossible to know which one of the two, the training set or the variable pool, is having the greater impact on the predictive power of the model. Moreover, with the consensus kNN method, the final prediction is the average of the models used for the consensus model, and so it is essential that the compounds in the training and test sets are the same for all models, if not, then the calculation of the consensus model becomes impossible.

4.5 Diversity and source of the data set

At present, one important reason for the unreliability of the SAR and QSAR models has been the small number of the experimental studies, at least in the case of ER. The usability of the models in the real life screening situations will increase as the structural diversity of the compounds used for the modelling increases. In this study, the limiting effect of too small structural variation in compounds used for the building of the model

was noted with kNN and consensus kNN models, for which the model applicability domain test (Tropsha et al. 2003) was performed in order to find out whether the test set compounds were too dissimilar with the compounds used for building the model and this would account for the poor predictive ability of the models. Indeed, some compounds were found to be too dissimilar, and the external predictive ability of the first consensus kNN model employing rat ER data set (**II**) improved clearly after these excessively dissimilar compounds were omitted (Pr-R^2 value increased from 0.65 to 0.84). However, in the second study employing consensus kNN, the predictive ability of the models did not improve to any major extent by omission of the compounds considered as being too dissimilar (Table 3 in publication **III**), and this indicates that consensus kNN is able to diminish the impact of those compounds.

As present, since there are not experimental values available for compounds covering a truly wide structural spectrum, the predictive power of the models will suffer, if there are some dissimilar structures in the data set used for modelling and prediction. This was noted in the classification tests performed in this research, as the classification ability of the methods tested was clearly better with the calf data set comprising structurally similar compounds than that with other data sets containing compounds with more diverse structures. However, some difficulties may also arise as the structural diversity of the compounds increases. In particular, it causes problems for the 3D-methods requiring alignment of the compounds before QSAR modelling. The alignment becomes more and more complicated and unreliable as the structural variation of the compounds increases and the structures become more complex. This favours the use of QSAR methods which do not need any structural alignments in the models, such as the consensus kNN technique developed and validated in this project.

The data source may also cause some problems and increase the unreliability of the models. The different methods used to perform the activity assays can give dissimilar activities for the same compound. This can be avoided by using only the activities obtained from the biological assays performed with the same method, but this reduces even more the small number of compounds for which the data is available. In the studies presented in this thesis, the difference between the source and particularly the difference between the ER (human α, human β, rat, mouse or calf) or the enzyme (CYP2A5 or CYP2A6) used for binding affinity measurements influenced the predictive powers of the models. With the CYP2A5 and CYP2A6 data sets, this has been discussed in section 4.2.2 and the situation is not so complicated, because both enzyme data sets contain the same compounds. In contrast, the ER data sets are more complicated, since each data set comprises different compounds, except for the data sets of human ERα and human ERβ, which have only two compounds which are not common for both data sets. Surprisingly, the predictive power of the consensus kNN models employing the ER data sets grouped according to the species (**III**), were clearly different for human ERα and human ERβ (Figure 10), but unfortunately we do not have any sound explanation to offer for this phenomenon. The low SDEP values in the external consensus kNN models obtained with the calf ER data set are clearly attributable to the smaller variation between the binding affinities of compounds included, whereas the high SDEP value with the rat ER data set can be explained by the greater variation in the binding affinities.

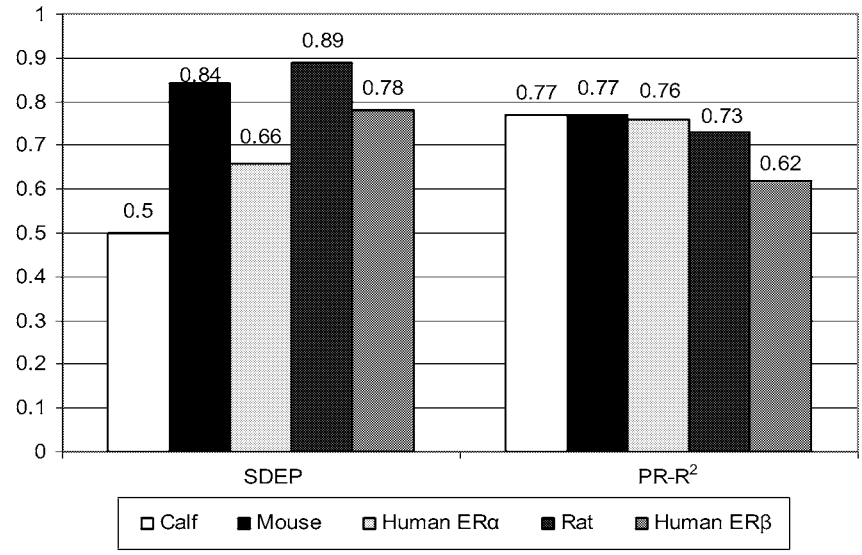


Figure 10. The SDEP and Pr-R^2 values for the consensus kNN models with the data sets based on the different species (**III**).

5 CONCLUSIONS

The following conclusions can be drawn based on the results obtained and observations made in this thesis:

- Spectroscopic CoSA with ^{1}H NMR chemical shifts, and especially PLS/EEVA and PLS/EVA methods did not handle enantiomers and symmetric compounds properly in the data set used for modelling, whereas CoSA with ^{13}C NMR chemical shifts performed better.
- SOMFA produced promising results with the tested estrogenic data sets, and it also handled enantiomers and symmetric compounds included among the test compounds, but the use of SOMFA is limited to structurally homogenous data sets due to the fact that alignment is needed for modelling.
- Estrogenic activity can be predicted with good accuracy employing consensus kNN method. This technique produces better results than have previously been obtained employing CoMFA, HQSAR, CODESSA and FRED/SKEYS methods with the same data set.
- LVQ and kNN were suitable methods for the classification of estrogenic compounds especially if there was a structurally homogenous data set.
- Naphthalene was oxidized by CYP2A5 and CYP2A6 enzymes to 1- and 2-naphthols, evidencing that compounds having no estrogenic activity can be oxidized by cytochrome P450 enzymes to compounds potentially possessing estrogenic activity.
- EEVA outperformed EVA and performed more reliably and constantly when employing the cytochrome P450 data set. Both EEVA and EVA performed better than CoMFA.
- Hansch-type QSAR is a very simple method to perform and it produced predictions which were comparable to the more complicated QSAR methods with congeneric data sets.
- The performance of QSAR methods was greatly influenced by the data set used, especially by the structural variability between the compounds.
- PLS/EEVA and consensus kNN can be valuable tools for preliminary testing of estrogenic activity of the compounds, especially in the situation where the activity of individual compounds is to be tested (i.e. in the development of drugs or new chemicals). LVQ and kNN can be employed in the same situation for classification purposes.
- Simulated annealing is a suitable variable selection method, but it is time-consuming when employing large data sets.
- Molecular dynamics calculations are too laborious and excessively time-consuming and thus cannot be considered as supporting computational tools in practical situations where one needs to evaluate large data sets.

6 REFERENCES

Anstead, G.M.; Carlson, K.E.; Katzenellenbogen, J.A. (1997). The estradiol pharmacophore: Ligand structure – estrogen receptor binding activity relationships and a model for binding site. Steroids 62, 268-303.

Barkhem, T.; Carlsson, B.; Nilsson, Y.; Enmark, E.; Gustafsson, J.-Å. Nilsson, S. (1998). Differential response of estrogen receptor α and estrogen receptor β to partial estrogen agonists/antagonists. Mol. Pharmacol. 54, 105–112.

Baurin, N.; Mozziconacci, J-C.; Aenoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. (2004). 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modelling and screening of the NCI database. J. Chem. Inf. Comput. Sci. 44, 276-285.

Beger, R.D.; Freeman, J.P.; Lay, J.O.; Wilkes, J.G.; Miller, D.W. (2000). C-13 NMR and electron ionization mass spectrometric data-activity relationship model of estrogen receptor binding. Toxicol. App. Pharm. 169, 17-25.

Beger, R.D.; Freeman, J.P.; Lay, J.O.; Wilkes, J.G.; Miller, D.W. (2001a). Use of C-13 NMR spectrometric data to produce a predictive model of estrogen receptor binding activity. J. Chem. Inf. Comput. Sci. 41, 219-224.

Beger, R.D.; Buzatu, D.A.; Wilkes, J.G.; Lay, J.O., Jr. (2001b). ^{13}C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroids binding the aromatase enzyme. J. Chem. Inf. Comput. Sci. 41, 1360-1366.

Beger, R.D.; Wilkes, J.G. (2001a). Models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding affinity to the aryl hydrocarbon receptor developed using ^{13}C NMR data. J. Chem. Inf. Comput. Sci. 41, 1322-1329.

Beger, R.D.; Wilkes, J.G. (2001b). Developing ^{13}C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. J. Comp.- Aid. Mol. Des. 15, 659-669.

Borgelt, C. Codes for the decision tree modelling (http://fuzzy.cs.uni-magdburg.de/~borgelt/doc/dtree/dtree.html)

Breinman, L.; Friedman, J.H.; Olshen, R.A.; Stone, J.C. (1984). Classification and regression trees. Wandsworth, Pacific Grove, CA, USA.

Brzozowski, A.M.; Pike, A.C.W.; Dauter, Z.; Hubbard, R.E.; Bonn, T.; Engström, O.; Öhman, L.; Greene, G.L.; Gustafsson, J.-Å.; Carlquist, M. (1997). Molecular basis of agonism and antagonism in the oestrogen receptor. Nature 389, 753-758.

Bursi, R.; Dao, T.; van Wijk, T.; De Gooyer, M.; Kellenbach, E.; Verwer, P. (1999). Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. J. Chem. Inf. Comput. Sci. 39, 861-867.

Charles, G.D.; Bartels, M.J.; Zacharewski, T.R.; Gollapudi, B.B.; Freshour, N.L.; Carney, E.W. (2000). Activity of benzo[a]pyrene and its hydroxylated metabolites in an estrogen receptor-alpha reporter gene assay. Toxicol. Sci. 55, 320-326.

Coleman, K.P.; Toscano, W.A. Jr.; Wiese, T.E. (2003). QSAR models of the *in vitro* estrogen activity of bisphenol A analogs. QSAR Comb. Sci. 22, 78-88.

Cooper, R.L.; Kavlock, R.J. (1997). Endocrine disruptors and reproductive development: A weight-of-evidence overview. J. Endocrinol. 152, 159–166.

Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. (1998). Comparative molecular-field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J. Am. Chem. Soc. 110, 5959-5967.

Danzo, B.J. (1998). The effects of environmental hormones on reproduction. Cell. Mol. Life Sci. 54, 1249–1264.

Devillers, J.; Domine, D.; Guillon, C.; Karcher, W. (1998). Simulating lipophilicity of organic molecules with a back-propagation neural network. J. Pharmacol. Sci. 87, 1086-1090.

Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. (1985). AM1: A new general purpose quantum-mechanical molecular model. J. Am. Chem. Soc. 107, 3902-3909.

EPA (U.S. Government Environmental Protection Agency) (1998). Endocrine disruptor screening and testing advisory committee (EDSTAC), final report. Available at: http://www.epa.gov/scipoly/oscpendo/edspoverview/finalrpt.htm.

Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Health Persp. 111, 1361-1375.

Fang, H.; Tong, W.; Shi, L.M.; Blair, R.; Perkins, R.; Branham, W.; Hass, B.S.; Xie, Q.; Dial, S.T.; Moland, C.L.; Sheehan, D.M. (2001). Structure-activity relationship for a large diverse set of natural, synthetic, and environmental estrogens. Chem. Res. Toxicol. 14, 280-294.

Ferguson, A.M.; Heritage, T.; Jonathon, P.; Pack, S.E.; Phillips, L.; Rogan, J.; Snaith, P.J. (1997). EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. J. Comp.- Aided Mol. Des. 11, 143-152.

