REPORT

Modelling endocrine disruption

Quantitative structure-activity relationships for three estrogenic end-points

Erik Furusjö, Ann-Sofie Allard, Sara Nilsson, Magnus Rahmberg, Anders Svenson

B1759

October 2007

This report approved 2007-12-14

Lars-Gunnar Lindfors
Scientific Director

IVL Swedish Environmental Research Institute

IVL Swedish Environmental Research Institute

Organization

IVL Swedish Environmental Research Institute Ltd.

Address

P.O. Box 21060

SE-100 31 Stockholm

Telephone

+46 (0)8-598 563 00

Report Summary

Project title

QSAR-verktyg för prognosticering av kemikaliers egenskaper

Project sponsor

CFs Miljöfond, SIVL

Author

Erik Furusjö, Ann-Sofie Allard, Sara Nilsson, Magnus Rahmberg, Anders Svenson

Title and subtitle of the report

Modelling endocrine disruption. Quantitative structure-activity relationships for three estrogenic end-points

Summary Hormone regulated processes govern the larval development and reproduction in aquatic vertebrates. The involvement of pollutants in these processes needs to be examined in the evaluation of consequences of their release into the aquatic environment. Prognosis models may be used as a pre-screening step to predict such properties. This work describes the development of models to predict the properties involved in estrogenic hormone activity, *i.e.* induction (in absolute and relative figures) and binding affinity to the human estrogen receptor α. Models were developed using molecular hologram descriptors, *i.e.* derived only from the chemical structure of substances. Statistical molecular design (SMD) was used to select substances for experimental testing of receptor induction. Substances were then divided in a set for model construction and a validation set. Models were calculated for a large number of systematically varied hologram configurations to find those with the best data based on the predictive ability (Q2-value). The models with the best predictive ability were those based on holograms that considered chirality in the chemical structure. No other systematic effect of hologram length and fragment length was observed. The best models for each biological response were then further refined by a pruning procedure that resulted in exclusion of descriptors that did not contribute positively to the model. The three estrogenic responses were predicted within a factor 10 (root mean square error of prediction 0.7 – 1.0), which should be sufficient for the pre-screening purpose. A broad applicability characterises the three models, only the structural formula of a new substance is needed to be able to perform the prediction.

Keyword

Hormone, QSAR, HQSAR, molecular hologram, PLS, multivariate, PModX, SMD, prediction

Bibliographic data

IVL Report B1759

The report can be ordered via

Homepage: www.ivl.se, e-mail: publicationservice@ivl.se, fax+46 (0)8-598 563 90, or via IVL, P.O. Box 21060, SE-100 31 Stockholm Sweden

Summary

Hormone regulated processes govern the larval development and reproduction in aquatic vertebrates. The involvement of pollutants in these processes needs to be examined in the evaluation of consequences of their release into the aquatic environment. Prognosis models may be used as a pre-screening step to predict such properties. This work describes the development of models to predict the properties involved in estrogenic hormone activity, *i.e.* induction (in absolute and relative figures) and binding affinity to the human estrogen receptor α. Models were developed using molecular hologram descriptors, *i.e.* derived only from the chemical structure of substances. Statistical molecular design (SMD) was used to select substances for experimental testing of receptor induction. Substances were then divided in a set for model construction and a validation set. Models were calculated for a large number of systematically varied hologram configurations to find those with the best data based on the predictive ability (Q2-value). The models with the best predictive ability were those based on holograms that considered chirality in the chemical structure. No other systematic effect of hologram length and fragment length was observed. The best models for each biological response were then further refined by a pruning procedure that resulted in exclusion of descriptors that did not contribute positively to the model. The three estrogenic responses were predicted within a factor 10 (root mean square error of prediction 0.7 – 1.0), which should be sufficient for the pre-screening purpose. A broad applicability characterises the three models, only the structural formula of a new substance is needed to be able to perform the prediction.

Sammanfattning

Hormonstyrda processer styr individers utveckling hos vattenlevande organismer och deras fortplantning. Påverkan av främmande ämnen på dessa processer är egenskaper som behöver kartläggas inför bedömningen av eventuella utsläpp till vattenmiljön och överhuvudtaget i samband med utvärdering av miljöegenskaper. I ett inledande skede kan modeller att beräkna sådana egenskaper utnyttjas. I detta arbete har multivariata modeller utvecklats för tre egenskaper som alla är knutna till östrogena effekter, induktion (såväl absoluta som relativa värden) och bindningsaffinitet med östrogenreceptor α hos människa. Modellerna bygger enbart på molekylära hologramdeskriptorer härledda från ämnens kemiska struktur. Modeller beräknades för ett stort antal olika hologramkonfigurationer för var och en av de tre biologiska egenskaperna för att söka de konfigurationer som gav bästa resultat. Urvalet baserades på modellernas prediktionsförmåga (Q2-värdet). Högsta värden erhöll modeller som tog hänsyn till kiralitet i ämnenas molekyler. Hologrammens eller fragmentens längd påverkade inte prediktionsförmågan. Med dessa modeller är det möjligt att beräkna dessa tre östrogena effekter med en genomsnittlig träffsäkerhet inom en faktor 10. Modellen för bindningsaffinitet gav något bättre prognoser än de andra två. Gemensamt för de tre modellerna är deras breda tillämpbarhet. Endast ett ämnes strukturformel behövs för att prognosticera dessa effekter.

Table of contents

1 Introduction - QSARs for endocrine disruption

During the last two decades there has been an increasing concern about substances interfering with the normal endocrine functions in both animals and humans. Most focus has been directed to the effects of environmental pollutants on sex steroid hormone regulated processes. These substances affect the reproductive function and development in vertebrates. The effects of environmental disrupters on endocrine functions are a prioritised area by the European Union.

Quantitative structure-activity relationships (QSAR) are modelling tools often used to prescreen properties to facilitate decision processes prior to experimental studies. Several attempts have been reported about the use of QSAR modelling of endocrine disrupters. Good and reliable QSAR models for prognosis of other chemical or biological properties, such as acute aquatic toxicity, have been developed. An US Environmental Pollution Agency (EPA) workshop [EPA 2000] raised some concerns related to QSAR for endocrine disrupter priority-setting. Much of the modelling work was based on *in vitro* studies, and since reliable receptor binding data is critical, it was questioned whether EPA had enough data to base QSARs on. EPA responded that the US Food and Drug Administration (FDA) had completed data collection for 220 compounds. Their goal was to expand the data set to include 500 compounds. EPA anticipated that this baseline data will be robust. In addition, these QSARs and high-throughput pre-screening (HTPS) methods focused on ligand receptor binding only, and did not capture other types of agonist and antagonistic activities.

In a preliminary study we attempted to model estrogen receptor binding affinity (RBA) from traditional molecular descriptors calculated by the Dragon software. The results showed that the models could distinguish between groups with low and high RBA. Within the groups, the predictive ability was poor, despite several attempts with variable selection, clustering of substances and non-linear extensions of partial least squares (PLS). We concluded that this set of descriptors, *i.e.* descriptors that can be calculated by the Dragon software, were not appropriate for quantitative prediction models for the human estrogen receptor (ER-α) RBA. The results agreed with previously published comparative studies of ER-α RBA modelling. Tong *et al.* [1998] showed that traditional QSAR descriptors are outperformed by holographic QSAR (HQSAR) and Comparative Molecular Field Analysis (CoMFA). Also Shi *et al.* [2001] showed the potential of HQSAR and CoMFA for endocrine disrupter modelling. However, it is worth noting that PLS was usually the regression method used for both methods. This means that the outlier detection features of PLS can potentially be applied to HQSAR and CoMFA models for endocrine disruption, which significantly would increase the reliability of the models. Similar arguments can be applied to modelling of other measures of endocrine disruption, *e.g.* agonistic and antagonistic activities.

The FDA developed a four-phase approach for priority setting of endocrine disruptors using QSAR [Tong *et al.* 2002]. The approach was quite complex and utilised a large number of models in each phase to estimate if substances have significant binding affinity for the estrogen receptor. The later phases in the approach were based on CoMFA models, while the earlier ones were simpler rule-based methods. It was claimed that testing of about 90 % of substances may be eliminated by the approach and that the risk of false negatives was small.

Gao *et al.* [1999a] investigated QSAR modelling of groups of structurally similar compounds to calculate relative binding affinity for different estrogen receptors with strong focus on

interpretation of the models in physico-chemical terms to understand ligand-receptor interactions. The models were based on a few traditional molecular descriptors, such as the Taft parameter for substituents, molecular weight and molecular volume, and in general provided good fits for the small number of substances analysed in each case. The authors claimed that the interpretation was facilitated by the low number of descriptors and their physico-chemical characters and meant that more complex models, *e.g.* CoMFA models, could not be interpreted in the same terms, partly due to the use of principal components.

Suzuki *et al.* [2001] constructed models for predictions of ER-α binding affinity using 60 substances and one-, two-, and three-dimensional descriptors, the latter including Weighted holistic invariant molecular (WHIM) descriptors and Theoretical linear solvation energy relationship (TLSER). Models based on 3D descriptors gave expectedly better predictions (cross validated predictive ability, Q2cv: 0.67-0.96) as compared to those based only on 1D and 2 D descriptors (Q2cv: 0.42-0.75). The values refer to arranging substances in subclusters, as clustering resulted in better predictions than models based on all 70 substances.

Models using PCA with descriptors and Bayensian classification of scores was used to classify substances as active or inactive [Gao *et al.* 1999b]. This method used only independent descriptors. 3D descriptors could not be used because the conformation of the active component was not always known. Only 13 descriptors were used involving connectivity indices and shape parameters, including indicator values showing certain structural elements. The models performed well, 85-90 % of the substances were classified correctly, fairly independent of where the limit between inactive and active substances was set.

Conventional QSARs based on c. 200 descriptors, half of which were quantum chemical descriptors, were compared to models using HQSAR and CoMFA [Tong *et al.* 2002]. These latter methods performed better than conventional models and HQSAR was easier to use, because only 2D descriptors were employed, *i.e.* conformation had not to be considered. QSAR models for ER-α RBA were constructed based on a diverse set of 130 substances using CoMFA and HQSAR [Shi *et al.* 2001]. Upon comparison of the two different methods, CoMFA was superior (Q2 (loo-cv)=0.65-0.71). Schmieder *et al.* [2000] used CoRePa with 29 alkylphenolic substances for modelling 3D structures of substances interfering in estrogen gene activation. They conclude that molecular conformation was especially important in flexible molecules.

In this study we report the development of holographic QSAR models based on published data and own experimental tests of the agonistic estrogenic effect on the human estrogen receptor α and published data on the relative binding affinity to the same receptor. To be able to perform this type of QSAR the following components are needed:
- Substances with known structure (and biological activity if the aim is to develop new models).
- Software for generating molecular holograms based on the structures of the substances.
- Software for multivariate data analysis.

2 Theory

2.1 Molecular descriptors

A QSAR model is a relation between chemical structure and a property of the chemical compound. The features of a chemical structure are captured by so called chemical descriptors that can be of a number of different types. In the following sections three different types of descriptors are presented that can be used for QSAR models for endocrine disruption.

2.1.1 Holographic QSAR – HQSAR

Holographic QSAR (HQSAR) uses molecular holograms as descriptors [Burden, Winkler 1999]. A molecular hologram is an array containing counts of molecular fragments. The required input data for the generation of the hologram is the two-dimensional chemical structure of the substance. The substance is then divided into smaller fragments of a predefined length. Each fragment is converted into a number by translating its representation to an SLN (Sybyl Line Notation) to an integer using a CRC (cyclic redundancy check) algorithm [Lowis, 1997]. Each of these integers corresponds to a bin in an integer array of fixed length L. The length of the array is normally between 50 and 500. Bin occupancies are incremented according to the fragments generated. Thus, all generated fragments are hashed into array bins in the range 1 to L, similar to a histogram. This array is called a molecular hologram, and the bin occupancies are the descriptor variables. An overview of the generation of molecular hologram is shown in Figure 1.

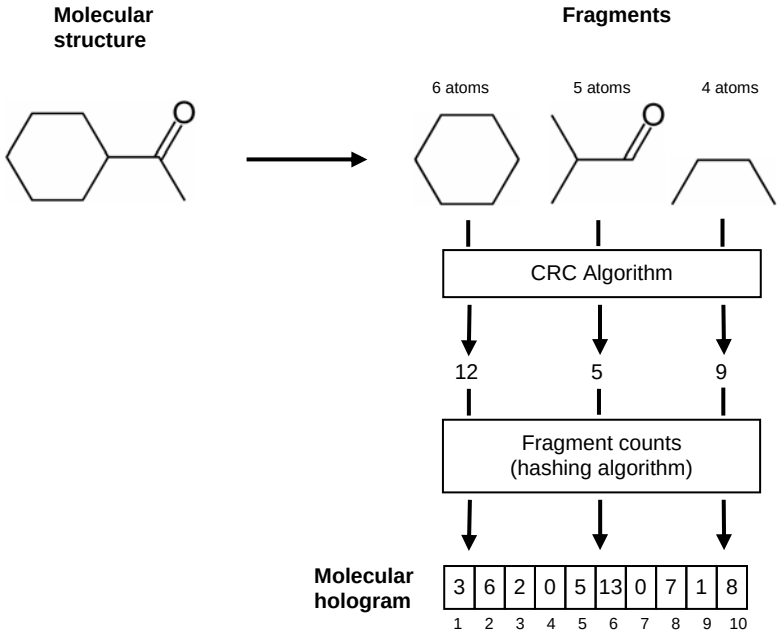


Figure 1. Schematic overview of molecular hologram generation. From Lowis [1997].

The outcome of the molecular hologram generation is dependent on a set of parameters [Lowis 1997]. These parameters are:

- **Hologram length**

- **Fragment length**
- **Fragment distinction**
  - **Atoms**, enables fragments to be resolved based on elemental atom types.
  - **Bonds**, enables fragments to be distinguished based on bond orders.
  - **Connectivity**, provides a measure of atomic hybridization states within fragments. That is, connectivity causes HQSAR to keep track of how many connections are made to constituent atoms, and the bond order of those connections.
  - **Hydrogens**, by default HQSAR ignores the hydrogen atoms during fragment generation; the hydrogen parameter overrides this behaviour
  - **Chirality**, enables fragments to be distinguished based on atomic and bond stereo chemistry.

The molecular hologram generated for the studied substances are related to the biological endpoint with multivariate projection methods, *i.e.* partial least squares (PLS) see section 2.2.2 and 2.2.3 for further information.

2.1.2 CoMFA and GRID

There exist specific QSAR descriptors that are based on a more physical model or understanding of the molecular interactions behind the measured biological response. Two methods that are closely related and based on superposition and alignment of molecular structures are Comparative molecular field analysis (CoMFA) and GRID [Livingstone 2000]. Both involve the use of a molecular probe and calculation of the interaction between the probe and the molecule that is being analysed. Interactions are measured at a (usually large) number of points in space defined by a grid placed around the molecular structure. PLS (see below) is usually used as the regression method in CoMFA.

CoMFA and GRID require that molecules be aligned relative to some common reference, *e.g.* the centre of mass. Aligning molecules with a similar structure is usually not that difficult, but a more diverse data set poses problems for all methods requiring alignment [Buydens *et al.* 1999]. CoMFA and GRID descriptors have not been used in the present work.

2.1.3 Common Reactivity Pattern

The modelling methods discussed above are general empirical regression methods that can in principle be applied to any regression problem and that can be used for QSAR modelling when applied to molecular descriptors and molecular properties of substances.

Another approach is Common Reactivity Pattern (CoRePa) [Mekenyan *et al.* 1997], which accounts for conformer flexibility in the structures. A brief description of CoRePa is as follows. A set of chemicals that are most (or sometimes least) active, *i.e.* that exceed (fall short of) a threshold for the biological activity in question, is selected. Then, a set of parameters that are hypothesised to be potentially important for the biological activity are identified. These are evaluated for a distribution of conformers for each compound to give a distribution of the parameter per substance. All distributions for a certain parameter are superimposed and common regions are identified. The

common regions identified (*i.e.* for different parameters) constitute the common reactivity pattern. CoRePa descriptors have not been used in the present work.

2.2 Modelling methods

Different modelling methods can be used to relate the structure of chemicals to environmental properties. Although simpler, univariate modelling have been used, the emphasis lately has been on multivariate regression methods, *e.g.* those based on latent variables, one of which, partial least squares (PLS), was used in this work. Some of these methods are described below.

2.2.1 Linear regression

The simplest QSAR models are univariate linear regression models of the form:

$$response = k_1 \times descriptor + k_0$$

These simple models are of limited use since such relationships are usually inadequate. An extension of the equation is:

$$response = k_0 + \sum_{i=1}^{p} k_i \times descriptor^i$$

where p is usually chosen as $p = 2$ or $p = 3$. The extension allows non-linear relations between the response and the single descriptor. However, a single descriptor is usually not sufficient to capture the properties of a substance, although successful applications have been reported for narrow groups of substances, usually with log K_{OW} as the descriptor.

Multiple linear regression (MLR) can be used to model the dependence of several descriptors according to the equation:

$$response = k_0 + \sum_{i=1}^{p} k_i \times x_i$$

where x_i is the ith descriptor. The number of descriptors, p, can vary widely from $p = 2$ to relatively large numbers. However, if many descriptors are used that contain similar information, *i.e.* are co-linear, problems with variance inflation occurs, which means that the models become very sensitive to small variations in the descriptors and that their predictive performance becomes poor. To solve this problem, different variable selection algorithms can be used to select a small set of variables with high information content. Another approach is to use multivariate projection methods, described in the next section, that handle, and even utilise, the co-linearity in the descriptor set.

2.2.2 Multivariate projection methods

Typical examples of multivariate projection methods are principal component analysis (PCA) and partial least squares (PLS). Sometimes this type of methods is denoted multivariate data analysis (MVA) methods, which is a rather non-descriptive name but nevertheless adopted here due to convention. More informative names are multivariate projection methods or latent variable methods.

The fundamental MVA method is PCA. A brief description of PCA is given here; more details are given in literature [Wold *et al.* 1987, Martens, Naes 1989, and Esbensen *et al.* 1996]. PCA decomposes a data matrix **X** (a table, in the current context the rows correspond to the substances while the columns correspond to descriptors) according to:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}$$

PCA can be considered a co-ordinate transformation from the original variable space to a model hyper-plane of much lower dimensionality that captures the variance in the data in the most efficient way. The scores, denoted **t** or **T**, are the co-ordinates in the new co-ordinate system and thus describe the objects (here: chemical substances). The loadings, denoted **p** or **P**, describe the relation between the latent variables (principal components) that span the model space and original variables.

The matrix **E** in the equation above contains the residuals, *i.e.* the part of the data not captured by the model hyper-plane. Substances that do not conform to the "pattern" found among the other substances will not be properly described by the model and thus have large residuals. This can be caused by corrupted data or that the substance in question is different from the others, which may indicate that a QSAR model based on the rest of the compounds will not be valid.

The substantial dimensionality reduction achieved by applying PCA to molecular descriptor data sets leads to enhanced interpretation abilities which facilitate classification and clustering of substances. This is utilised in a methodology known as statistical molecular design (SMD), *cf.* below.

PCA is not a regression method and cannot be used for finding quantitative relationships between descriptors and responses. The most common multivariate regression method is PLS.

2.2.3 Partial least squares

PLS is a latent variable based regression method described in several references [Martens, Naes 1989, Esbensen *et al.* 1996, Geladi, Kowalski 1986]. PLS has several benefits compared to ordinary multiple linear regression:

- Co-linearity is handled in a natural way and even utilised to find a robust estimate of the data structure. This means that variable selection methods are of less importance than in MLR.
- The latent variable approach means that outlier diagnostics can be obtained both for training and prediction substances.

The prediction outlier diagnostics obtained has no counterpart in MLR or the non-linear regression methods, such as artificial neural networks (ANN) discussed below, and are the greatest advantage of latent variable regression methods according to us. For a new sample it is possible to calculate a probability that it belongs to the same population as the model was estimated from and thus that the model is likely to yield a valid prediction. It should be noted that, as shown below, it is quite possible for a model to yield good predictions although the sample is classified as not belonging to the model. The opposite, that the sample is classified as belonging to the model and poorly predicted is uncommon. This is the behaviour required for risk assessment of substances, since a false prediction that is not detected may lead to a substance being erroneously classified as likely to be non-toxic and thus that further testing of the substance is given low priority.

2.2.4 Model validation and model accuracy measures

It is important to be able to measure model performance for different reasons, including ranking of models and estimating the reliability of predictions, when the model is used on new substances. An accuracy measure is essential in order to be able to trust and use a model prediction.

The data used to estimate the model, the training set, cannot be used to reliably estimate model performance. Two validation methods are commonly used:

- **Cross validation**. In cross validation the model is estimated a number of times. In each round, a part of the training substances are kept out. The toxicities of these substances are then predicted by the model and compared to the known (reference) values. The procedure is repeated until all samples have been kept out exactly once and cross validation prediction errors have been obtained for all substances.
- **Test set validation**. Test set validation is used when there are enough data available to exclude some of it, called the test set, from the model estimation and use it solely for validation. The model is estimated from the remaining data, the training set.

Test set validation is the most reliable method to estimate the true model performance, since if the test set is adequately selected, it is exactly equal to future model use; substances that are completely unknown to the model are predicted. Cross-validation is a reasonable substitute method if the amount of data is limited but the reliability is lower; slightly over-optimistic results are usually obtained.

For multivariate modelling methods and some other modelling methods there is a further complication. Validation is usually used both for model complexity selection (*e.g.* the number of PLS components in PLS regression). Since the model complexity selection is usually based on a prediction error criterion this can lead to so-called selection bias, which means that over-optimistic estimates of model performance are obtained. One way to deal with this problem that has been used in this work is to use cross validation to select model complexity and test set validation to estimate model performance. This means that selection bias is avoided and that very reliable estimates of model performance can be obtained.

Model performance can be measured by different metrics:

- **R^2** (or R^2Y) is the part of the variance explained in the training data, *i.e.* without validation. Thus, it does not give information about model performance for new substances. If R^2 is 1 the model explains the data perfectly, if R^2 is zero it is as good to guess a random number as to use the model.
- **Q^2** is the validation counterpart to R^2. It measures the part of the variance explained in the validation data. Q^2 can be calculated both for cross-validation, in which case it is sometimes denoted Q^2_{CV}, and for test set validation.
- **RMSEP** (root mean square error of prediction) is a measure of the prediction error and has the same unit as the response predicted by the model. It is calculated similarly to a standard deviation and can be used roughly as a standard deviation of predictions. In the formula, y is the reference value and $\hat{y}$ is the predicted value.

$$RMSEP = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

- **RMSEE** (root mean square error of estimation) the non-validated version of RMSEP, *i.e.* corresponding to R^2.

2.2.5 Outliers in QSAR models

An outlier in a QSAR model is a substance that is in some way different from the rest (majority) of the substances used to estimate the QSAR model and for which the model is not valid. The difference can be caused by different features in the chemical structure, which is closely related to the discussion above on classification of substances prior to modelling.

The common explanation of a model outlier is that it is badly predicted (has a large y residual) but this is a somewhat limited definition since a good prediction may be purely due to chance, although the substance class in question is not at all present in the training data. In multivariate statistics, it is common to define three types of outliers:

- X/Y outliers are outliers in the normal meaning, *i.e.* substances for which the relationship between the descriptors (X variables) and the environmental property (Y variable) is not valid, *e.g.* due to different toxicity mechanisms.
- X outliers. In short, a substance is an X outlier if the molecular descriptors for this substance do not conform to the "pattern" (covariance structure) in the (rest of the) training data. A different pattern in the descriptors indicates that the substance is different from the training data and thus that the prediction is likely to be inaccurate, *i.e.* a substance that is an X outlier is likely to be an X/Y outlier as well.
- Y outliers are only defined for training or test samples. They are substances for which the reference value of the response is bad for some reason.

It is important to note that outliers can be present both during training (model estimation) and model use (prediction). Naturally, since no Y value is normally available during prediction (this is why the model is used to estimate the property in question), Y cannot be present and X/Y outliers cannot be detected directly.

However, if multivariate prediction methods are used X outliers can be detected during prediction from the X residuals of the projection (also known as: distance to model in X space). This is a significant advantage of multivariate projection methods, like PLS, that facilitates automatic detection of outliers during the use of a QSAR model. This possibility is a property of the PLS method and not of the descriptors used. Thus, the advantage is present regardless of the molecular descriptors used although the success is of course dependent on the information content in the descriptors.