Fertuck, K.C.; Matthews, J.B.; Zacharewski, T.R. (2001). Hydroxylated benzo[a]pyrene metabolites are responsible for in vitro estrogen receptor-mediated gene expression induced by benzo[a]pyrene, but do not elicit uterotrophic effects in vivo. Toxicol. Sci. 59, 231-240.

Fujimoto, T.; Kitamura, S.; Sanoh, S.; Sugihara, K.; Yoshihara, S.; Fujimoto, N.; Ohta, S. (2003). Estrogenic activity of an environmental pollutant, 2-nitrofluorene, after metabolic activation by rat liver microsomes. Biochem. Bioph. Res. Comm. 303, 419-426.

Gaido, K.W.; Maness, S.C.; Waters, K.M. (1999). Exploring the biology and toxicology of estrogen receptor β. CIIT Activities 19, 1-10.

Gantchev, T.G.; Ali, H.; van Lier, J.E. (1994). Quantitative structure-activity relationships/compative molecular field analysis (QSAR/CoMFA) for receptor-binding properties of halogenated estradiol derivatives. J. Med. Chem. 37, 4164-4176.

Gao, H.; Katzenellenbogen, J.A.; Garg, R.; Hansch, C. (1999). Comparative QSAR analysis of estrogen receptor ligands. Chem. Rev. 99, 723-744.

Geladi, P.; Kowalski, B. (1986). Partial least-squares regression: A tutorial. Anal. Chim. Acta 185, 1-17.

Golbraikh, A.; Tropsha, A. (2002a). Beware of q^2! J. Mol. Graph. Model. 20, 269–276.

Golbraikh, A.; Tropsha, A. (2002b). Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J. Comp.- Aided Mol. Des. 16, 357-369.

Gramatica, P.; Pilutti, P.; Papa, E. (2004). Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modelling. J. Chem. Inf. Comput. Sci. 44, 1794-1802.

Gustafsson, J.-Å. (1999). Estrogen receptor β – a new dimension in estrogen mechanism of action. J. Endocrinol. 163, 379–383.

Halgren, T.A. (1992). Representation of Vanderwaals (VDW) interactions in molecular mechanics force-fields – potential form, combination rules, and VDW parameters. J. Am. Chem. Soc. 114, 7827-7843.

Hansch, C.; Muir, R.M.; Fujita, T.; Maloney, P.P.; Geiger, E.; Streich, M. (1963). The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. J. Am. Chem. Soc. 85, 2817-2824.

Hansch C, Fujita T. (1964). ρ-δ-π analysis. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc. 86, 1616-1626.

Hansch, C.; Leo, A. (1995). Exploring QSAR. Fundamentals and applications in chemistry and biology. American Chemical Society, Washington, D.C.

Hansch, C.; Mekapati, S.B.; Kurup, A.; Verma, R.P. (2004). QSAR of cytochrome P450. Drug Metabol. Rev. 36, 105-156.

Hanson, R.N.; Lee, C.Y.; Friel, C.J.; Dilis, R.; Hughes, A.; DeSombre, E.R. (2003). Synthesis and evaluation of 17α-20E-21-(4-substituted phenyl)-19-norpregna-1,3,5(10),20-tetraene-3,17β-diols as probes for the estrogen receptor α hormone binding domain. J. Med. Chem. 46, 2865-2876.

Hong, H.; Tong, W.; Fang, H.; Shi, L.; Xie, Q.; Wu, J.; Perkins, R.; Walker, J.D.; Branham, W.; Sheehan, D.M. (2002). Prediction of estrogen receptor binding for 58000 chemicals using an integrated system of a tree-based model with structural alerts. Environ. Health Perspect. 110, 29-36.

Jerina, D.M.; Daly, J.W. Zaltzman-Nirenberg, P.; Udenfriend, S. (1970). 1,2-Naphthalene oxide as an intermediate in the microsomal hydroxylation of naphthalene. Biochem. 9, 147-155.

Juvonen, R.; Gynther, J.; Pasanen, M.; Alhava, E.; Poso, A. (2000). Pronounced differences in inhibition potency of lactone and non-lactone compounds for mouse and human coumarin 7-hydroxylases (CYP2A5 and CYP2A6). Xenobiotica 30, 81-92.

Jobling, S. (1998). Review of suggested testing methods for endocrine-disrupting chemicals. Pure Appl. Chem. 70, 1805-1827.

Kangas, J.; Kohonen, T. (1996). Developments and applications of the self-organizing map and related algorithms. Math. Comput. Simulat. 41, 3-12.

Kirkpatrick, S.; Gelatt, C.D. Jr.; Vecchi, M.P. (1983). Optimization by simulated annealing. Science 220, 671-680.

Kitamura, S.; Sanoh, S.; Kohta, R.; Suzuki, T.; Sugihara, K.; Fujimoto, N.; Ohta, S. (2003). Metabolic activation of proestrogenic diphenyl and related compounds by rat liver microsomes. J. Health Sci. 49, 298-310.

Klebe, G.; Abraham, U.; Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J. Med. Chem. 37, 4130-4146.

Kohonen, T. (1998). The self-organizing map. Neurocomputing 21, 1-6.

Kohonen, T. (2001) Self-organizing maps. Springer Series in Information Sciences, 30, Springer, Heidelberg, 3rd ed.

Kollman, P.A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D.A.; Cheatham, T.E. III. (2000). Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. Acc. Chem. Res. 33, 889–897.

Kong, E.H.; Pike, A.C.W.; Hubbard, R.E. (2003). Structure and mechanism of the oestrogen receptor. Biochem. Soc. Transact. 31, 56-59.

Kuiper, G.J.M.; Carlsson, B.; Grandien, K.; Enmark, E.; Häggblad, J.; Nilsson, S.; Gustafsson, J.-Å. (1997). Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors α and β. J. Endocrinol. 138, 863-870.

Li, M.Y.; Du, L.P.; Wu, B.; Xia, L. (2003). Self-organizing molecular field analysis on alpha(1a)-adrenoceptor dihydropyridine antagonists. Bio-org. Med. Chem. 11, 3945-3951.

Li, J.J.; Li, S.A. (1998). Breast cancer: Evidence for xeno-oestrogens involvement in altering its incidence and risk. Pure Appl. Chem. 70, 1713-1723.

Lintelmann, J.; Katayama, A.; Kuhihara, N.; Shore, L.; Wenzel, A. (2003). Endocrine disruptors in the environment (IUPAC Technical Report). Pure Appl. Chem. 75, 631–681.

Lipnick, R.L. (1989). Hans Horst Meyer and the lipoid theory of narcosis, Trends Pharmacol. Sci., 10, 265-269.

Manallack, D.; Pitt, W.; Gancia, E.; Montana, J.; Livingstone, D.; Ford, M.; Whitley, D. (2002). Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks. J. Chem. Inf. Comput. Sci. 42, 1256-1262.

Martinek, T.A.; Otvos, F.; Dervarics, M.; Toth, G.; Fulop, F. (2005). Ligand-based prediction of active conformation by 3D-QSAR flexibility descriptors and their application in 3+3D-QSAR models. J. Med. Chem. 48, 3239-3250.

McLachlan, J.A. (2001). Environmental signaling: What embryos and evolution teach us about endocrine disrupting chemicals. Endocrin. Rev. 22, 319–341.

Mikamo, E.; Harada, S.; Nishikawa, J.; Nishihara, T. (2003). Methoxychlor induces CYP2C11 to convert itself into hormonally active metabolites. J. Health Sci. 49, 229-232.

Morgan, J.N.; Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. J. Am. Stat. Assoc. 58, 415-434.

Napolitano, E.; Fiaschi, R.; Carlson, K.E.; Katzenellenbogen, J.A. (1995). 11β-substituted estradiol derivatives, potential high-affinity carbon-11-labeled probes for the estrogen receptor: A structure-affinity relationship study. J. Med. Chem. 38, 429-434.

Oostenbrink, B.C.; Pitera, J.W.; van Lipzig, M.M.H.; Meerman, J.H.N.; van Gunsteren, W.F. (2000). Simulations of the estrogen receptor ligand-binding domain: Affinity of natural ligands and xenoestrogens. J. Med. Chem. 43, 4594-4605.

Pearlman, D.A.; Case, D.A.; Caldwell, J.W.; Ross, W.R.; Cheatham, T.E. III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. (1995). AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. Comp. Phys. Commun. 91, 1-41.

Perkins, R.; Fang, H.; Tong, W.; Welsh, W.J. (2003). Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology. Environ. Toxicol. Chem. 22, 1666-1679.

Ponder, J.W.; Case, D.A. (2003). Force fields for protein simulations. Adv. Prot. Chem. 66, 27-85.

Poso, A.; Gynther, J.; Juvonen, R. (2001). A comparative molecular field analysis of cytochrome P450 2A5 and 2A6 inhibitors. J. Comp.- Aided Mol. Des. 15, 195–202.

Rauber, T. W. A toolbox for analysis and visualization of sensor data in supervision (http://www.inf.ufes.br/~thomas/home/tooldiag.html).

Reel, J.R.; Lamb, J.C.; Neal, B.H. (1996). Survey and assessment of mammalian estrogen biological assays for hazard characterization. Fund. Appl. Toxicol. 34, 288–305.

Robinson, D.D.; Winn, P.J.; Lyne, P.D.; Richards, W.G. (1999). Self-organizing molecular field analysis: A tool for structure-activity studies. J. Med. Chem. 42, 573-583.

Roggo, Y.; Duponchel, L.; Huvenne, J.P. (2003). Comparison of supervised pattern recognition methods with McNemar's statistical test. Application to qualitative analysis of sugar beet by near-infrared spectroscopy. Anal. Chim. Acta 477, 187-200.

Sadler, B.R.; Cho, S.J.; Ishaq, K.S.; Chae, K.; Korach, K.S. (1998). Three-dimensional quantitative structure-activity relationship study of nonsteroidal estrogen receptor ligands using comparative molecular field analysis/cross-validated r^2-guided region selection approach. J. Med. Chem. 41, 2261-2267.

Saliner, A.G.; Amat, L.; Carbó-Dorca, R.; Schultz, W.T.; Cronin, M.D.T. (2003). Molecular quantum similarity analysis of estrogenic activity. J. Chem. Inf. Comput. Sci. 43, 1166-1176.

Sanoh, S.; Kitamura, S.; Sugihara, K.; Ohta, S. (2002). Cytochrome P450 1A1/2 mediated metabolism of trans-stilbene in rats and humans. Pharm. Bull. 25, 397-400.

Schmieder, P.K.; Ankley, G.; Mekenyan, O.; Walker, J.D. (2003). Quantitative structure-activity relationship models for prediction of estrogen receptor binding affinity of structurally diverse chemicals. Environ. Toxicol. Chem. 22, 1844-1854.

Schultz, T.W.; Cronin, M.T.D.; Walker, J.D.; Aptula, A.O. (2003). Quantitative structure-activity relationships (QSARs) in toxicology: A historical perspective. J. Mol. Struct. (Theochem) 622, 1-22.

Shao, J. (1993). Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88, 486-494.

Shelby, M.D.; Newbold, R.R.; Tully, D.B.; Chae, K.; Davis, V.L. (1996). Assessing environmental chemicals for estrogenicity using a combination of in vitro and in vivo assays. Environ. Health Perspect. 104, 1296-1300.

Shi, L.M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W.S. Dial, S.L.; Moland, C.L.; Sheehan, D.M. (2001). QSAR models using a large diverse set of estrogens. J. Chem. Inf. Comput. Sci. 41, 186-195.

Sippl, W. (2000). Receptor-based 3D QSAR analysis of estrogen receptor ligands – merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. J. Comp.-Aided Mol. Des. 14, 559-572.

Sippl, W. (2002). Binding affinity prediction of novel estrogen receptor ligands using receptor-based 3-D QSAR methods. Bio-org. Med. Chem. 10, 3741-3755.