2.2.6 Outlier detection

Outlier detection during prediction, *i.e.* to detect substances that do not fit in the model and thus have a high risk of being poorly predicted, is very important. Since the aim of the prognosis model is typically screening of new substances and prioritising further testing it is essential to avoid false negative predictions. On the other hand, to predict false positives is less serious, since this will be revealed by the testing performed as a result of the QSAR prediction. However, also such malpredictions decrease the efficiency of the screening and prioritisation and should, naturally, be avoided if possible.

Outlier detection during prediction aims at completely avoiding grossly erroneous predictions. If the substance in question risks being badly predicted, this should be detected and the prediction should be considered unreliable and not used. Other methods for screening and prioritisation should then be used, *e.g.* other QSAR models or testing toxicity.

When PLS regression is used as the modelling method, as in this work, two measures can be considered when judging whether or not a new substance belongs to the model. The first is the distance to the model plane (also called residual magnitude) and the second is the distance between the model centre and the projection in the model plane. In the SIMCA software, the distance to the model plane of a prediction is known as DModXPS (Distance to Model in X space for the Prediction Set), while also considering the distance in the model plane leads to the statistic DModXPS+. From these distances and the corresponding distances in the training set, it is possible to calculate a probability that a (new) substance belongs to the model. These probabilities are known as PModXPS and PModXPS+, respectively, in the software.

In a previous report it was concluded that a PModXPS+ value of 0.001, *i.e.* 0.1 % theoretical risk of erroneously classifying a valid prediction as an outlier, was an appropriate risk level and gave a reasonable number of erroneous outlier indications [Furusjö *et al.* 2001].

3 Method

3.1 Descriptor Calculation

The calculation of molecular holograms was performed with the Sybyl 6.9 software from Tripos Inc. The following different settings for the parameters described in 2.1.1 were used for the molecular hologram calculation:
Fragment distinction:

1. Atoms + bonds + connectivity
2. Atoms + bonds + connectivity + chirality
3. Atoms + bonds + connectivity + hydrogens
4. Atoms + bonds + connectivity + hydrogens + chirality

Fragment length:

1. 2-5
2. 4-7
3. 6-9
4. 7-10

Hologram length:

1. 53
2. 83
3. 257
4. 401

A molecular hologram with the code 123 means that fragments in the hologram have been distinct with atoms, bonds and connectivity, and the fragment length is 4-7 and that the hologram length is 257. The total number of different holograms calculated is 64 (4^3).

3.2 Statistical Molecular Design – SMD

Statistical molecular design (SMD) was introduced with the purpose to apply experimental design methodology in QSAR modelling [Eriksson, Johansson 1996, Andersson *et al.* 2000, Eriksson *et al.* 2000]. The goal of experimental design is to select a training set for modelling that contains maximal information given the number of experiments that can be performed. In QSAR, the experiments correspond to substances but their properties (molecular descriptors) cannot be designed since they can not be controlled independently in practically all cases.

SMD uses a large number of candidate structures for which the response (y) variable does not need to be measured or known. Molecular descriptors are calculated or measured for all candidate substances and PCA is performed on the data set. The principal components that are combinations of the molecular properties are referred to as the principal properties (PP) of the data set, since they are the combinations that explain the variation among the molecules in an optimal way. The design is then performed with respect to the principal properties by selecting a subset of substances that are most efficient in spanning the substance (or PCA model) space and thus are the best selection of training set for a QSAR model. The selection can be done manually from the score plots if the number of principal components (properties) is three to four or less. An algorithm based on optimal filling of the experimental space, D-optimality, is necessary when a higher number of PCs are used. Such an algorithm can be used for low-dimensional models as well, but it is often sufficient to select samples manually.

The usefulness of SMD and multivariate techniques for exploration of principal properties has been shown by Giraud *et al.* [2000].

3.3 Estrogenicity test

3.3.1 Chemicals

Chemical products were used in tests of *in vitro* estrogenicity at the purity obtained from the suppliers without further purification. Their CAS registration numbers are given in Appendix 1 and 3.

The following products were obtained from Sigma-Aldrich Sweden AB, Stockholm: 5α-androstane-3,17-dione (A8255, purity >99 %); 4-androstene-3,17-dione (A9630, purity 98 %); (corticosterone (C2505, purity >92 %); dihydrotestosterone (androstane-17β-ol-3-one, A8380, purity not stated); esculetin (E2631, purity not stated); 17β-estradiol (E8875, purity >98 %); estriol (E1253, purity 99 %); estrone (E9750, purity 99 %); formestane (F2552, purity not stated); hexestrol (H7753, purity 98 %); 4-Hydroxytamoxifen (H7904, purity >98 %); 11-ketotestosterone (K8250, purity not stated); 17α-metyltestosterone (M7252, purity not stated); morin (M4008, purity not stated); S-naringenin (N1251, purity 95 %); norethynodrel (N-7253, purity not stated); 19-norprogesterone (N2390, purity not stated); quercetin dihydrate (Q0125, purity >98 %); pregnenolone (P9129, purity 98 %); progesterone (P8783, purity >99 %); raloxifene hydrochloride (R1402, purity not stated); rutin (R5143, purity 95 %); tamoxifen (T5648, purity >99 %); and trenbolone (T3925, purity >98 %).

Androsterone (5α -Androstan-3α-ol-17-one, 21,901-0, purity not stated); bisphenol A (23,965-8, purity >99 %); fraxetin (7,8-dihydroxy-6-methoxycoumarin, 25,491-6, purity 98 %), 5,6-didehydroandrosterone (12,578-4, purity 97 %); 17α-ethinylestradiol (28,586-2, purity not stated);

and mestranol (85,587-1, purity not stated) were products from Aldrich Chem. Co. obtained through Sigma-Aldrich Sweden AB, Stockholm. Phenolphthalein (1,07233.055) was from Merck KgaA, Darmstadt.

The following stilbene precursors and derivatives: di-piperonyl stilbene; 3',4'-dimethoxy-4-chlorostilbene; 3',4'-methylenedioxy-2,4-dichlorostilbene; 2,4-dichlorophenyl-3,4-piperonylstilbene; 3,5-dimethoxy-4-hydroxy-trans-stilbene carboxylic acid; 3,5-dimethoxy-trans-stilbene carboxylic acid; 1-(4-chlorophenyl)-2-(3,4-methylenedioxyphenyl)-ethane; 1-(2-chlorophenyl)-2-(3,4-piperonyl)-ethanol; 2'-chlorophenyl-2-(3´-4´-dimethoxyphenyl)-ethanol; 1-(2,4-dichlorophenyl)-2-hydroxy-2-(3,4-methylenedioxyphenyl)-ethanone; 1-(2-chlorophenyl)-2-hydroxy-2-(3,4-dimethoxyphenyl)-ethanone; 1-(2-chlorophenyl)-2-hydroxy-2-(3,4-methylenedioxyphenyl)-ethanone; and the following flavanone precursors and derivatives: chalcone-3,4-dimethoxyacetophenone; chalcone-4-chloroacetophenone; chalcone-O-acetate-vanillin acetovanillone were kind gifts from Dr. Alasdair Neilson. They were synthesized by procedures in literature, with slight modifications, at the purity obtained by repeated recrystallisations.

The polycyclic aromatic musk fragrances: cashmeran, celestolide, galaxolide, musk moskene (DE NIT 132), musk tibetene (NIT 133 purity 99 %), musk xylene, phantolid, tonalide, and traseolide were obtained from LGC Promochem AB, Stockholm. Musk ketone (60720 purity >97 %) was the product from Fluka Chem. Co. obtained from Sigma-Aldrich Sweden AB, Stockholm.

3.3.2 Yeast estrogen screen (YES) test

The estrogenicity test was performed with a recombinant yeast strain, *Saccharomyces cerevisiae*, with the human estrogen receptor α gene incorporated in the main chromosome [Routledge, Sumpter, 1996]. The yeast cell also contained plasmids carrying the estrogen response element and the reporter gene *lacZ* coding for β-galactosidase. This enzyme, released into the culture medium, will catalyse the conversion of the chromogenic substrate, chlorophenol red-β-D-galactopyranoside, CPRG, into a red product, which is measured by spectrophotometry in an automatic plate reader (Spectracount, Packard). The composition of media and microtitre plate procedure followed published descriptions [Routledge, Sumpter, 1996; Beresford *et al.*, 2000, Svenson *et al.* 2003]. Each test was run in triplicate. 17β-Estradiol in ethanol was run on each plate as a positive control using a serial dilution factor of 1.8. Otherwise solutions of tested substances were diluted using a factor 2.0. The plates were incubated for three-four days in darkness at 30 °C until the positive control was fully developed. After incubation each plate was shaken for 30 s and left to settle an hour before the absorbance was read at 540 nm.

As a control of cell growth, light absorption (scattering) was measured at 670 nm. Wells with reduced cell growth were not included in the calculation of estrogenicity.

3.3.3 Calculation of EC_{50}

Dose-response curves measured as absorbance at 540 nm at different concentrations of controls and tested substances were evaluated by a non-linear, exponential fit to the experimental data using the Solver program in Microsoft Excel. EC_{50}-values and slopes (s) of dose curves were derived from a minimization of the sum of deviations of the non-linear fit and the experimental data calculated according to:

$$A_{calc.} = A_{min} + (A_{max}-A_{min})*(C_i/EC_i)^s/(1+(C_i/ECi)^s)$$

EC_{50} values of tested substances were expressed in molar units and transformed to negative tenth-powers of millimolar concentrations (pEC_{50}, mM). Data was also expressed relative to the estrogenicity of the positive control 17β-estradiol.

4 Results and discussion

4.1 SMD

SMD was used to select substances for laboratory testing of estrogenicity in own experiments. The molecular hologram, *i.e.* with chiral HQSAR descriptors, indicated with the numerical code 433, was used. A PCA model was calculated for all substances for which holograms with chiral HQSAR descriptors were available, including substances with and without known *in vitro* estrogenicity. Three substances, betulin, sitosterol and kepone were removed during modelling since they were extremes in the PCA model. Kepone would cause waste disposal problems in the experimental testing.

Four principal components were used. Five regions were identified in the score plots for PC1/PC2 and PC3/PC4 as shown in Figure 1. The substances appearing in the five regions are listed in Appendix 4.

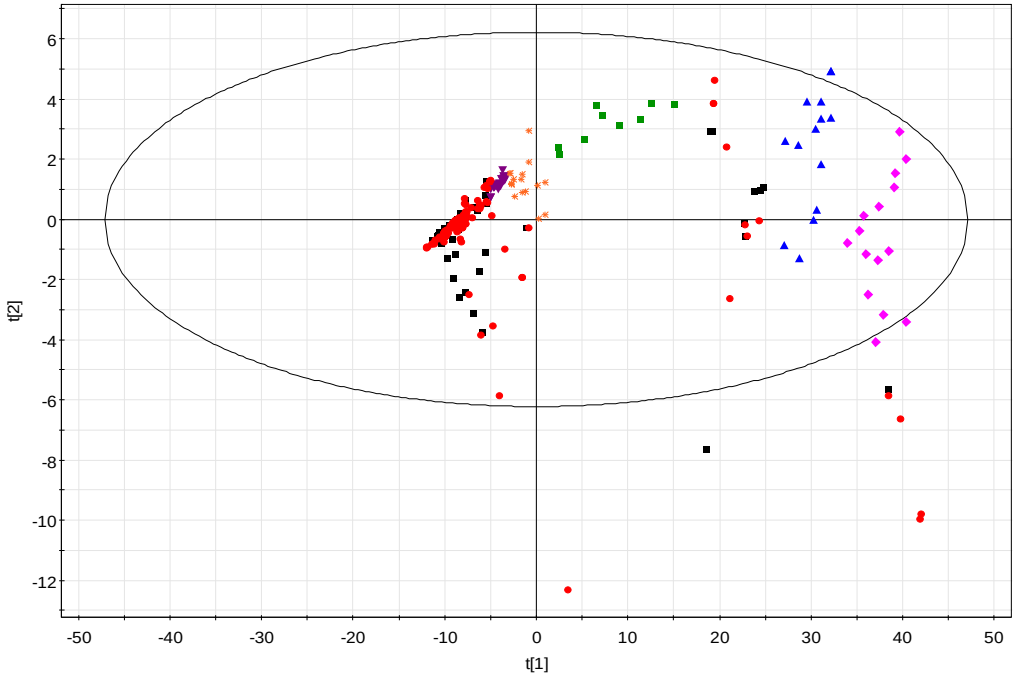


Figure 2. PCA score plot for principal component 1 and 2. ▼- region 1, * - region 2, ■ - region 3, ▲ - region 4, ♦ - region 5, ● – not assigned to any region, ● - no biological response values available.

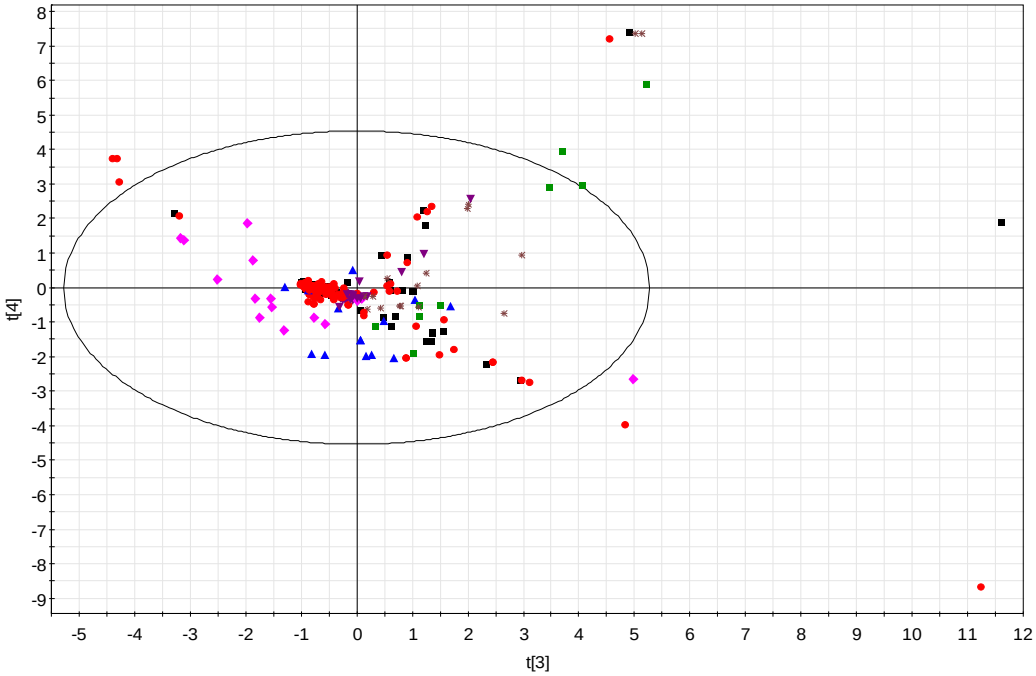


Figure 3. PCA score plot for principal component 3 and 4.▼- region 1, * - region 2, ■ - region 3, ▲ - region 4, ♦ - region 5, ● – not assigned to any region, ● - no biological response values available.

By using the score plots above a representative subset of substances from all regions was selected for YES tests. The substances that were selected are presented in Table 1.

Table 1. Substances selected for YES tests

Region	Substance
1	4-Hydroxytamoxifen
"	Fenolftalein
"	Tamoxifen
2	Raloxifene HCl
"	Hexestrol
3	Rutin
4	Norethynodrel
"	Formestane
"	4-Androstene-3,17-dione
"	5,6-Didehydroandrosterone
5	5a-Androstane-3,17-dione
"	Dihydrotestosterone
"	Androsterone
"	Metyltestosterone
"	Esculetin
"	Quercetin
"	Stilbene carbocylic acid, B 3,5-dimethoxy-4-hydroxy
"	Stilbene carboxylic acid, A 3,5-dimethoxy-
"	Musk xylene
"	Trenbolone

4.2 Estrogenicity end-points

Three estrogenic response end-points were used for modelling. Data were obtained from literature or generated by own testing using SMD for selection of substances. Furthermore three groups of substances were selected for testing: musk fragrances, chlorinated and other stilbene and chalcone derivatives and precursors. The stilbenes and chalcones were selected because of their possible occurrence in pulp bleaching processes and for musk fragrances there was a general lack of experimental data.

4.2.1 Yeast estrogen screen (YES) test, EC_{50}, absolute values

A number of sources report estrogenic responses of single compounds tested with recombinant yeast strains transfected with genes for the human estrogen receptor α. Test procedures, although basically similar, vary in performance and different factors influencing the test results have been identified [Beresford et al. 2000]. Usually 17β-estradiol has been used as a positive control, and a span in EC_{50} between 1.5 and 126 pM [*cf.* Versonnen *et al.* 2002, Schultz *et al.* 1998] emphasizes a need for selection criteria. Own experiments have gained a substantial set of data for the positive control, from which an average and measures of variation may be derived. Because of the shape of dose curves, a log-normal average of the EC_{50} was calculated using a published assay procedure [Routledge, Sumpter 1996] with some modifications [Beresford et al. 2000, Svenson et al. 2003]. The average EC_{50} for 17β-estradiol in 288 tests was 62.4 pM. The limits of one standard deviation were 33-118 pM.

Thus values of EC_{50} reported in literature with positive controls within these limits were accepted, transformed into p(EC_{50}, mM). Own experimental data were generated with this criterion. Chemical products consisting of mixtures of position isomers or stereoisomers were not included. On the other hand preparations possibly containing small impurities of highly potent hormones that could cause the test response have not been considered. All own data were obtained from tests with commercial preparations of the best purity available.

Data is presented in Appendix 1. Some of the substances that were tested in the yeast estrogen screen test were either not responsive at the concentrations administered or showed dose curves that were severely disturbed by cytotoxic effects. These results are compiled in Appendix 3. The compounds were not included in the modelling of estrogenicity.

4.2.2 Yeast estrogen screen (YES) test, EC_{50}, relative values

Estrogenic responses reported with control values within or outside the limits of estrogenicity for the reference substance, 17β-estradiol, 33-118 pM, were transformed to relative molar potencies (molar ratios between EC_{50}s of a substance and 17β-estradiol), log transformed and used for modelling. Reports not stating the estrogenicity of 17β-estradiol were not included. Data on relative estrogenic response in the yeast screen assay are compiled in Appendix 1.

4.2.3 Relative binding affinity (RBA)

A considerable amount of data has been published on the binding affinity of different substances to the estrogen receptors. Various sources report interference with estrogenic receptors prepared from different organs and a number of different mammalian species. Different types of estrogenic

receptors appear also in the same tissues. Because of a variation in specificity in receptors of different types, probably also from different organs and species, data were selected for the human estrogen α receptor. This estrogenic response was also used for modelling. Data was reported as the binding affinity relative to that of 17β-estradiol expressed in percentage on molar basis. The percentage values were log transformed. Data used are given in Appendix 2.

4.3 Modelling results

Before viewing the modelling results it can be interesting to note the distribution and span of the toxicity values in the data sets used for modelling. These are shown in the histograms in Figure 4 below and values are also given in Appendix 1 and 2.

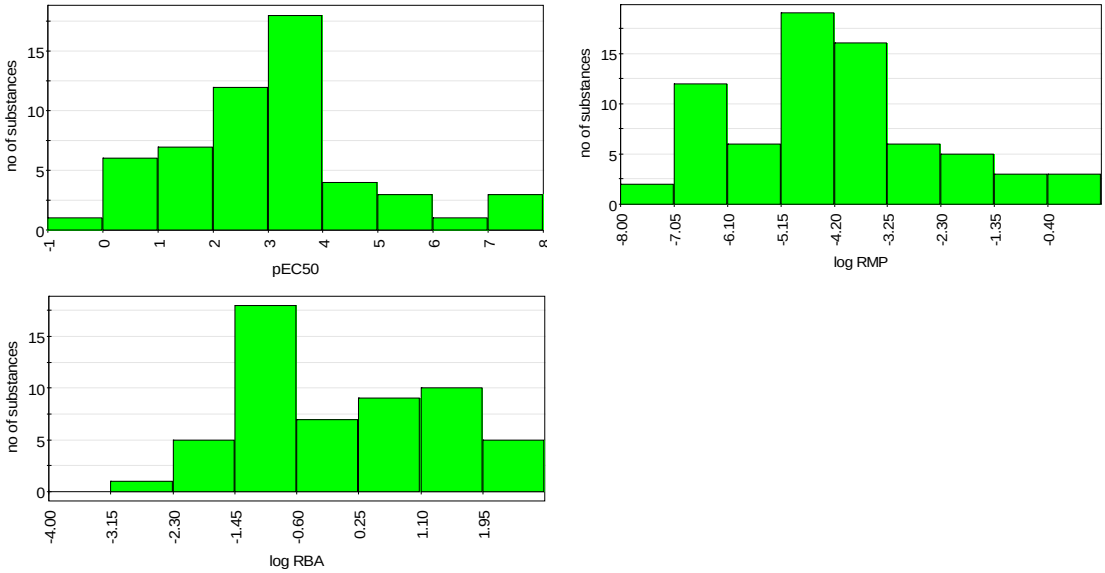


Figure 4. Histograms showing the distribution of pEC_{50} (upper left), log RMP (upper right) and log RBA (lower left) in the data.

The histograms show that all three responses were fairly normally distributed. p(EC50) varied between -0.1 and 7.7 (median 3.0), . log RMP between -7.5 and 0.5 (median -4.4), and log RBA between -2 to 2.5 (median -0.2). The spans were considerable, 4.5, 7.8 and 8.0 tenth-powers, respectively for the three data sets.

4.3.1 Screening of hologram configurations

To find the hologram configurations that constitute the best data for modelling, all configurations from 111 to 444 were investigated. To relate the chemical structure of the substances summarized in the descriptors to the biological response of interest (p(EC50), log RMP or log RBA) PLS models for every combination for each response were calculated. The best models for p(EC50), log RMP and log RBA had a Q2 value of 0.43, 0.47 and 0.54 respectively (for a summary of the best hologram configurations, see Table 2 and for Q2 for all hologram configurations, see Appendix 5 – Hologram configurations and corresponding Q2).

Table 2. Hologram configurations and corresponding Q2 values for the best models for each biological response.

	Hologram configuration	Q2	Q2 rank
p(EC50)	441	0.425	**1**
	424	0.419	**2**
	433	0.402	**3**
	324	0.39	**4**
log RMP	233	0.473	**1**
	222	0.469	**2**
	434	0.45	**3**
	211	0.434	**4**
log RBA	413	0.544	**1**
	424	0.508	**2**
	414	0.507	**3**
	411	0.493	**4**

Common for p(EC50) and log RBA was that the models with the highest Q2 value were based on data containing information about atoms, bonds, connectivity, hydrogens and chirality (*i.e.* hologram configuration 4xx). The best models for log RMP used combination 2xx or 4xx. Both 4xx and 2xx takes atoms, bonds, connectivity and chirality into account when the molecular holograms are generated. The fact that they consider chirality separates them from 1xx and 3xx.

The models for the hologram configurations with the best Q2 values for each biological response variable were selected to be further refined.

It is also worth mentioning that for the log RBA models the substances dieldrin and kepone were often missing or outliers and therefore excluded. An example is shown in Figure 5 below.