Smith, P.A.; Sorich, M.J.; McKinnon, R.A.; Miners, J.O. (2003). Pharmacophore and quantitative structure-activity relationship modeling: Complementary approaches for the rationalization and prediction of UDP-glucuronosyltransferase 1A4 substrate selectivity. J. Med. Chem. 46, 1617-1626.

Sonnenschein, C.; Soto, A.M. (1998). An updated review of environmental estrogen and androgen mimics and antagonists. J. Steroid Biochem. Mol. Biol. 65, 143-150.

Sugihara, K.; Kitamura, S.; Sanoh, S.; Ohta, S.; Fujimoto, N.; Maruyama, S.; Ito, A. (2000). Metabolic activation of the proestrogens trans-stilbene and trans-stilbene oxide by rat liver microsomes. Toxicol. Appl. Pharmacol. 167, 46–54.

Tetko, I. (2002). Neural networks studies 4. Introduction to associative neural networks. J. Chem. Inf. Comput. Sci. 42, 717-728.

Tetko, I.; Tanchuk V. (2002). Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J. Chem. Inf. Comput. Sci. 42, 1136-1145.

Todeschini, R.; Consonni, V.; Pavan, M. DRAGON. Software for the calculation of molecular descriptors (http://www.talete.mi.it/dragon.htm).

Todeschini, R.; Consonni, V. (2000). Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry Vol. 11; Wiley-VCH, Weinheim.

Tong, W.; Perkins, R. (1997). QSAR models for binding of estrogenic compounds to estrogen receptor α and β subtypes. J. Endocrinol. 138, 4022-4025.

Tong, W.; Perkins, R.; Strelitz, R.; Collantes, E.R.; Keenan, S.; Welsh, W.J.; Branham, W.S.; Sheehan, D.M. (1997). Quantitative structure-activity relationships (QSARs) for estrogen binding to the estrogen receptor: Predictions across species. Environ. Health Perspect. 105, 1116-1124.

Tong, W.; Lowis, D.R.; Perkins, R.; Chen, Y.; Welsh, W.J.; Goddette, D.W.; Heritage, T.W.; Sheehan, D.M. (1998). Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. J. Chem. Inf. Comput. Sci. 38, 669-677.

Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. (2003). Decision forest: Combining the predictions of multiple independent decision tree models. J. Chem. Inf. Comput. Sci. 43, 525-531.

Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. (2004). Assessment of prediction confidence and domain extrapolation of two structure–activity relationship models for predicting estrogen receptor binding activity. Environ. Health Perspect. 112, 1249–1254.

Tropsha, A.; Gramatica, P.; Gombar, V. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. 22, 69-76.

Tuppurainen, K. (1999). EEVA (electronic eigenvalue): A new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. SAR QSAR Environ. Res. 10, 39-46.

Tuppurainen, K.; Ruuskanen, J. (2000). Electronic eigenvalue (EEVA): A new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. A QSAR approach to the Ah receptor binding affinity of polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs), Chemosphere 41, 843-848.

Tuppurainen, K.; Viisas, M.; Laatikainen, R.; Peräkylä, M. (2002). Evaluation of a novel electronic eigenvalue (EEVA) molecular descriptor for QSAR/QSPR studies: Validation using a benchmark steroid dataset. J. Chem. Inf. Comput. Sci. 42, 607-613.

Tyler, C.R.; Routledge, E.J (1998). Oestrogenic effects in fish in English rivers with evidence of their causation. Pure Appl. Chem. 70, 1795-1804.

van der Kraak, G. (1998). Observations of endocrine effects in wildlife with evidence of their causation. Pure Appl. Chem. 70, 1785-1794.

van Lipzig, M.M.H.; ter Laak, A.M.; Jongejan, A.; Vermeulen, N.P.E.; Wamelink, M.; Geerke, D.; Meerman, J.H.N. (2004). Prediction of ligand binding affinity and orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulation and the linear interaction energy method. J. Med. Chem. 47, 1018-1030.

van Lipzig, M.M.H.; Vermeulen, N.P.E.; Gusinu, R.; Legler, J.; Frank, H.; Seidel, A.; Meerman, J.H.N. (2005a). Formation of estrogenic metabolites of benzo[a]pyrene and chrysene by cytochrome P450 activity and their combined and supra-maximal estrogenic activity. Environ. Toxicol. Pharm. 19, 41–55.

van Lipzig, M.M.H.; Commandeur, J.N.; de Kanter, F.J.J.; Damsten, M.C.; Vermeulen, N.P.E.; Maat, E.; Groot, E.J.; Brouwer, A.; Kester, M.H.A.; Visser, T.J.; Meerman, J.H.N. (2005b). Bioactivation of dibrominated biphenyls by cytochrome P450 activity to metabolites with estrogenic activity and estrogen sulfotransferase inhibition capacity. Chem. Res. Toxicol. 18, 1691-1700.

Vorobjev, Y.N.; Almagro, J.C.; Hermans, J. (1998). Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum models. Proteins 32, 399–413.

Votano, J.R.; Parham, M.; Hall, L.H.; Kier, L.B.; Oloff, S.; Tropsha, A.; Xie, Q.A.; Tong, W. (2004). Three new consensus QSAR models for the prediction of Ames genotoxicity. Mutagenesis 19, 365-377.

Vračko, M. (1997). Structure mutagenicity study of 12 trimethylimidazopyridine isomers using orbital energies and "spectrum-like representation" as descriptors. J. Chem. Inf. Comp. Sci. 37, 1037-1043.

Waller, C.L.; Minor, D.L.; McKinney, J.D. (1995). Using three-dimensional quantitative structure-activity relationships to examine estrogen receptor binding affinities of polychlorinated hydroxybiphenyls. Environ. Health Perspect. 103, 702-707.

Waller, C.L.; Oprea, T.I.; Chae, K.; Park, H-K.; Korach, K.S.; Laws, S.C.; Wiese, T.E.; Kelce, W.R.; Gray, L.E. (1996). Ligand-based identification of environmental estrogens. Chem. Res. Toxicol. 9, 1240-1248.

Waller, C.L.A. (2004). A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. J. Chem. Inf. Comput. Sci. 44, 758-765.

Weatherman, R.V.; Fletterick, R.J.; Scanlan, T.S. (1999). Nuclear-receptor ligands and ligand-binding domains. Annu. Rev. Biochem. 68, 559-581.

Wiese, T.E.; Polin, L.A.; Palomino, E.; Brooks, S.C. (1997). Induction of the estrogen specific mitogenic response of MCF-7 cells by selected analogues of estradiol-17β: A 3D QSAR study. J. Med. Chem. 40, 3659-3669.

Wold, S. (1991). Validation of QSARs. Quant. Struct.-Act. Relat. 10, 191-193.

Wold, S.; Esbensen, K.; Geladi, P. (1987). Principal component analysis. Chemom. Intell. Lab. Syst. 2, 37-52.

Wolinski, K.; Hinton, J.F.; Pulay, P. (1990). Efficient implementation of the gauge- independent atomic orbital method for NMR chemical shift calculations. J. Am. Chem. Soc. 112, 8251-8260.

Wurtz, J-M.; Egner, U.; Heinrich, N.; Moras, D.; Mueller-Fahrnow, A. (1998). Three-dimensional models of estrogen receptor ligand binding domain complexes, based on related crystal structures and mutational and structure-activity relationship data. J. Med. Chem.41, 1803-1814.

Xing, L.; Welsh, W.J.; Tong, W.; Perkins, R.; Sheehan, D.M. (1999). Comparison of estrogen receptor α and β subtypes based on comparative molecular field analysis (CoMFA). SAR QSAR Environ. Res. 10, 215-237.

Zacharewski, T. (1998). Identification and assessment of endocrine disruptors: Limitations of in vivo and in vitro assays. Eviron. Health Perspect. 106 (Suppl. 2), 577-582.

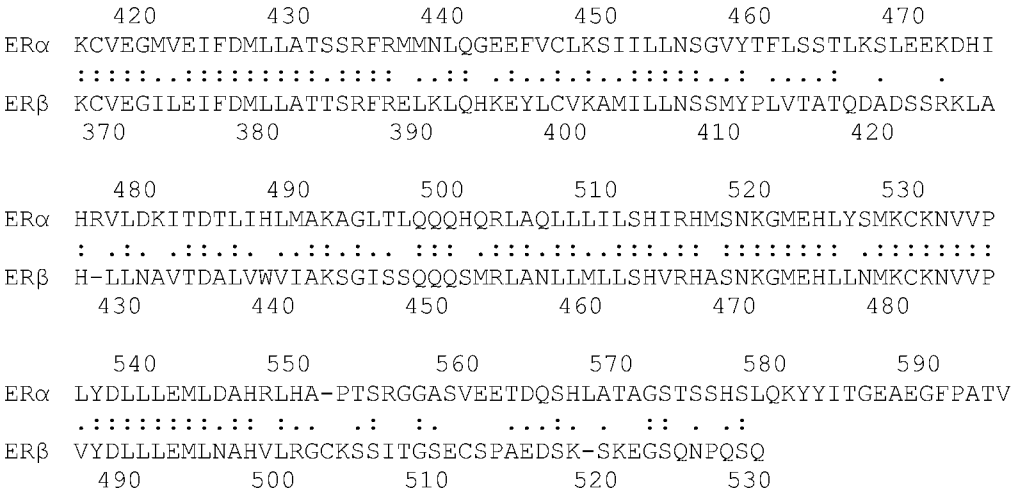
Zheng, W.; Tropsha, A. (2000). Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbour principle. J. Chem. Inf. Comput. Sci. 40, 185-194.

APPENDICES

Appendix 1. Amino acid sequences of the human ERα and ERβ.

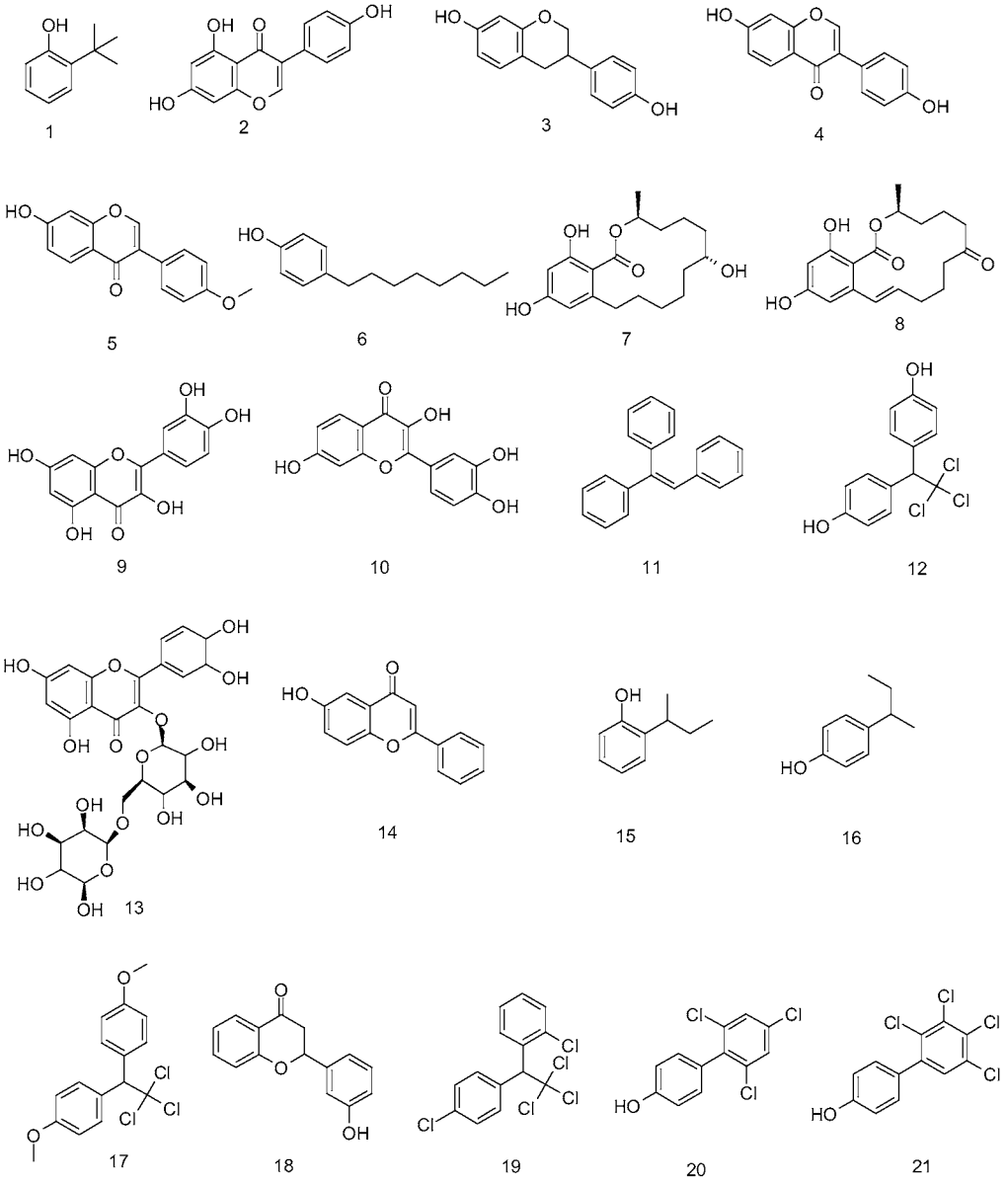
Alignment was performed with protein information resource (PIR) pairwise alignment tool with function SSEARCH [version 3.4t24 July 21, 2004] (http://pir.georgetown.edu/pirwww/search/pairwise.html). The sequences used for the alignment were retrieved from the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) (http://kr.expasy.org/) with the ID numbers of P03372 for ERα and Q92731 for ERβ.

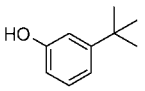
```
            10        20        30        40        50        60
ERα  TMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGEVYLDSSKPAVYNYPEGAAY
                                :  :  .:::.  : .:. ::   : .. :  :.
ERβ            MDIKNSPSSLNSPSSYNCSQSILPLEH--GSIYIPSS--YVDSHHEYPAM
                       10        20          30          40

             70         80        90       100       110
ERα  EFNAAAAANAQVYGQ-TGLPYGPGSEAAAFGSNGLGGFPPLNSVSPSPLMLLHPPPQLSP
      : . :. : .. .. :.:  ::: .                  . :: .:   : .:::
ERβ  TFYSPAVMNYSIPSNVTNLEGGPGRQ-----------------TTSPNVLWPTPGHLSP
       50        60        70                         80

   120       130       140       150         160        170
ERα  FLQPHGQQVPYYLENEPSGYTVREAGPPAFYRP-NSDN-RRQGGRERLAS-TNDKGSMAM
     ..  : :    : : . : .   ::       : : .. .:. . .: :: ..  ::
ERβ  LVV-HRQLSHLYAEPQKSPWC--EARSLEHTLPVNRETLKRKVSGNRCASPVTGPGS---
     90        100         110       120       130       140

     180       190       200       210       220       230
ERα  ESAKETRYCAVCNDYASGYHYGVWSCEGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKS
     .....::::.:::::::::::::::::::::::::::::.:::::::::::::::::
ERβ  --KRDAHFCAVCSDYASGYHYGVWSCEGCKAFFKRSIQGHNDYICPATNQCTIDKNRRKS
          150       160       170       180       190       200

      240       250       260       270       280       290
ERα  CQACRLRKCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVGSAGDMRAANLWPSPL
     ::::::::::::::.: : :..: : :.....:. :.     ..  ::  . ..   .:
ERβ  CQACRLRKCYEVGMVKCGSRRERCGYRLVRRQRSADE-----QLHCAGKAKRSG-GHAPR
           210       220       230             240        250

      300       310       320       330       340       350
ERα  MIKRSKKNSLALSLTADQMVSALLDAEPP-ILYSEYDPTRPFSEASMMGLLTNLADRELV
     .     .. :  .:. .:.: .::.:::: .: :.  :. ::.:::::  ::.:::.:::
ERβ  V-----RELLLDALSPEQLVLTLLEAEPPHVLISR--PSAPFTEASMMMSLTKLADKELV
             260       270       280         290       300

       360       370       380       390       400       410
ERα  HMINWAKRVPGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFAPNLLLDRNQG
     :::.:::..::::.:.: :::.:::  :.:.::.::.:::..:::::.:::.:.::::.:
ERβ  HMISWAKKIPGFVELSLFDQVRLLESCWMEVLMMGLMWRSIDHPGKLIFAPDLVLDRDEG
     310       320       330       340       350       360
```


420 430 440 450 460 470
ERα KCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSGVYTFLSSTLKSLEEKDHI
ERβ KCVEGILEIFDMLLATTSRFRELKLQHKEYLCVKAMILLNSSMYPLVTATQDADSSRKLA
370 380 390 400 410 420
480 490 500 510 520 530
ERα HRVLDKITDTLIHLMAKAGLTLQQQHQRLAQLLLILSHIRHMSNKGMEHLYSMKCKNVVP
ERβ H-LLNAVTDALVWVIAKSGISSQQQSMRLANLLMLLSHVRHASNKGMEHLLNMKCKNVVP
430 440 450 460 470 480
540 550 560 570 580 590
ERα LYDLLLEMLDAHRLHA-PTSRGGASVEETDQSHLATAGSTSSHSLQKYYITGEAEGFPATV
ERβ VYDLLLEMLNAHVLRGCKSSITGSECSPAEDSK-SKEGSQNPQSQ
490 500 510 520 530

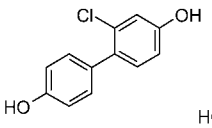
Appendix 2. The structures and the activities of the compounds used in studies **II**, **III** and **IV**.

Figure A2. Structures of the compounds used for the construction of the kNN (rat data from compounds 1-245 and compounds 340-351) (**II**) and consensus kNN (compounds 1-245) (**III**) models. The inactive compounds used in the classification study (**IV**) are numbered 246-339.