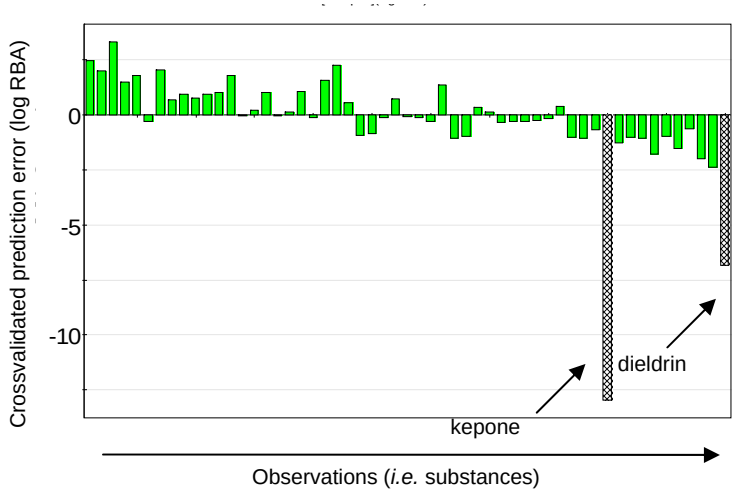


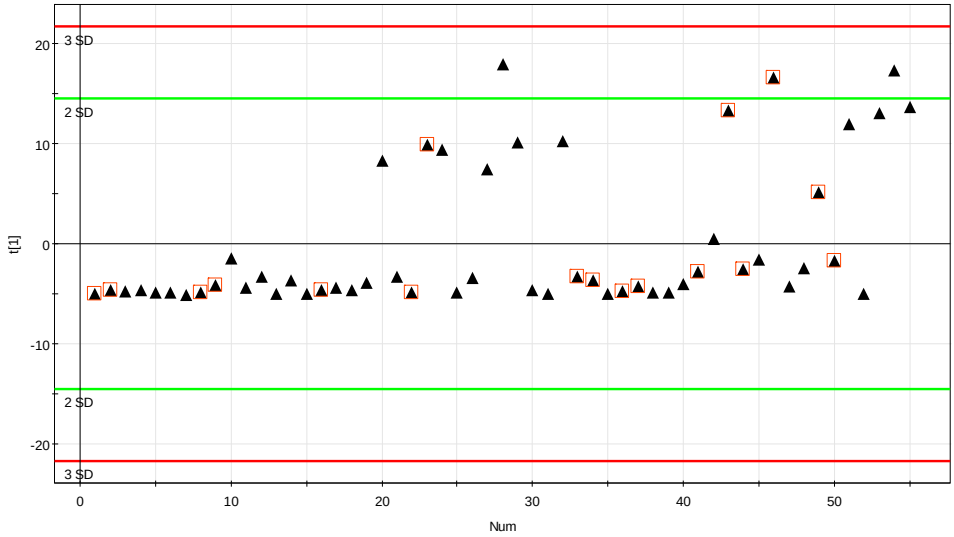
Figure 5. An example of a model where kepone and dieldrin were outliers and therefore excluded. The bar diagram shows the prediction error for each observation (substance) in the workset computed from the model with that specific observation removed.

Kepone and dieldrin had considerable prediction errors and were therefore excluded from further modelling.

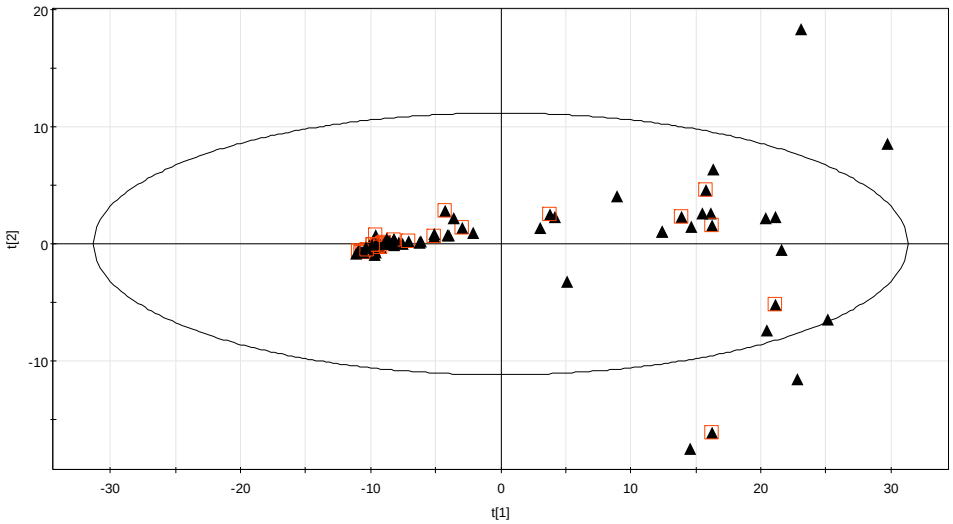
4.3.2 Selection of representative validation set

To evaluate and compare the models, one third of the observations were excluded and used as validation set for *semi*-external validation. The observations were not used in the construction of the models, but included in the data set in the selection of variables to be included in the pruned models. A validation set makes it possible to calculate RMSEP (Root Mean Square Error of Prediction). An RMSEP value of for example X means that the model can predict the value of an observation that is new to the model with a precision of ±X.

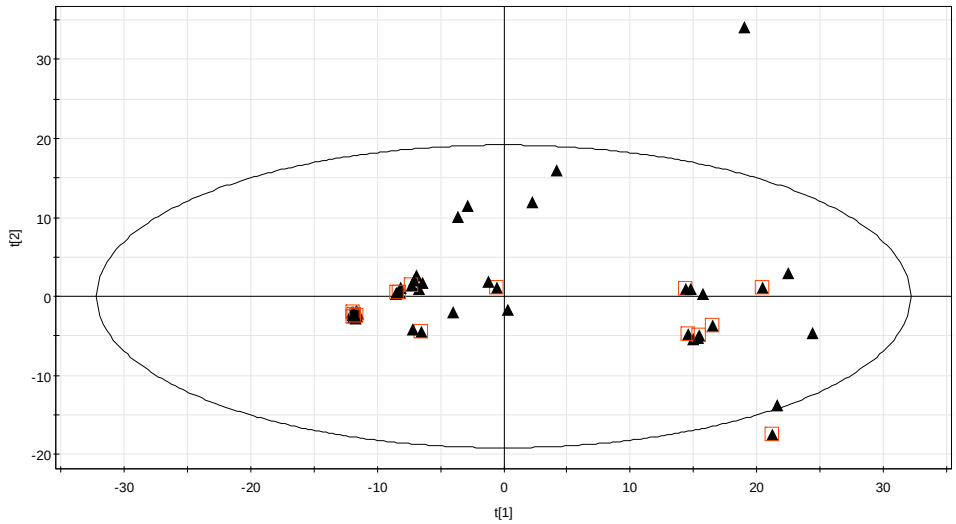
The samples that formed the validation sets were selected in such a way that they were as representative as possible for the whole data sets. This was done by calculating PCA models for p(EC_{50}), log RMP and log RBA and, by visual inspection, selecting substances that were spread in a representative way in the score plot for each model (Figure 6).



a. p(EC_{50}) (principal component 1)



b. log RMP (principal component 1 and 2)



c. log RBA (principal component 1 and 2)

Figure 6. PCA score plots showing the observations selected as validation set (marked with squares) and the observations selected as training set (not marked with squares) for the p(EC_{50}) (a), log RMP (b) and log RBA models. Each triangle represents one substance.

The substances that were selected for validation according to Figure 6 are presented in further detail in Appendix 1 and 2.

4.3.3 Models for p(EC_{50}), log RMP and log RBA

The four models with the highest Q2 value for each biological response in the screening were further refined by exclusion of variables (descriptors) that did not contribute positively to the model. By the exclusion procedure the number of variables were lowered from 53-401 to 10-115. The descriptors that the pruned models were calculated on are shown in Appendix 6 – Descriptors in pruned models. The chemical interpretation of the included hologram fragments in the descriptors used was not carried out. The Q2 value was increased for all the selected combinations, and the average Q2 increase was 27%. To investigate the presence of possible outliers and to evaluate the effect of excluding them, the PModXPS+ based outlier detection method (see chapter 2.2.6) was used. In the cases where outliers were detected with the PModXPS+ based outlier detection method and the outliers were removed from the test-set, the RMSEP values of the unpruned models decreased with 8 % and the RMSEP values of the pruned models increased with 2 %. Thus, the outlier detection tool was not as successful as was previously shown [Furusjö *et al*, 2001]. This might be due to the fact that the substances used in this case were more homogenous.

A summary of the characteristics of the original and the pruned models for every selected combination for each response is shown in Table 3.

Table 3. Properties of the pruned models compared to the unpruned models. The hologram configuration(s) resulting in the best model(s) for each response are written in bold. Original – the models based on all the variables, Pruned – the models based on pruned data.

		Original						Pruned					
	Conf	**A**	**Q2**	**RMS EE**	**RMS EP**	**RMS EP***	**Var**	**A**	**Q2**	**RMS EE**	**RMS EP**	**RMS EP***	**Var**
p(EC_{50})	441	4	0.425	1.19	1.48	-	53	3	0.470	1.24	1.47	1.49 (1)	17
	424	2	0.419	1.09	1.36	1.07 (4)	401	2	0.555	1.08	1.10	1.21 (4)	99
	433	3	0.402	1.21	1.45	1.52 (2)	257	2	0.513	1.19	1.37	-	26
	324	2	0.390	1.12	1.28	1.24 (5)	401	2	0.497	1.13	1.11	0.99 (4)	115
log RMP	233	2	0.473	1.28	1.05	-	257	3	0.539	1.12	1.06	1.10 (3)	72
	222	3	0.469	1.19	1.06	1.01 (4)	83	3	0.649	1.03	0.82	0.84 (1)	16
	434	2	0.450	1.34	1.11	-	401	2	0.526	1.27	1.08	1.11 (2)	63
	211	2	0.434	1.15	1.31	1.17 (2)	53	2	0.604	1.08	1.08	1.14 (3)	23
log RBA	413	2	0.544	0.64	1.15	1.07 (2)	253	3	0.734	0.59	0.64	-	10
	424	3	0.508	0.62	0.90	0.88 (1)	401	2	0.668	0.64	0.71	0.72 (1)	94
	414	2	0.507	0.64	1.25	1.00 (3)	364	1	0.694	0.66	0.79	0.78 (2)	30
	411	3	0.493	0.74	1.17	-	53	1	0.538	0.84	1.08	-	13

A: Number of components
RMSEP*: RMSEP when substances with PModX+ < 0.1% were removed from the validation set, number of outliers excluded is showed within brackets
Var: Number of variables included in the model

The hologram configuration 424 was one of the best configurations for modelling of p(EC_{50}). The original model, where no variables were excluded, had an RMSEP of 1.36, and the pruned model had an RMSEP of 1.10. The PModXPS+ outlier detection detected four outliers. By visual inspection of the observed vs. predicted plot (Figure 7) it seemed that the predictions of the outliers, especially for three of the substances, were as good as for the other substances. A permutation validation was performed to investigate if the pruning resulted in an increased degree of over-fit, which was not the case. The over-fit rather decreased after the pruning of the data set.

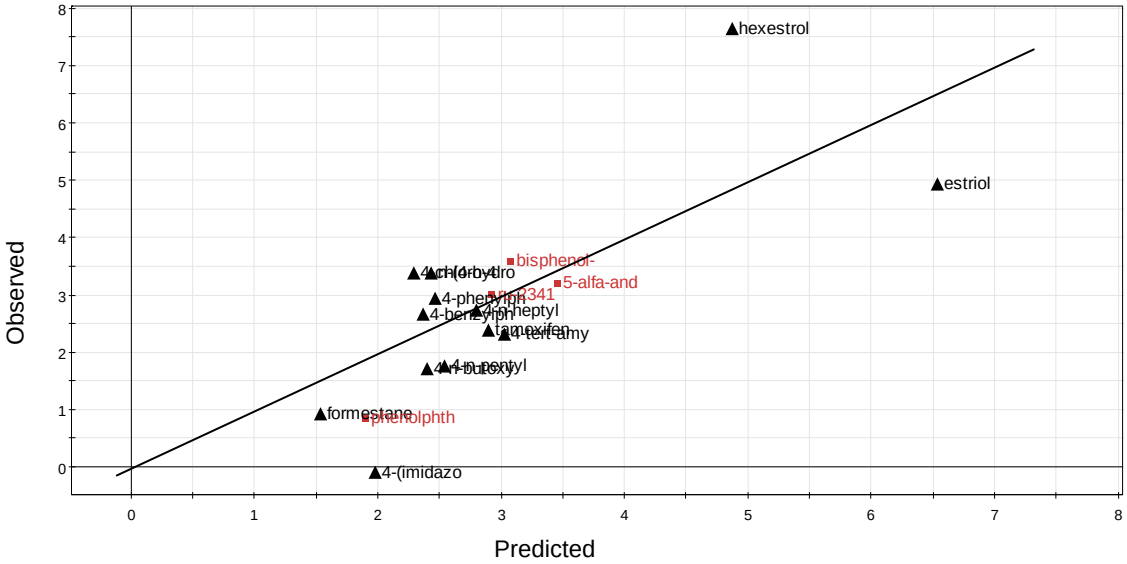


Figure 7. Observed vs predicted values for p(EC50) (*i.e.* y-axis: measured p(EC50) values, x-axis: by the model predicted p(EC50)) hologram configuration 424 (pruned model). Substances in red (■) were considered to be outliers by the PModX+ detection function and substances in black (▲) were not considered as outliers.

For log RMP, the hologram configuration 222 resulted in the best models. The RMSEP of the original model, based on all the variables, was 1.06, and pruning of that model resulted in a decrease in RMSEP to 0.84. The observed vs. predicted plot for the pruned model is shown in Figure 8. The permutation validation did not indicate that the pruning caused an increased over-fit of the model. Exclusion of the outlier that was detected by the PModXPS+ method increased the RMSEP value.