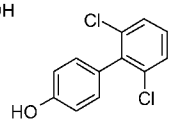




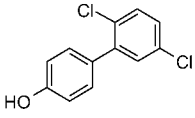
22



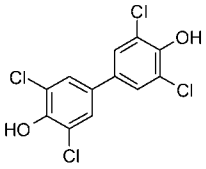
23



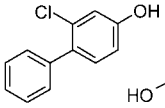
24



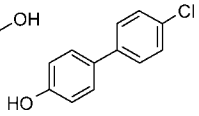
25



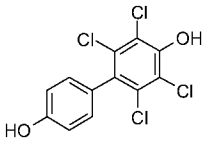
26



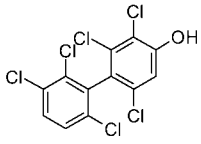
27



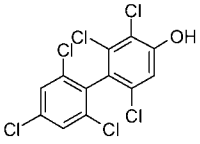
28



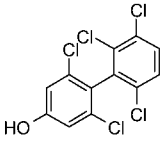
29



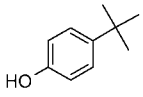
30



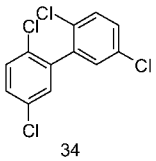
31



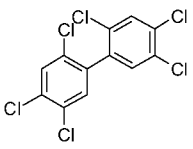
32



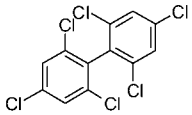
33



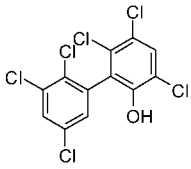
34



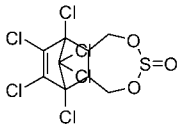
35



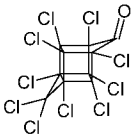
36



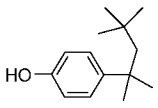
37



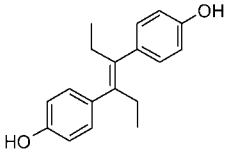
38



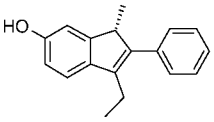
39



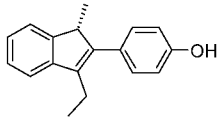
40



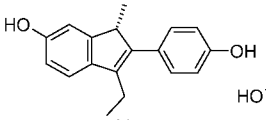
41



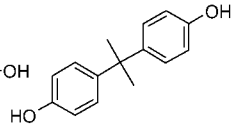
42



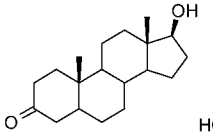
43



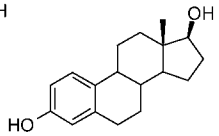
44



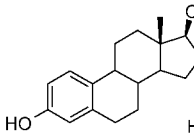
45



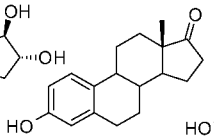
46



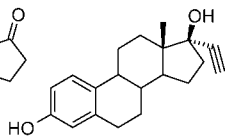
47



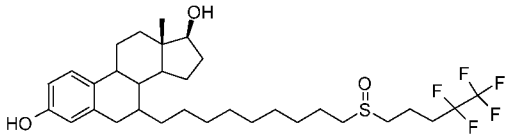
48



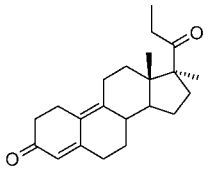
49



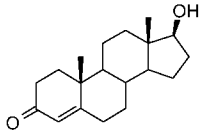
50



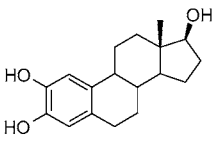
51



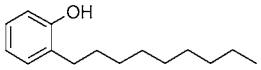
52



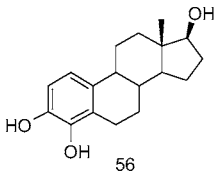
53



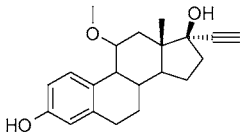
54



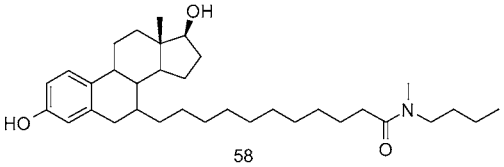
55



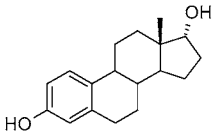
56



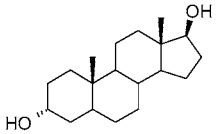
57



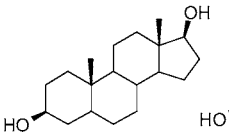
58



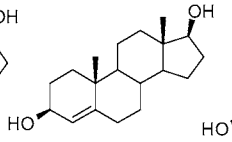
59



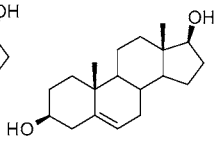
60



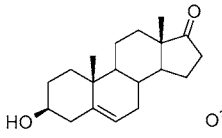
61



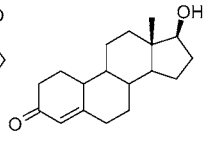
62



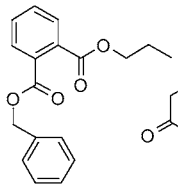
63



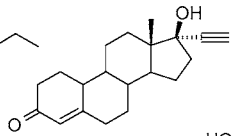
64



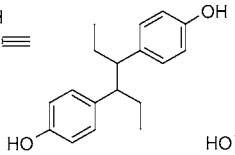
65



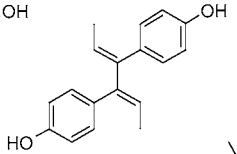
66



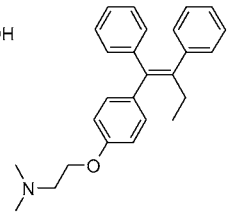
67



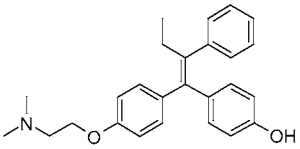
68



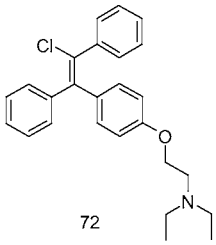
69



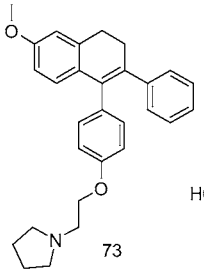
70



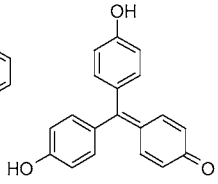
71



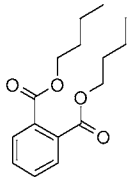
72



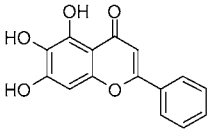
73



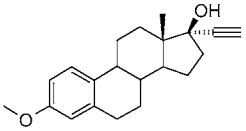
74



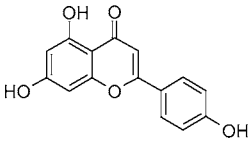
75



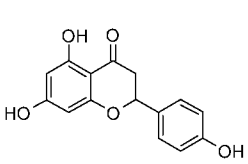
76

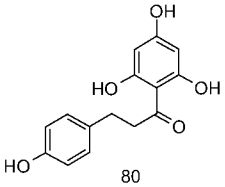


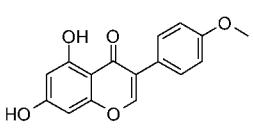
77

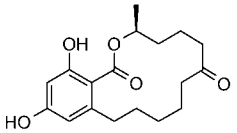


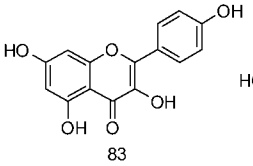
78

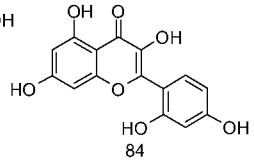

79

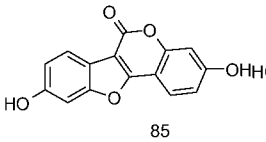

80

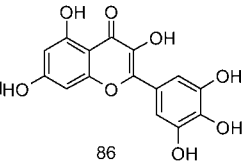

81

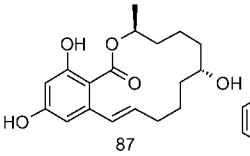

82

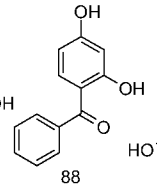

83

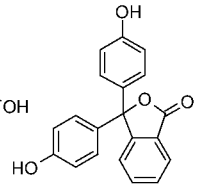

84

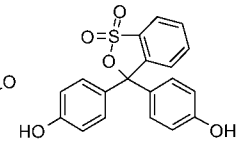

85

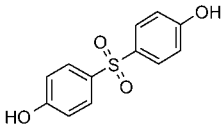

86

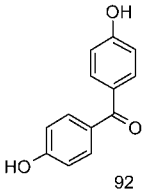

87

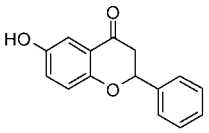

88

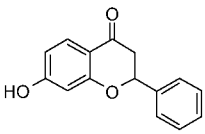

89

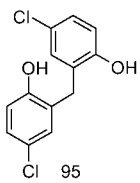

90

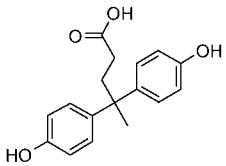

91

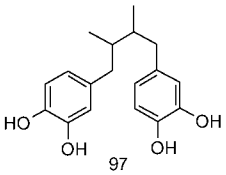

92

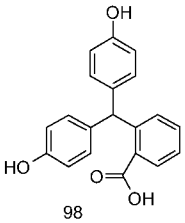

93

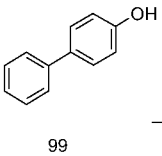

94

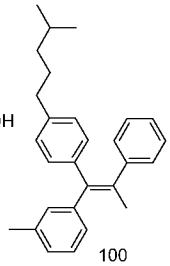

95

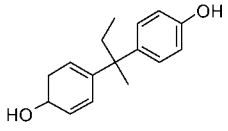

96

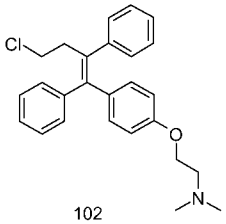

97

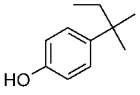

98

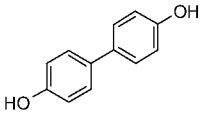

99

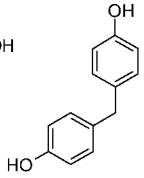

100

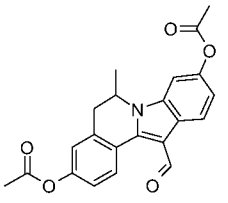

101

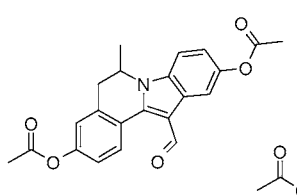

102

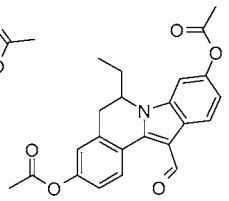

103

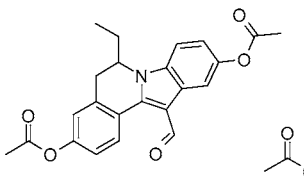

104

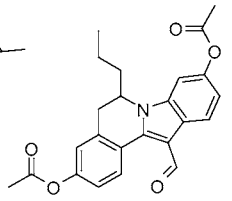

105

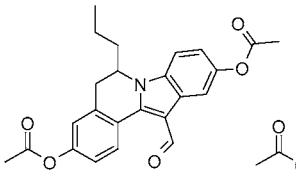

106

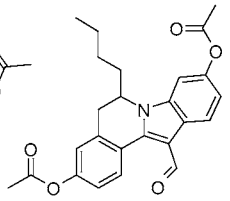

107

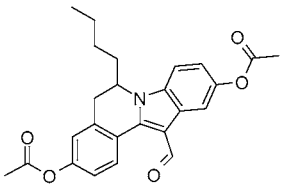

108

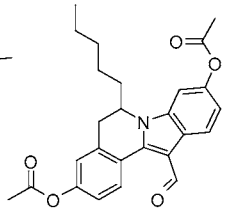

109

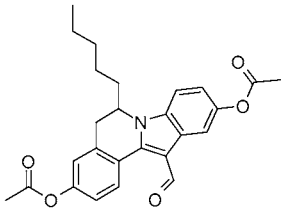

110

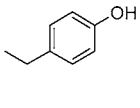

111

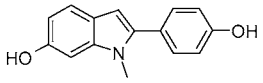

112

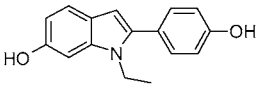

113

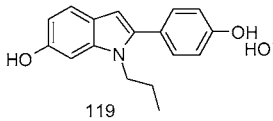

114

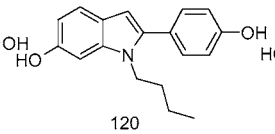

115

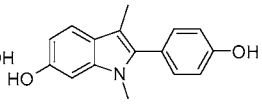

116

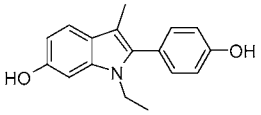

117

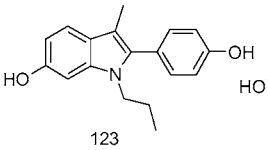

118

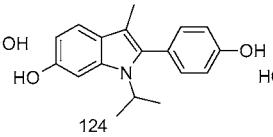

119

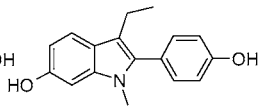

120

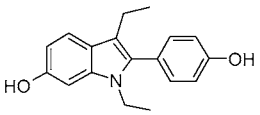

121

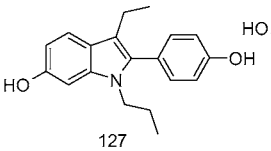

122

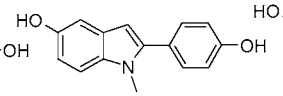

123

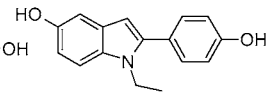

124

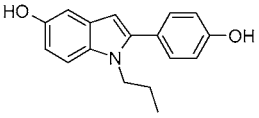

125

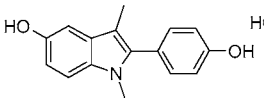

126

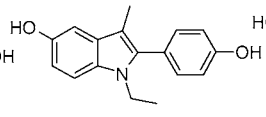

127

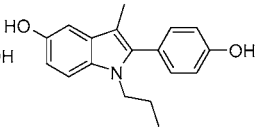

128

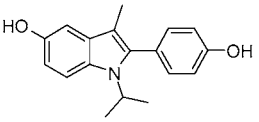

129

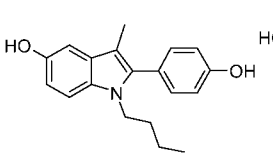

130


131

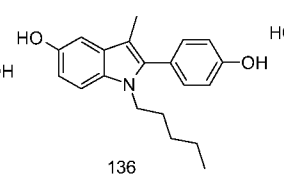

132


133

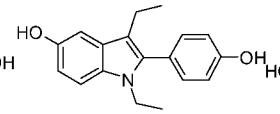

134



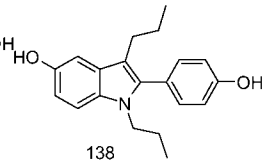
135



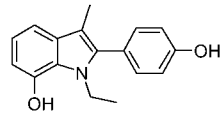
136



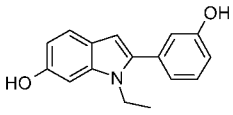
137



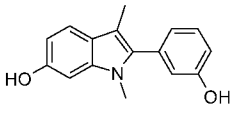