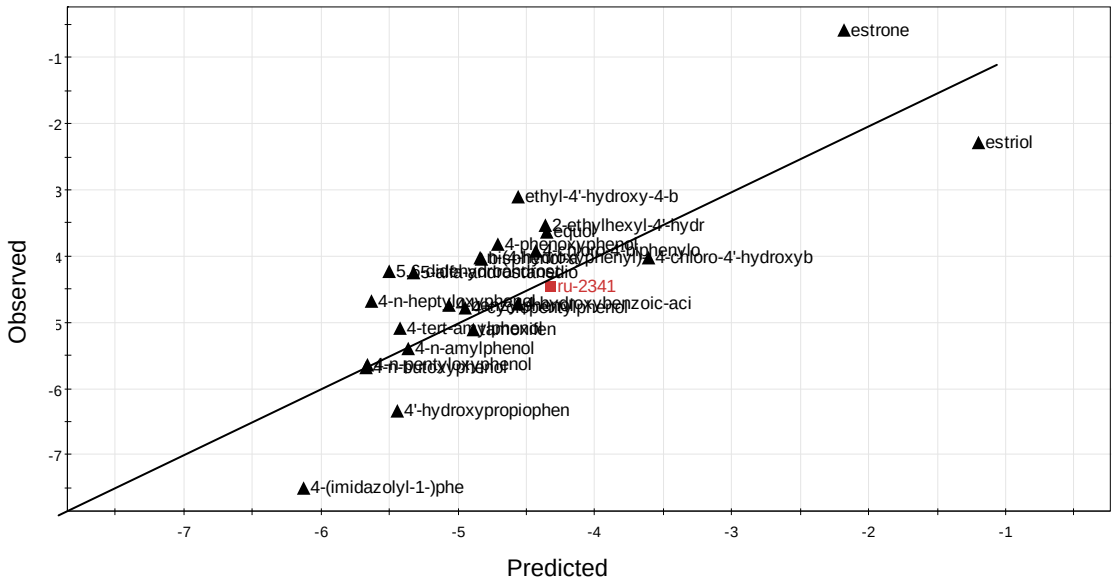


Figure 8. Observed vs. predicted values for log RMP (*i.e.* y-axis: measured log RMP values, x-axis: by the model predicted log RMP values), hologram configuration 222 (pruned model). Substances in red (■) were considered to be outliers by the PModX+ detection function and substances in black (▲) were not considered as outliers.

Hologram configuration 424 resulted in the best models for log RBA. The original model had an RMSEP of 0.90 and the pruned model (Figure 9) of 0.71. As in the log RMP case one substance was detected as an outlier in the pruned model, but exclusion of the outlier did not increase the RMSEP. The permutation validation indicated a reduced probability for over-fit of the model.

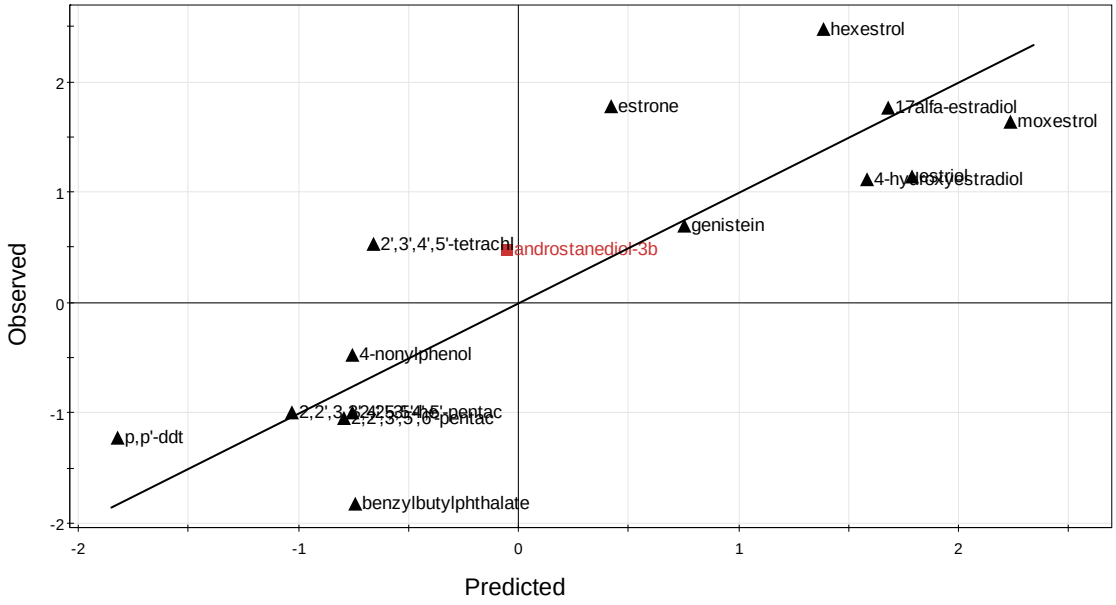


Figure 9. Observed vs predicted values for log RBA (*i.e.* y-axis: measured log RBA values, x-axis: model predicted log RBA values), hologram configuration 424 (pruned model). Substances in red (■) were considered to be outliers by the PModX+ detection function and substances in black (▲) were not considered as outliers.

RMSEP was chosen as a measure of model performance. The RMSEP value is dependent on how the validation set is selected and therefore varies with different content in the validation set. In this investigation validation sets were selected based on the position of the substances in PCA score plots, but there are also other methods for validation set selection that could be interesting to evaluate.

The RMSEP-value calculations were based on semi-external validation, where the substances that made up the validation set were also involved in the pruning procedure, since they contained information that would be unfavourable to exclude from the modelling procedure. If more substances with known biological responses were available, an ordinary external validation could be done and a more information gained.

The PModXPS+ method for outlier detection showed a tendency to classify substances as outliers even though they were proven to be relatively correctly estimated by the model. It might still be a useful tool when screening substances with unknown biological response, since it is better if a method classifies "normal" substances as outliers than classifies outliers as "normal". The PModXPS+ function could probably be improved by including more substances in the training set, since the model would then be trained on a broader diversity of substance structures.

5 Conclusions

Using molecular hologram descriptors and the multivariate method PLS, it should be possible to predict estrogenic properties in compounds before experimental testing. Three properties, *i.e.* binding affinity and induction of the human receptor estrogen α (in absolute and relative values), may be predicted using models developed in this work. To generate molecular holograms a specific program is needed. The best models predicted these properties within a factor 10 (RMSEP within 0.7 – 1.0), which should be sufficient for screening purposes. The models for binding affinity performed slightly better than the others and models based on chirality of compounds better than those not considering this property. No systematic effect of hologram length and fragment length could be noted. The chemical interpretation of the included hologram fragments in the descriptors used was not carried out. The method for outlier detection had a tendency to classify "normal" samples as outliers, but might still be a useful tool for detection of substances that do not belong to the model in a screening case where the structural span is larger.

6 Acknowledgements

The project was financially supported by the Environmental Foundation of the Swedish Association of Graduate Engineers (CFs Miljöfond) and the Foundation for the Swedish Environmental Research Institute (SIVL).

7 References

Andersson, P. M., Sjöström, M., Wold, S., Lundstedt, T. 2000. Comparison between physicochemical and calculated molecular descriptors, J. Chemom. 14, 629-642.

Beresford N., Routledge E. J., Harris C., Sumpter J.P. 2000. Issues arising when interpreting results from an in vitro assay for estrogenic activity. Toxicol. Appl. Pharmacol. 162, 22-33.

Bolger R., Wiese T. E., Ervin K., Nestich S., Checowich W. 1998. Rapid screening of environmental chemicals for estroen receptor binding capacity. Environ. Hlth. Persp. 106, 551-557.

Burden, F. R., Winkler, D. A. 1999. New QSAR methods applied to structure-activity mapping in combinatorial chemistry, J. Chem. Inf. Comput. Sci. 39, 236-242.

Buydens, L. M. C., Reijmers, T. H., Beckers, M. L. M., Weherens, R. 1999. Molecular data-mining: a challenge for chemometrics, Chemom. Intell. Lab. Syst. 49, 121-133.

EPA 2000. Summary of the endocrine disruptor priority-setting workshop. EPA Contract 68-W6-0022, Work Assignment 5-20.

Eriksson, L., Johansson, E. 1996. Multivariate design and modelling in QSAR, Chemom. Intell. Lab Syst. 34, 1-19.

Eriksson, L., Johansson, E., Müller, M., Wold, S. 2000. On the selection of the training set in environmental QSAR analysis when compounds are clustered, J. Chemom. 14, 599–616.

Esbensen, K, Schönkopf, S., Midtgaard, T. 1996. Multivariate analysis in practice, Camo A/S, Trondheim.

Furusjö E., Andersson M., Rahmberg M., Svenson A. 2001. Estimating environmentally important properties of chemicals from the chemical structure. IVL report B1517.

Gao H., Katzenellenbogen, J.A., Garg R., Hansch C. 1999a. Comparative QSAR Analysis of Estrogen Receptor Ligands. Chem. Rev. 99, 723-744.

Gao H., Williams C., Labute P., Bajorath J. 1999b Binary Quantitative Structure-Activity Relationship (QSAR) Analysis of Estrogen Receptor Ligands. J. Chem. Inf. Comput. Sci. 39, 164-168.

Geladi, P., Kowalski, B. R. 1986. Partial least squares regression: a tutorial, Anal. Chim. Acta 185, 1-17.

Giraud E., Lutmann C., Lavelle F., Riou J-F., Mailliet P., Laoui A. 2000. Multivariate data analysis using D-optimal designs, partial least squares and response surface modelling: a directional approach for the analysis of farnesyltransferase inhibitors. J. Med. Chem. 43, 1807-1816.

Kuiper G. 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. Endocrinology 138, 863-870.

Kuiper G. G. J. M., Lemmen J. G., Carlsson B., Corton J. C., Safe S. H., van der Saag P. T., van der Burg B., Gustafsson J-Å. 1998. Interaction of estrogenic chemicals and phytoestrogens with estrogen receptor β. Endocrinology 139, 4252-4263.

Livingstone, D. J. 2000. The characterisation of chemical structure using molecular properties. A survey, J. Chem. Inf. Comput. Sci. 40, 195-209.

Lowis D, 1997. HQSAR: A new, highly predictive QSAR technique. Tripos Technical Notes. Vol. 1 Number 5. HQSAR. Tripos Inc.

Martens, H., Naes, T., 1989. Multivariate calibration, John Wiley & Sons, Chichester.

Mekenyan, O., Ivanov, J., Karabunarliev, S., Bradbury, S. P., Ankley, G. T., Archer, W. 1997. A computationally-based hazard identification algorithm that incorporates ligand flexibility. 1 Identification of potential androgen receptor ligands, Env. Sci. Technol. 31, 3702-3711.

Routledge E. J., Parker J., Odum J., Ashby J., Sumpter J. P. 1998. Some alkyl hydroxy benzoate preservatives (parabens) are estrogenic. Toxicol. Appl. Pharmacol. 153, 12-19.

Routledge E. J., Sumpter J. P. 1996. Estrogenic activity of surfactants and some of their degradation products assessed using a recombinant yeast screen. Environ. Toxicol. Chem. 15, 241-248.

Saito M., Tanaka H., Takahashi A., Yakou Y. 2002. Comparison of yeast-based estrogen receptor assays. Wat. Sci. Technol. 46, 349–354.

Schmieder P. K., Aptula A. O., Routledge E. J., Sumpter J. P., Mekenyan O. G. 2000 Estrogenecity of alkylphenolic compounds: A 3-D structure-activity evaluation of gene activation. Environ. Toxicol. Chem. 19, 1727-1740.

Schultz T. W., Kraut D. H., Sayler G. S., Layton A. C. 1998. Estrogenicity of selcted biphenyls evaluated using a recombinant yeast assay. Environ. Toxicol. Chem. 17, 1727-1729.

Schultz T. W., Sinks G. D., Cronin M. T. D. 2000. Effect of substituent size and dimensionality on potency of phenolic xenoestrogens evaluated with a recombinant yeast assay. Environ. Toxicol. Chem. 19, 2637-2642

Shi L. M., ang 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.

Suzuki, T., Ide K., Ishida M., Shapiro S. 2001. Classification of environmental estrogens by physicochemical properties using principal component analysis and hierarchical cluster analysis. J. Chem. Inf. Comput. Sci. 41, 718-726.

Svenson A., Allard A-S., Ek M. 2003. Removal of estrogenicity in Swedish municipal sewage treatment plants. Wat. Res. 37, 4433-4443.

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

Tong W., Perkins R., Fang H., Hong H., Xie Q., Branham W., Sheehan D.M., Farr Anson J. 2002. Development of Quantitative Structure-Activity Relationships (QSARs) and Their Use for Priority Setting in the Testing Strategy of Endocrine Disruptors. Reg. Res. Persp. 1:3, 1-16.

Versonnen B. J., Arijs K., Verslycke T., Lema W., Janssen C. J. 2003. In vitro and in vivo estrogenicity and toxicity of o-, m-, and p-dichlorobenzene. Environ. Toxicol. Chem. 22, 329-335.

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

Appendices

Appendix 1 - Receptor induction data

Estrogens used for construction and testing of QSAR models, their yeast estrogen screen test potency (p(EC_{50}, mM) and relative molecular potency* (RMP). Values in bold face were selected as validation set for models constructed on values in normal face.

Substance	CAS No	p(EC_{50}, mM)	log RMP*	Source
17α-Ethinylestradiol	57-63-6	7,682	0,473	this study
Hexestrol	84-16-2	**7,650**	0,335	this study
17β-Estradiol	50-28-2	7.241	0.000	this study
2-Hydroxyestradiol	362-05-0		-0,523	Saito & al. 2002
Estrone	53-16-7	6,626	**-0,583**	this study
17α-Estradiol	57-91-0		-1.000	Saito & al. 2002
2',4',6'-Trichloro-4-biphenylol	14962-28-8		-1,606	Schultz & al. 1998
Norethynodrel	68-23-5	5,652	-1,663	this study
Mestranol	72-33-3	5,445	-1,671	this study
2',3',4',5'-Tetrachloro-4-biphenylol	67651-34-7		-2,000	Schultz & al. 1998
Estriol	50-27-1	**4,932**	**-2,277**	this study
2-Hydroxyestrone	362-06-1		-2,301	Saito & al. 2002
4-(1-Adamantyl)-phenol	29799-07-3	5,068	-2,340	Schultz & al. 2000
4-Hydroxytamoxifen	68047-06-3	4,513	-2,675	this study
2',5'-Dichloro-4-biphenylol	53905-28-5		-3,004	Schultz & al. 1998
Ethyl-4'-hydroxy-4-biphenyl carboxylate	50670-76-3	4,298	**-3,109**	Schultz & al. 2000
Quercetin dihydrate	117-39-5	4,315	-3,149	this study
Dihydrotestosterone	521-18-6	3.884	-3.289	this study
Benzyl-4-hydroxy-benzoic acid	94-18-8	3,971	-3,437	Schultz & al. 2000
Isoamyl-4-hydroxy-benzoate	6521-30-8	3,932	-3,476	Schultz & al. 2000
5α -Androstan-3α-ol-17-one (=androsterone)	53-41-8	3.676	-3.498	this study
2-Ethylhexyl-4'-hydroxy-benzoate	5153-25-3	3,866	**-3,541**	Schultz & al. 2000
Nonyl-4-hydroxybenzoic acid	38713-56-3	3,783	-3,625	Schultz & al. 2000
Equol	531-95-3		**-3,638**	Breinholt & Larsen, 1998
4-tert-Octylphenol	140-66-9	3,752	-3,656	Schultz & al. 2000
Phenyl-4-hydroxybenzoate	17696-62-7	3,642	-3,766	Schultz & al. 2000
4-Phenoxyphenol	831-82-3	3,582	**-3,826**	Schultz & al. 2000
4'-Chloro-4-biphenylol	28034-99-3		**-3,921**	Schultz & al. 1998
N-(4-Hydroxyphenyl)-2-naphthylamine	93-45-8	**3,382**	**-4,026**	Schultz & al. 2000
4-Chloro-4'-hydroxy-benzophenone	42019-78-3	**3,377**	**-4,031**	Schultz & al. 2000
Bisphenol A	80-05-7	**3,597**	**-4,041**	this study
Raloxifene hydrochloride	84449-90-1	3,335	-4,107	this study
Naringenin	480-41-1		-4,115	Breinholt & Larsen, 1998
4-Benzyloxyphenol	103-16-2	3,265	-4,143	Schultz & al. 2000
5,6-Didehydroandrosterone	53-43-0	2.944	**-4.229**	this study
5a-Androstane-3,17-dione	846-46-8	**3,219**	**-4,244**	this study
Genistein	446-72-0		-4,337	Breinholt & Larsen, 1998
4'-Hydroxyoctano-phenone	2589-73-3	3,053	-4,355	Schultz & al. 2000
Benzyl-4-hydroxyphenyl ketone	2491-32-9	3,045	-4,363	Schultz & al. 2000
4-Hydroxybenzophenone	1137-42-4	3,007	-4,401	Schultz & al. 2000

Substance	CAS No	p(EC_{50}, mM)	log RMP*	Source
Trenbolone	10161-33-8	**3,010**	**-4,453**	this study
4-Phenylphenol	92-69-3	**2,939**	-4,469	Schultz & al. 2000
4-Hydroxybenzoic acid, propyl ester	94-13-3		-4,477	Routledge & al. 1998
4,4'-Biphenol	92-88-6		-4,634	Schultz & al. 1998
4-n-Heptyloxyphenol	13037-86-0	**2,726**	**-4,682**	Schultz & al. 2000
4-n-Octylphenol	1806-26-4	2,724	-4,684	Schultz & al. 2000
4-Hydroxybenzoic acid, n-butyl ester	94-26-8	2,697	**-4,711**	Schultz & al. 2000
4-Benzylphenol	101-53-1	**2,674**	**-4,734**	Schultz & al. 2000
Metyltestosterone	58-18-4	2.426	-4.747	this study
4-Cyclopentylphenol	1518-83-8	2,618	**-4,790**	Schultz & al. 2000
4-Hexyloxyphenol	18979-55-0	2,377	-5,031	Schultz & al. 2000
4-tert-Amylphenol	80-46-6	**2,322**	**-5,085**	Schultz & al. 2000
Tamoxifen	10540-29-1	**2.386**	**-5.102**	this study
4-n-Amylphenol	14938-35-3	2,022	**-5,386**	Schultz & al. 2000
3',5'-Dimethoxy-4'-hydroxy-trans-stilbene carboxylic acid		2,059	-5,405	this study
4-n-Pentyloxyphenol	18979-53-8	**1,762**	**-5,646**	Schultz & al. 2000
4-Androstene-3,17-dione	63-05-8	1,637	-5,678	this study
4-n-Butoxyphenol	122-94-1	**1,726**	**-5,682**	Schultz & al. 2000
Phloretin	60-82-2		-6,036	Breinholt & Larsen, 1998
3',5'-dimethoxy-trans-stilbene carboxylic acid		1,246	-6.081	this study
4-tert-Butylbenzoic acid	98-73-7	1,124	-6,085	this study
N-Benzyl-4-hydroxyaniline	103-14-0	1,203	-6,205	Schultz & al. 2000
4-Hydroxybenzoic acid, ethyl ester	120-47-8	1,124	-6,284	Schultz & al. 2000
4'-Hydroxypropio-phenone	70-70-2	1,080	**-6,328**	Schultz & al. 2000
4-Hydroxybenzoic acid, methyl ester	99-76-3		-6,398	Routledge & al. 1998
2,6-Diisopropyl-naphthalene	24157-81-1		-6,423	Vinggaard & al. 2000
Musk xylene	81-15-2	0,961	-6,503	this study
Formestane	566-48-3	**0,923**	-6,519	this study
4-Propylphenol	10210-17-0	0,824	-6,584	Schultz & al. 2000
Phenolphthalein	77-09-8	**0,849**	-6,593	this study
4-Propoxyphenol	18979-50-5	0,785	-6,623	Schultz & al. 2000
2-Hydroxybiphenyl	90-43-7		-6,728	Vinggaard & al. 2000
Rutin	153-18-4	0,673	-6,769	this study
Esculetin	305-01-1	0,386	-6,928	this study
4-Chloro-3-methylphenol	59-50-7		-7,276	Vinggaard & al. 2000
4-(Imidazolyl-1-)phenol	10041-02-8	**-0,097**	**-7,505**	Schultz & al. 2000

17β-Estradiol reference compound

Appendix 2 - Binding affinity data

Relative binding affinity data on substances used for construction and testing of human estrogen receptor α RBA QSAR models. Values in bold face were selected as validation set for models constructed on values in normal face.

Substance	CAS No	log RBA	Source
Diethylstilbestrol	56-53-1	2,670	Kuiper & al, 1997
Hexestrol	84-16-2	**2,480**	Kuiper & al, 1997
4-Hydroxytamoxifen	68047-06-3	2,410	Kuiper & al, 1997
Dienestrol	84-17-3	2,348	Kuiper & al, 1997
17β-Estradiol	50-28-2	2,000	Kuiper & al, 1997
Coumestrol	479-13-0	1,973	Kuiper & al, 1997
ICI-164384		1,929	Kuiper & al, 1997
16α-Bromo-17β-estradiol	54982-79-5	1,881	Kuiper & al, 1998
Raloxifene	84449-90-1	1,839	Kuiper & al, 1998
Estrone	53-16-7	**1,778**	Kuiper & al, 1997
17α-Estradiol	57-91-0	**1,763**	Kuiper & al, 1997
Nafoxidine	1845-11-0	1,643	Kuiper & al, 1997
Moxestrol	34816-55-2	**1,633**	Kuiper & al, 1997
17-Epiestriol	547-81-9	1,462	Kuiper & al, 1998
Clomifene	911-45-5	1,398	Kuiper & al, 1997
β-Zearalanol	26538-44-3	1,204	Kuiper & al, 1997
Estriol	50-27-1	**1,146**	Kuiper & al, 1997
4-Hydroxyestradiol	5976-61-4	**1,114**	Kuiper & al, 1997
Zearalenone	17924-92-4	1,000	Kuiper & al, 1998
2-Hydroxyestradiol	362-05-0	0,845	Kuiper & al, 1997
Tamoxifen	10540-29-1	0,845	Kuiper & al, 1997
5-Androstenediol	521-17-5	0,778	Kuiper & al, 1997
Genistein	446-72-0	**0,699**	Kuiper & al, 1997
2',3',4',5'-Tetrachloro-4-biphenylol	67651-34-7	**0,531**	Kuiper & al, 1997
3β-Androstanediol	571-20-0	**0,477**	Kuiper & al, 1997
2',4',6'-Trichloro-4-biphenylol	14962-28-8	0,380	Kuiper & al, 1997
2-Hydroxyestrone	362-06-1	0,301	Kuiper & al, 1998
HPTE		0,230	Bolger & al, 1998
16-Keto-17β-estradiol	566-75-6	0,114	Kuiper & al, 1998
Norethrynodrel	68-23-5	-0,155	Kuiper & al, 1997
4-Androstenediol	1156-92-9	-0,301	Kuiper & al, 1997
o,p'-DDT	789-02-6	-0,317	Bolger & al, 1998
2,2',3',4',6'-Pentachloro-4-biphenylol	59512-50-4	-0,523	Kuiper & al, 1998
2,2',4',6'-Tetrachloro-4-biphenylol	150304-08-8	-0,523	Kuiper & al, 1998
4-Nonylphenol	104-40-5	**-0,477**	Bolger & al, 1998
4-t-Octylphenol	140-66-9	-0,761	Bolger & al, 1998
Daidzein	486-66-8	-0,699	Kuiper & al, 1998
2',3,4',6'-Tetrachloro-4-biphenylol	150304-08-8	-0,745	Kuiper & al, 1998
2',3,3',4',6'-Pentachloro-4-biphenylol		-0,886	Kuiper & al, 1998
2',3,3',4',5'-Pentachloro-4-biphenylol		-0,959	Kuiper & al, 1998
2,2',3,3',4',5,5'-Heptachloro-4-biphenylol	158076-64-3	**-1,000**	Kuiper & al, 1998
2,2',3,4',5,5',6-Heptachloro-4-biphenylol	158076-68-7	-1,000	Kuiper & al, 1998
2,2',3',4',5'-Pentachloro-4-biphenylol		**-1,000**	Kuiper & al, 1998
2,2',3',4,4',5,5'-Heptachloro-4-biphenylol		-1,046	Kuiper & al, 1998
2,2',3',5',6'-Pentachloro-4-biphenylol		**-1,046**	Kuiper & al, 1998
2,2',3,3',4',5-Hexachloro-4-biphenylol	158076-62-1	-1,155	Kuiper & al, 1998
3α-Androstanediol	1852-53-5	-1,155	Kuiper & al, 1997
Norethindrone	68-22-4	-1,155	Kuiper & al, 1997
2',3,3',5',6'-Pentachloro-4-biphenylol		-1,222	Kuiper & al, 1997