138



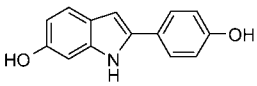
139



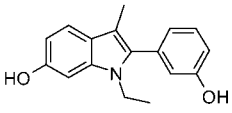
140



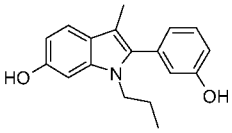
141



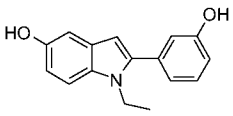
142



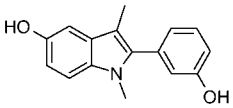
143



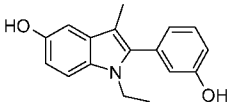
144



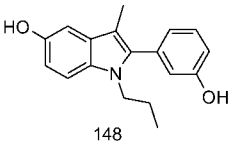
145



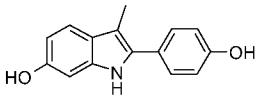
146



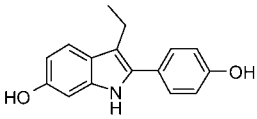
147



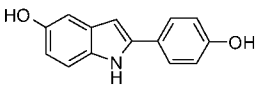
148



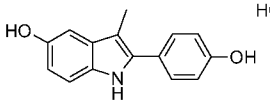
149



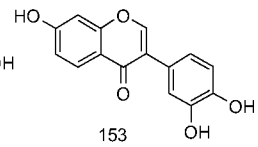
150



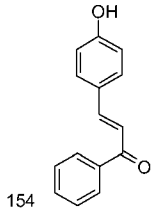
151



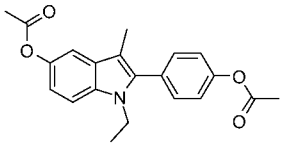
152



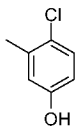
153



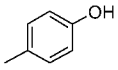
154



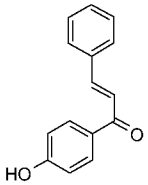
155



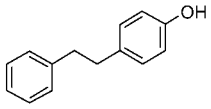
156



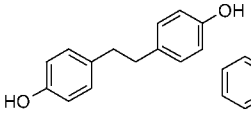
157



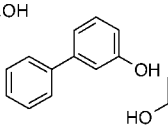
158



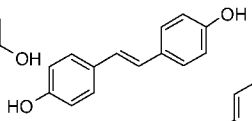
159



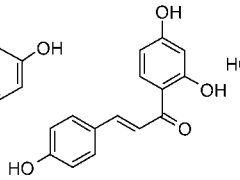
160



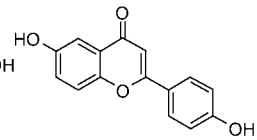
161



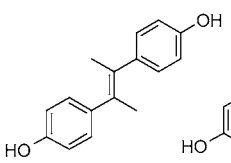
162



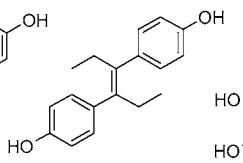
163



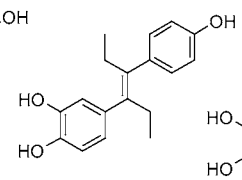
164



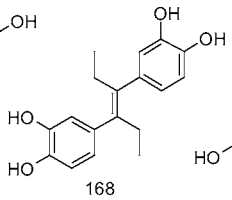
165



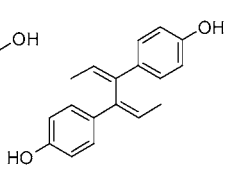
166



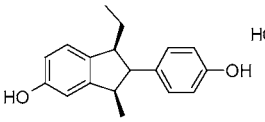
167



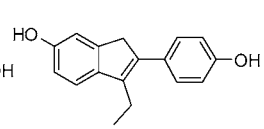
168



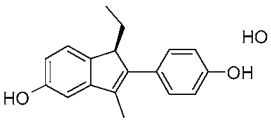
169



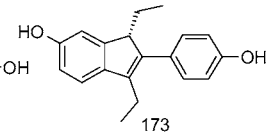
170



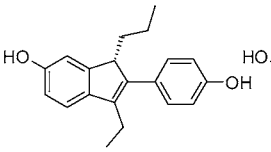
171



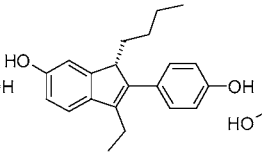
172



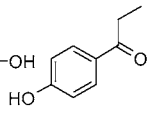
173



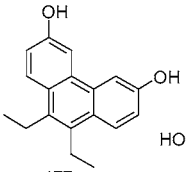
174



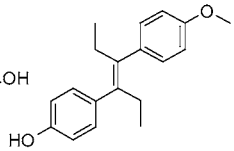
175



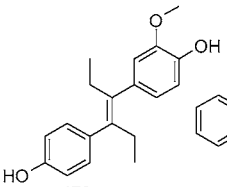
176



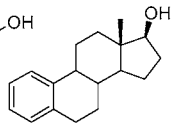
177



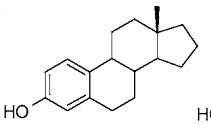
178



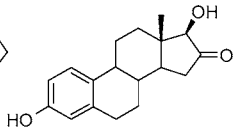
179



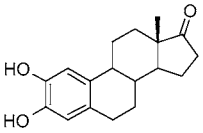
180



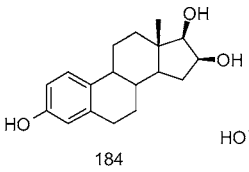
181



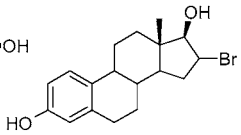
182



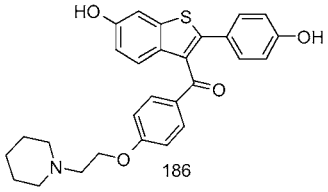
183



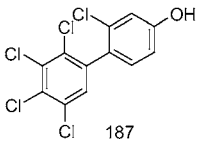
184



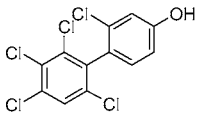
185



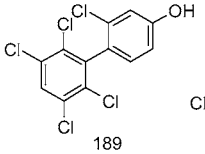
186



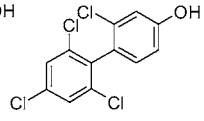
187



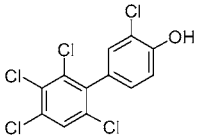
188



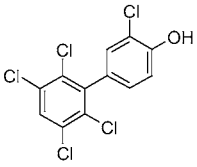
189



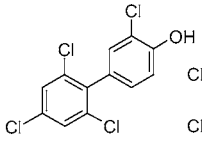
190



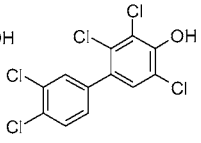
191



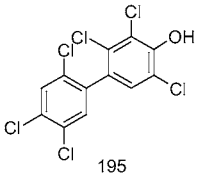
192



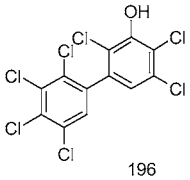
193



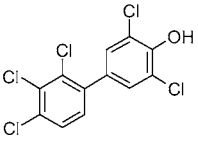
194



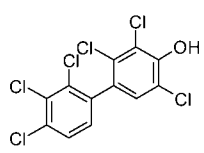
195

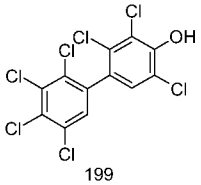


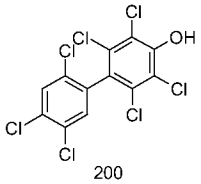
196

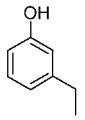


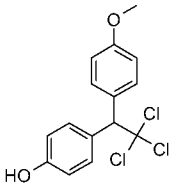
197

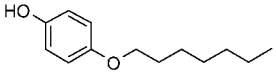

198

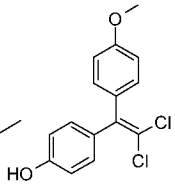

199

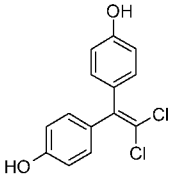

200

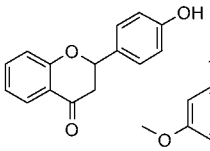

201

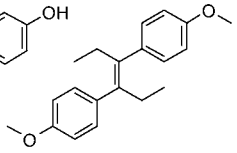

202

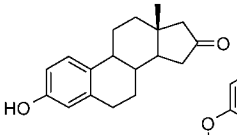

203

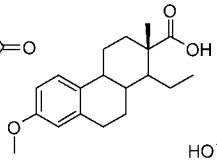

204

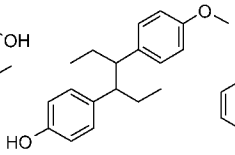

205

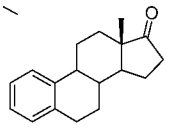

206

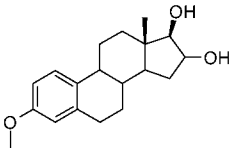

207

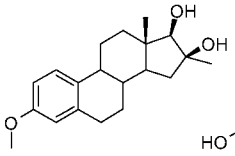

208

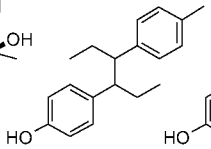

209

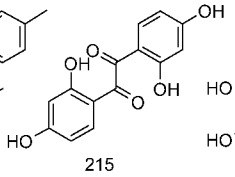

210

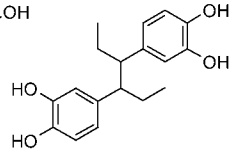

211

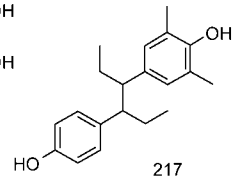

212

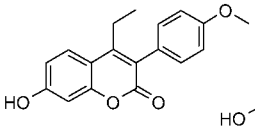

213

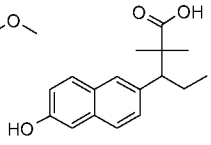

214

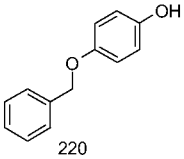

215

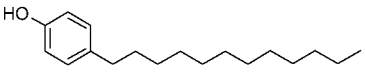

216

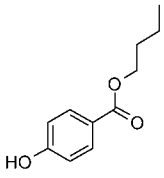

217

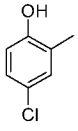

218

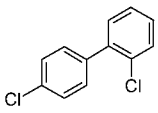

219

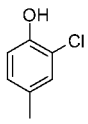

220

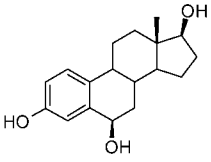

221

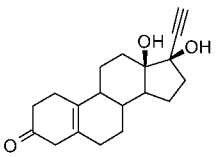

222

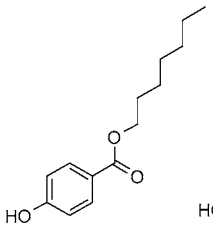

223


224

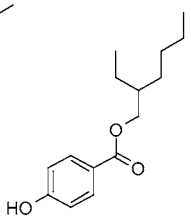

225


226

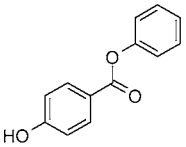

227



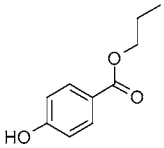
228



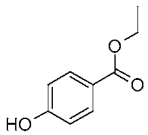
229



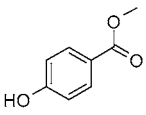
230



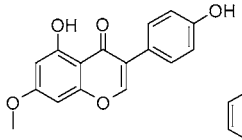
231



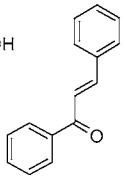
232



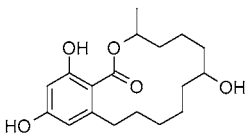
233



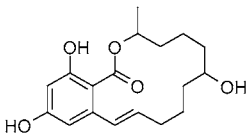
234



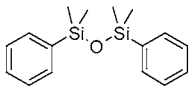
235



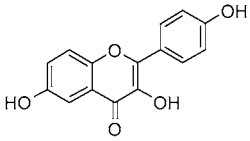
236



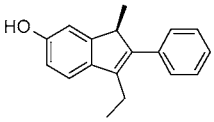
237



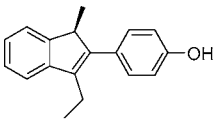
238



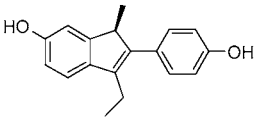
239



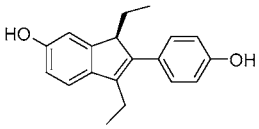
240



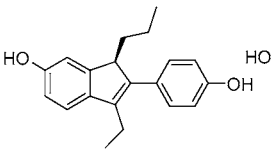
241



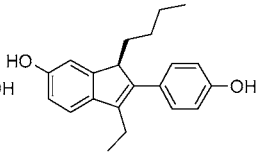
242



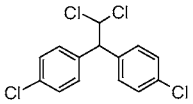
243



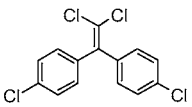
244



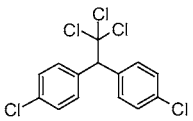
245



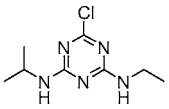
246



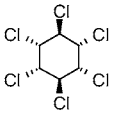
247



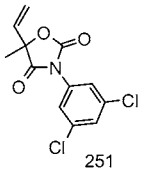
248



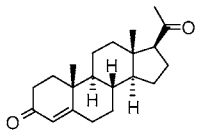
249



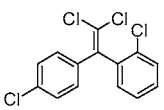
250



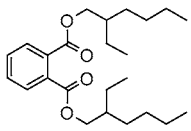
251



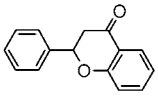
252



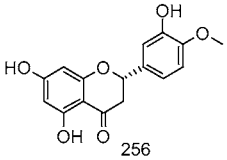
253