Substance	CAS No	log RBA	Source
Chlordecone (Kepone)	143-50-0	-1,222	Kuiper & al, 1998
p,p'-DDT	50-29-3	**-1,222**	Bolger & al, 1998
5α-Dihydrotestosterone	521-18-6	-1,301	Kuiper & al, 1997
Bisphenol A	80-05-7	-1,301	Kuiper & al, 1997
Dehydroepiandrosterone	53-43-0	-1,398	Kuiper & al, 1997
2,2',3,4',5,5'-Hexachloro-4-biphenylol	145413-90-7	-1,523	Kuiper & al, 1997
2,3,3',4',5'-Pentachloro-4-biphenylol		-1,523	Kuiper & al, 1997
4-n-Octylphenol	1806-26-4	-1,699	Kuiper & al, 1998
Benzylbutylphthalate	85-68-7	**-1,824**	Bolger & al, 1998
19-Nortestosterone	434-22-0	-2,000	Kuiper & al, 1997
2',3,3'4',5-Tetrachloro-4-biphenylol		-2,000	Kuiper & al, 1997
Methoxychlor	72-43-5	-2,000	Kuiper & al, 1997
5α-Dihydrotestosterone	521-18-6	-2,018	Bolger & al, 1998
Dieldrin	60-57-1	-3,049	Bolger & al, 1998

* 17β-Estradiol reference compound

Appendix 3 - Receptor induction data below detection limit

Substances tested with no response in the yeast screen. Experimental data that was not used in model construction.

Substance	CAS Registration No	Type of substance	EC_{50} (mg/L)	Maximal tested concentration (mg/L)
Chalcone-3,4-dimethoxyacetophenone		Flavanone derivative	>> 149	
Chalcone-4-chloroacetophenone		Flavanone derivative	>> 137	
Chalcone-O-acetate-vanillin acetovanillone		Flavanone derivative	>> 525	
Cashmeran	33704-61-9	Musk fragrance	>> 245	
Celestolide	13171-00-1	Musk fragrance	>> 145	
Galaxolide	1222-05-5	Musk fragrance	>> 120	24
Musk ketone	81-14-1	Musk fragrance	> 100	
Musk moskene	116-66-5	Musk fragrance	> 1000	1000
Musk tibetene	145-39-1	Musk fragrance	>> 380	
Phantolid	15323-35-0	Musk fragrance	>> 145	
Tonalide	1506-02-1	Musk fragrance	>> 990	
Traseolide	68140-48-7	Musk fragrance	>> 35	
Fraxetin	574-84-5	Phytoestrogen	>> 250	
Morin	480-16-0	Phytoestrogen	> 10	60
S-Naringenin	93602-28-9	Phytoestrogen	> 100	500
Corticosterone	50-22-6	Steroid hormone	>>500	
11-Ketotestosterone	564-35-2	Steroid hormone	> 50	100
Norprogesterone	472-54-8	Steroid hormone	> 85	85
Pregnenolone	145-13-1	Steroid hormone	> 30	30
Progesterone	57-83-0	Steroid hormone	>> 1000	
2,4-Dichlorophenyl-3,4-piperonylstilbene		Stilbene derivative	> 30	77.5
3',4'-Dimethoxy-4-chlorostilbene		Stilbene derivative	>> 525	
3',4'-Methylenedioxy-2,4-dichlorostilbene		Stilbene derivative	>> 725	
1-(4-Chlorophenyl)-2-(3,4-methylenedioxyphenyl)-ethane		Stilbene derivative	>> 285	
1-(2-Chlorophenyl)-2-(3,4-piperonyl)-ethanol		Stilbene derivative	>> 655	
2'-Chlorophenyl-2-(3´-4´-dimethoxyphenyl)-ethanol		Stilbene derivative	>> 515	
1-(2,4-Dichlorophenyl)-2-hydroxy-2-(3,4-dimethoxyphenyl)-ethanone		Stilbene derivative	>> 45	
1-(2,4-Dichlorophenyl)-2-hydroxy-2-(3,4-methylenedioxyphenyl)-ethanone		Stilbene derivative	>> 197	
1-(2-Chlorophenyl)-2-hydroxy-2-(3,4-methylenedioxyphenyl)-ethanone		Stilbene derivative	>> 164	

Appendix 4 – SMD selection

Substances in the five regions identified by SMD

Region 1	**Region 2**	**Region 3**	**Region 4**	**Region 5**
3-(4-Phenol)-2'-methane-indene	3,3-bis(4-Hydroxyphenyl)-pentane	Cashmeran	11β-Hydroxy androstenedione	11-Keto androsterone
3-(4-Phenol)-2'trifluormethane-indene	4,4-bis-(4-Hydroxyphenyl)-heptane	Celestolide	11-Hydroxy androstenedione	17α-Hydroxy progesterone
3-(4-Phenol)-3'-nitro-indene	4-tert-Heptylphenol	Equilenin	11-Ketotestosterone	4-Androstenediol
3-(4-Phenol)-4'-bromo-indene	4-tert-Hexylphenol	Galaxolide	17α-Methylestradiol	5α-Androstane dione
3-(4-Phenol)-4'-hydroxy-indene	4-tert-Octylphenol-diethoxylate	Musk moskene	1-Dehydro testosterone	5β-Androstane-3β-11β-diol-17-one
3-(4-Phenol)-4'-nitrilo-indene	4-tert-Octylphenol-tertethoxylate	Nafoxidine	5,6-Didehydro androsterone	Androstane-3,17-dione
3-(4-Phenol)-4'-nitro-indene	3,3'-Dihydroxyhexestrol	Phantolid	Androstenedione	Corticosterone
3-(4-Phenol)-6-hydroxy-indene	Hexestrol	RU-2341	Estradiol-17β-glucuronide	ICI182780
3-(4-Phenol)-indene-2	Indenestrol-A	Rutin	Formestane	Methyltestosterone
3-(4-Phenol)-indene	Indenestrol-A-(R-enantiomer)		Moxestrol	Mibolerone
3-Benzyl-4'-hydroxy-indene	Indenestrol-A-(S-enantiomer)		Norethynodrel	Norprogesterone
3-Benzyl-4'-nitro-indene	Indenestrol-B		Norluton	Pregnenolone
3-Benzyl-indene	Musk ketone		Nortestonate	Progesterone
3-Ethyl-4'-hydroxy-indene	Musk tibetene		RU-2453	RU-1364
3-Ethyl-6-hydroxy-indene	Nordihydroguaiaretic acid			Testosterone
3-Phenyl-4'-hydroxy-indene	Procymidone			
Clomifene	Raloxifene			
2,2-bis-(4-Hydroxyphenyl)-propane				
Bisphenol-B				
Droloxifene				
4-Hydroxytamoxifen				
Musk xylene				
Phenolphthalein				
RU-56187				
Tamoxifen				
Toremifene				
Zindoxifene				

Appendix 5 – Hologram configurations and corresponding Q2

Q2 values for PLS-models for p(EC50), log RMP and log RBA based on all combinations of hologram configurations. The hologram configurations are also ranked from 1 to 64 according to Q2 value – the higher the Q2 the better the rank. The hologram configurations marked with grey and bold face correspond to the four highest Q2 values for each response.

	p(EC50)		log RMP		log RBA	
Hologram configuration	Q2	Q2 rank	Q2	Q2 rank	Q2	Q2 rank
111	0.257	42	0.377	28	0.315	54
112	0.056	63	0.184	62	0.366	40
113	0.125	61	0.196	61	0.306	56
114	0.199	52	0.228	58	0.33	52
121	0.382	7	0.382	24	0.271	58
122	0.335	22	0.328	49	0.337	51
123	0.149	59	0.349	40	0.401	29
124	0.206	51	0.286	56	0.419	23
131	0.255	43	0.298	55	0.412	25
132	0.36	14	0.398	20	0.444	14
133	0.358	15	0.401	19	0.412	26
134	0.286	37	0.342	45	0.432	18
141	0.267	40	0.328	48	0.443	15
142	0.3	33	0.355	38	0.459	9
143	0.341	20	0.379	27	0.406	28
144	0.309	32	0.346	41	0.412	27
211	0.237	48	**0.434**	**4**	0.354	49
212	0.087	62	0.271	57	0.427	19
213	0.217	49	0.322	51	0.448	12
214	0.24	47	0.342	44	0.452	11
221	0.211	50	0.361	37	0.358	45
222	0.376	9	**0.469**	**2**	0.366	41
223	0.164	56	0.395	22	0.439	17
224	0.253	45	0.416	13	0.464	7
231	0.24	46	0.31	54	0.37	39
232	0.38	8	0.423	11	0.464	8
233	0.373	10	**0.473**	**1**	0.383	34
234	0.355	17	0.41	15	0.441	16
241	0.259	41	0.342	43	0.364	42
242	0.369	11	0.409	16	0.394	32
243	0.352	18	0.407	17	0.4	30
244	0.333	24	0.386	23	0.447	13
311	0.157	57	0.212	59	0.359	43
312	0.126	60	0.152	64	0.455	10
313	0.05	64	0.154	63	0.425	21
314	0.169	55	0.211	60	0.359	44

Hologram configuration	p(EC50) Q2	p(EC50) Q2 rank	log RMP Q2	log RMP Q2 rank	log RBA Q2	log RBA Q2 rank
321	0.254	44	0.411	14	0.144	61
322	0.327	26	0.323	50	0.379	35
323	0.357	16	0.37	31	0.397	31
324	**0.39**	**4**	0.351	39	0.425	22
331	0.296	34	0.34	46	0.267	60
332	0.289	36	0.342	42	0.325	53
333	0.314	29	0.372	30	0.305	57
334	0.335	21	0.372	29	0.418	24
341	0.322	27	0.381	25	0.375	36
342	0.277	39	0.318	52	0.315	55
343	0.347	19	0.367	33	0.387	33
344	0.32	28	0.379	26	0.352	50
411	0.152	58	0.333	47	**0.493**	**4**
412	0.183	54	0.362	36	0.474	5
413	0.186	53	0.31	53	**0.544**	**1**
414	0.31	31	0.406	18	**0.507**	**3**
421	0.283	38	0.368	32	0.126	63
422	0.311	30	0.365	34	0.136	62
423	0.333	23	0.397	21	0.469	6
424	**0.419**	**2**	0.424	10	**0.508**	**2**
431	0.363	13	0.433	6	0.118	64
432	0.291	35	0.364	35	0.427	20
433	**0.402**	**3**	0.433	5	0.371	38
434	0.384	6	**0.45**	**3**	0.357	46
441	**0.425**	**1**	0.43	8	0.269	59
442	0.332	25	0.421	12	0.375	37
443	0.368	12	0.428	9	0.355	47
444	0.388	5	0.431	7	0.355	48

Appendix 6 – Descriptors in pruned models

Descriptors used in the pruned models for the estrogenic endpoints p(EC_{50}), log RMP and log RBA.

p(EC50)			log RMP	log RBA		
Descriptor configuration 424			Descriptor configuration 222	Descriptor configuration 424		
6	140	274	3	11	122	254
10	143	275	6	15	124	269
17	151	278	8	19	126	276
24	152	280	13	20	135	280
32	155	285	14	21	136	281
34	161	287	24	23	139	283
37	162	290	42	24	142	290
47	163	296	46	28	149	291
48	166	299	50	31	151	293
51	167	300	62	36	152	296
56	168	309	64	37	153	297
57	169	311	75	43	155	314
60	178	313	76	48	159	320
61	179	317	78	53	163	324
73	181	320	81	54	169	326
77	190	325	82	55	178	328
83	202	331		58	179	329
88	208	336		60	187	336
90	209	339		61	191	342
91	222	342		67	199	344
94	223	345		71	202	346
98	229	348		76	205	356
103	232	350		80	207	357
106	238	360		84	215	360
110	240	364		94	225	366
113	244	368		95	228	373
115	245	370		96	239	380
116	246	371		99	242	381
117	249	380		106	245	384
124	257	381		113	247	388
125	258	383		115	248	394
132	260	393		120		
134	265	398				