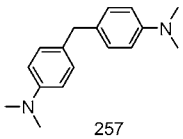
254



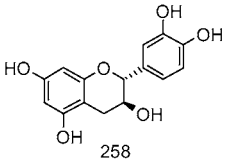
255



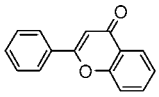
256



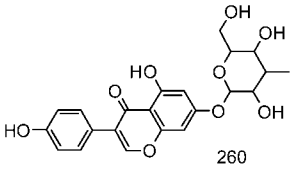
257

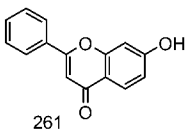


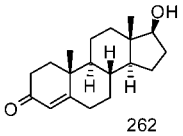
258

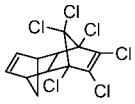


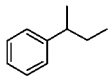
259

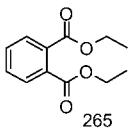

260

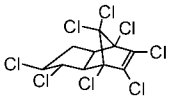

261

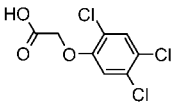

262

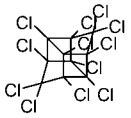

263

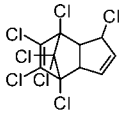

264

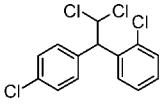

265

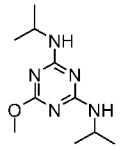

266

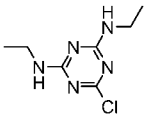

267

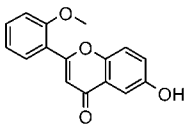

268

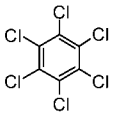

269

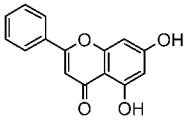

270

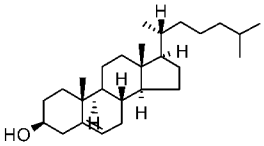

271

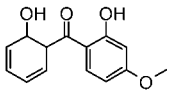

272

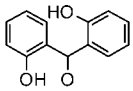

273

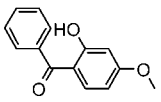

274

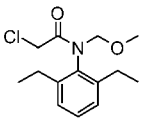

275

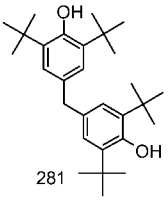

276

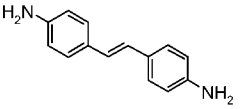

277

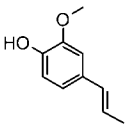

278

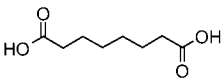

279

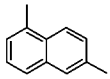

280

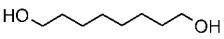

281

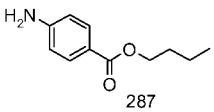

282

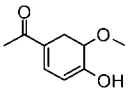

283

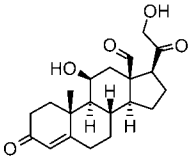

284

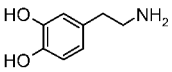

285

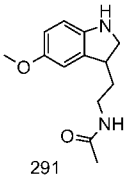

286

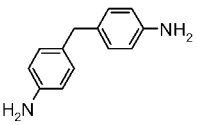

287


288

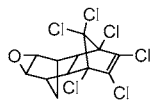

289

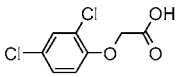

290

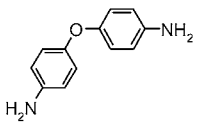

291

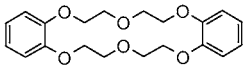

292

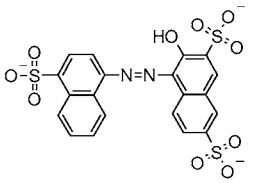

293

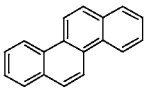

294

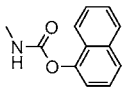

295

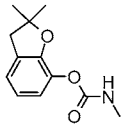

296

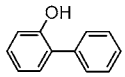

297

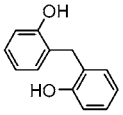

298

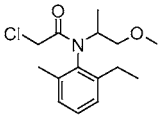

299


300

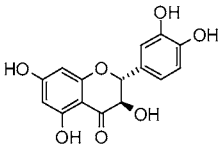

301

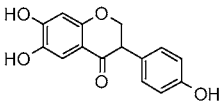

302

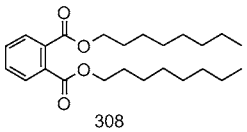

303

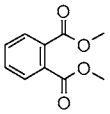

304

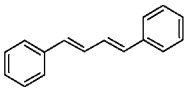

305

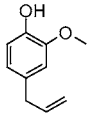

306

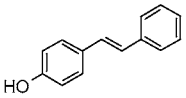

307

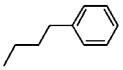

308

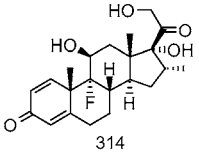

309

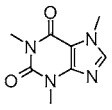

310

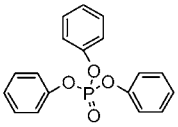

311

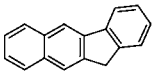

312

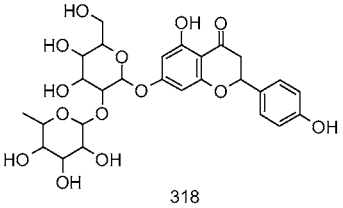

313


314

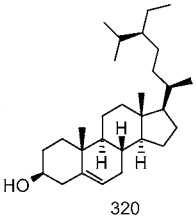

315

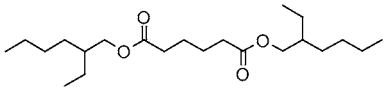

316

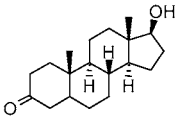

317

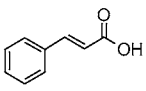

318


319

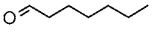

320

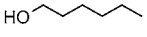

321

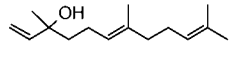

322


323

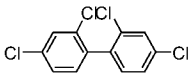

324


325

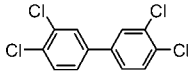

326



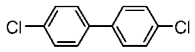
327



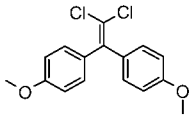
328



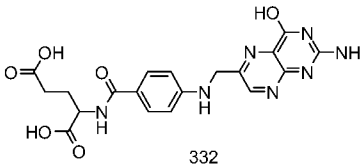
329



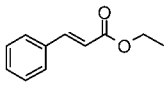
330



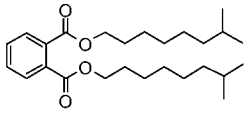
331



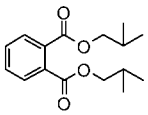
332



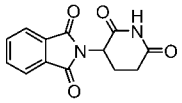
333



334



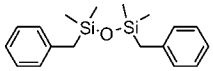
335



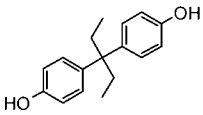
336

337

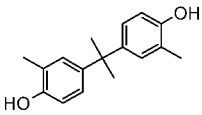
338



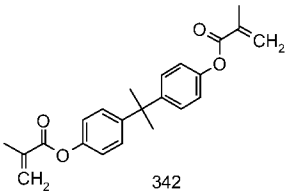
339



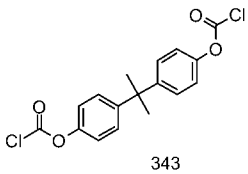
340



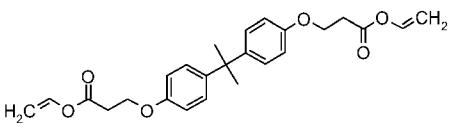
341



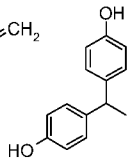
342



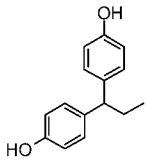
343



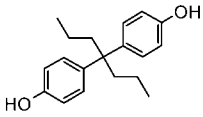
344



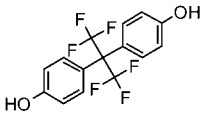
345



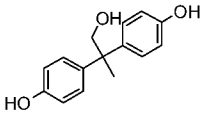
346



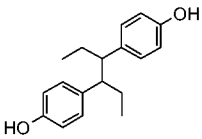
347



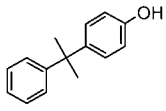
348



349



350



351

Table A2. CAS numbers and *log*RBA values of the compounds used for the construction of the kNN (rat data from compounds 1-245 and compounds 340-351) (**II**) and consensus kNN (compounds 1-245) (**III**) models. Active compounds used in the classification studies (**IV**) are compounds 1-245 and the inactive compounds are numbered 246-339 (N.D. = activity below detection limit).

Number	CAS	Calf	Human α	Human β	Mouse	Rat
1	88-18-6				-2.95	
2	446-72-0		0.60	1.94	-0.17	-0.36
3	531-95-3					-0.82
4	486-66-8		-1.00	-0.30		-1.65
5	485-72-3					-2.98
6	1806-26-4		-1.70	-1.15		-2.31
7	55331-29-8					1.48
8	17924-92-4		0.85	0.70	1.64	
9	117-39-5		-2.00	-1.40		
10	528-48-3					-2.35
11	58-72-0					-2.78
12	2971-36-0				0.72	-0.60
13	153-18-4					-4.09
14	6665-83-4					-3.41
15	89-72-5					-3.54
16	99-71-8					-3.37
17	72-43-5				-2.42	
18						-2.78
19	789-02-6		-2.00	-1.70	-1.04	-2.85
20			0.38	0.67	0.74	
21			0.53	0.86	0.76	-0.64
22	585-34-2				-3.18	
23					0.36	
24					-0.06	
25	53905-28-5				-0.13	-1.44
26	13049-13-3				-0.87	-3.25
27	92-04-6				-1.40	-2.77
28	28034-99-3				-1.59	-2.18
29					-1.69	
30					-1.43	
31					-1.31	
32					-0.78	
33	98-54-4				-2.79	-3.61
34	35693-99-3				-1.37	
35	35065-27-1				-1.51	
36	33979-03-2				-0.70	

37					-1.39	
38	115-29-7				-3.36	
39	143-50-0		-1.22	-1.00	-0.73	-1.89
40	140-66-9		-2.00	-1.52	-0.70	-1.82
41	56-53-1		2.37	2.34	2.57	2.60
42					-0.28	
43					0.38	
44					1.78	
45	80-05-7		-2.00	-2.00	-0.74	-2.11
46	521-18-6		-1.30	-0.77	-1.58	-2.89
47	50-28-2	2.00	2.00	2.00	2.00	2.00
48	50-27-1		1.15	1.32	1.27	0.90
49	53-16-7		1.78	1.57	1.78	0.86
50	57-63-6				2.94	2.28
51	129453-61-8	0.79			2.64	1.57
52	34184-77-5				-0.65	
53	58-22-0				-2.04	
54	362-05-0		0.85	1.04		1.47
55	25154-52-3		-1.30	-1.05	-0.50	-1.53
56	5976-61-4		1.11	0.85		1.82
57	34816-55-2		1.63	0.70		1.14
58		0.75	1.93	2.22		1.16
59	57-91-0		0.85	0.30		0.49
60	1852-53-5		-1.15	-0.52		-2.67
61	571-20-0		0.48	0.85		-0.92
62	1156-92-9		-0.30	-0.22		
63	521-17-5		1.23	0.00		
64	54-43-0		-1.40	-1.15		
65	434-22-0		-2.00	-0.64		
66	85-68-7				-2.46	
67	68-22-4		-1.15	-2.00		
68	84-16-2	1.34	2.48	2.37		2.48
69	84-17-3		2.35	2.61	1.30	1.57
70	10540-29-1	-0.44	0.60	0.48		0.21
71	68047-06-3	0.83	2.41	2.37		2.24
72	911-45-5		1.40	1.08		-0.14
73	1845-11-0		1.64	1.20		-0.14
74	603-45-2					-1.50
75	84-74-2				-2.58	
76	491-67-8					-3.05
77	72-33-3					0.35
78	520-36-5		-0.52	0.78		-1.55
79	480-41-1		-2.00	-0.96		-2.13

80	60-82-2		-0.70	-0.15		-1.16
81	491-80-5					-2.37
82	5975-78-0					0.32
83	520-18-3		-1.00	0.48		-1.61
84	480-16-0					-3.09
85	479-13-0		1.30	2.15	0.45	-0.05
86	529-44-2					-2.75
87	71030-11-0					1.63
88	131-56-6					-2.61
89	77-09-8					-1.87
90	143-74-8					-3.25
91	80-09-1					-3.07
92	611-99-4					-2.46
93	4250-77-5					-3.05
94	6515-36-2					-3.73
95	97-23-4					-2.45
96	126-00-1					-3.13
97	500-38-9					-1.51
98	81-90-3					-3.67
99	92-69-3				-2.00	-3.04
100	82413-20-5					1.18
101	77-40-7					-1.07
102	89778-26-7					0.14
103	80-46-6					-3.26
104	92-88-6			-1.52	-2.00	
105	620-92-8					-3.02
106		0.23				
107		0.59				
108		0.11				
109		0.52				
110		-0.30				
111		0.41				
112		-0.15				
113		0.32				
114		-0.22				
115		0.28				
116	123-07-9					-4.17
117		0.58				
118		1.20				
119		0.93				
120		0.63				
121		1.00				
122		1.52				

123		1.11				
124		1.11				
125		0.77				
126		1.32				
127		1.28				
128		-0.10				
129		0.76				
130		1.26				
131		0.66				
132		0.98				
133		1.20				
134		0.54				
135		0.66				
136		0.36				
137		1.36				
138		0.23				
139		-1.70				
140		0.23				
141		-0.26				
142		-2.00				
143		0.48				
144		0.54				
145		0.23				
146		-0.22				
147		0.34				
148		0.87				
149		-1.22				
150		-0.89				
151		-2.00				
152		-1.22				
153	485-63-2					-2.35
154	20426-12-4					-2.55
155		1.32				
156	59-50-7					-3.38
157	106-44-5					-4.50
158	2657-25-2					-2.43
159	6335-83-7					-2.69
160	6052-84-2					-1.44
161	580-51-8					-3.44
162	659-22-3					-0.55
163	961-29-5					-1.26
164	63046-09-3					-0.82
165	552-80-7				1.52	1.16

166					-0.10	
167					2.00	
168					1.40	
169					-0.52	
170					0.30	
171					1.15	
172					2.00	
173					2.47	
174					2.36	
175					2.25	
176	70-70-2				-1.00	
177					-0.80	
178					1.30	1.31
179					1.00	
180	2529-64-8					-0.30
181	53-63-4					1.14
182	566-75-6		0.11	-0.05		
183	362-06-1		0.30	-0.70		
184	1228-72-4		1.46	1.90		
185			1.88	1.00		
186	84449-90-1		1.84	1.20		
187			-1.00	-0.89		
188			-0.23	-0.70		
189			-1.05	-1.52		
190			-0.52	-0.30		
191			-0.89	-0.92		
192			-1.22	-1.40		
193			-0.74	-0.64		
194			-1.52	-1.70		
195			-1.52	-1.40		
196			-1.05	-1.00		
197			-2.00			
198			-1.15	-1.22		
199			-1.00	-1.00		
200			-1.00	-1.00		
201	620-17-7					-3.87
202	28463-03-8					-0.89
203	13037-86-0					-2.88
204	75938-34-0					-0.63
205	14868-03-2					0.42
206	6515-37-3					-2.65
207	7773-34-3					-1.25
208	3601-97-6					-0.29

209	15372-34-6					-2.74
210	13026-26-1					0.97
211	53-45-2					-2.20
212	3434-79-5					-1.65
213	5108-94-1					-1.48
214						0.60
215	5394-98-9					-0.68
216	79199-51-2					1.19
217						0.95
218	5219-17-0					-0.05
219	65118-81-2					-0.02
220	103-16-2					-3.44
221	104-43-8					-1.73
222	94-26-8					-3.07
223	1570-64-5					-3.67
224	34883-43-7					-3.61
225	6640-27-3					-3.66
226	1229-24-9					-0.15
227	68-23-5		-0.16	-0.66		-0.67
228	1085-12-7					-2.09
229	5153-25-3					-1.74
230	94-18-8					-2.54
231	94-13-3					-3.22
232	120-47-8					-3.22
233	99-76-3					-3.44
234	552-59-0					-2.74
235	94-41-7					-2.82
236	42422-68-4		1.20	1.15		-0.19
237						-0.69
238	56-33-7					-3.16
239						-0.35
240					0.68	
241					1.17	
242					2.88	
243					1.04	
244					1.26	
245					0.90	
246	3424-82-6					N.D.
247	117-81-7					N.D.
248	487-26-3					N.D.
249	520-33-2					N.D.
250	101-61-1					N.D.
251	154-23-4					N.D.

252	525-82-6					N.D.
253	529-59-9					N.D.
254	6665-86-7					N.D.
255	481-30-1					N.D.
256	309-00-2					N.D.
257	135-98-8					N.D.
258	84-66-2					N.D.
259	57-74-9					N.D.
260	93-76-5					N.D.
261	2385-85-5					N.D.
262	76-44-8					N.D.
263	53-19-0					N.D.
264	1610-18-0					N.D.
265	122-34-9					N.D.
266						N.D.
267	118-74-1					N.D.
268	72-54-8					N.D.
269	72-55-9					N.D.
270	50-29-3					N.D.
271	1912-24-9					N.D.
272	58-89-9					N.D.
273	50471-44-8					N.D.
274	57-83-0					N.D.
275	480-40-0					N.D.
276	57-88-5					N.D.
277	131-53-3					N.D.
278	835-11-0					N.D.
279	131-57-7					N.D.
280	15972-60-8					N.D.
281	118-82-1					N.D.
282	54760-75-7					N.D.
283	97-54-1					N.D.
284	505-48-6					N.D.
285	575-43-9					N.D.
286	629-41-4					N.D.
287	94-25-7					N.D.
288	121-33-5					N.D.
289	52-39-1					N.D.
290	51-61-6					N.D.
291	73-31-4					N.D.
292	101-77-9					N.D.
293	95-57-8					N.D.
294	60-57-1					N.D.

295	94-75-7					N.D.
296	101-80-4					N.D.
297	14187-32-7					N.D.
298	915-67-3					N.D.
299	218-01-9					N.D.
300	63-25-2					N.D.
301	1563-66-2					N.D.
302	90-43-7					N.D.
303						N.D.
304	51218-45-2					N.D.
305	108-95-2					N.D.
306	480-18-2					N.D.
307	17817-31-1					N.D.
308	117-84-0					N.D.
309	131-11-3					N.D.
310	886-65-7					N.D.
311	97-53-0					N.D.
312	6554-98-9					N.D.
313	104-51-8					N.D.
314	50-02-2					N.D.
315	58-08-2					N.D.
316	115-86-6					N.D.
317	243-17-4					N.D.
318	10236-47-2					N.D.
319	90-00-6					N.D.
320	83-46-5					N.D.
321	103-23-1					N.D.
322	571-22-2					N.D.
323	621-82-9					N.D.
324	100-51-6					N.D.
325	111-71-7					N.D.
326	111-27-3					N.D.
327	7212-44-4					N.D.
328	2437-79-8					N.D.
329	32598-13-3					N.D.
330	2050-68-2					N.D.
331	2132-70-9					N.D.
332	59-30-3					N.D.
333	103-36-6					N.D.
334	28553-12-0					N.D.
335	84-69-5					N.D.
336	50-35-1					N.D.
337	98-01-1					N.D.

338	470-82-6					N.D.
339	1833-27-8					N.D.
340						-0.74
341	79-97-0					-0.60
342	3253-39-2					-2.82
343	2024-88-6					-1.64
344						-3.30
345						-3.05
346						-0.82
347						-0.82
348	1478-61-1					0.00
349						-2.12
350	5776-72-7					0.56
351	599-64-4					-2.30

Kuopio University Publications C. Natural and Environmental Sciences

C 173. Tarvainen, Mika. Estimation Methods for Nonstationary Biosignals.
2004. 138 p. Acad. Diss.

C 174. Yppärilä, Heidi. Depth of sedation in intensive care patients. A neuropsysiological study.
2004. 104 p. Acad. Diss.

C 175. Sohlberg, Antti. Molecular small animal imaging with pinhole single-photon emission computed tomography.
2004. 104 p. Acad. Diss.

C 176. Kumlin, Timo. Studies on cancer-related effects of 50 Hz magnetic fields.
2004. 64 p. Acad. Diss.

C 177. Seppänen, Kari. Does mercury promote lipid peroxidation?: in vivo and in vitro studies concerning mercury and selenium in lipid peroxidation and coronary heart disease.
2004. 49 p. Acad. Diss.

C 178. Riikonen, Johanna. Modification of the growth, photosynthesis and leaf structure of silver birch by elevated CO2 and O3.
2004. 126 p. Acad. Diss.

C 179. Frank, Christian. Functional profiling of the xenobiotic nuclear receptors CAR and PXR.
2004. 87 p. Acad. Diss.

C 180. Rytkönen, Esko. High-frequency vibration and noise in dentistry.
2005. 80 p. Acad. Diss.

C 181. Seppänen, Aku. State estimation in process tomography.
2005. 117 p. Acad. Diss.

C 182. Ibrahim, Mohamed Ahmed. Plant essential oils as plant protectants and growth activators.
2005. 143 p. Acad. Diss.

C 183. Vuorinen, Terhi. Induced volatile emissions of plants under elevated carbon dioxide and ozone concentrations, and impacts on indirect antiherbivore defence.
2005. 98 p. Acad. Diss.

C 184. Savinainen, Juha. Optimized methods to determine ligand activities at the cannabinoid CB1 and CB2 receptors.
2005. 83 p. Acad. Diss.

C 185. Luomala, Eeva-Maria. Photosynthesis, chemical composition and anatomy of Scots pine and Norway spruce needles under elevated atmospheric CO2 concentration and temperature.
2005. 137 p. Acad. Diss.

C 186. Heikkinen, Lasse M. Statistical estimation methods for electrical pricess tomography.
2005. 147 p. Acad. Diss.

C 187. Riihinen, Kaisu. Phenolic compounds in berries.
2005. 97 p. Acad. Diss.