4D-Fingerprint Categorical QSAR Models for Skin Sensitization Based on the Classification of Local Lymph Node Assay Measures

ARTICLE *in* CHEMICAL RESEARCH IN TOXICOLOGY · FEBRUARY 2007

Impact Factor: 3.53 · DOI: 10.1021/tx6002535 · Source: PubMed

7 AUTHORS, INCLUDING:

Yufeng Jane Tseng
National Taiwan University
52 PUBLICATIONS **468** CITATIONS
SEE PROFILE

Frank Gerberick
Procter & Gamble
196 PUBLICATIONS **6,917** CITATIONS
SEE PROFILE

Published in final edited form as:
Chem Res Toxicol. 2007 January ; 20(1): 114–128. doi:10.1021/tx6002535.

4D-Fingerprint Categorical QSAR Models for Skin Sensitization Based on Classification Local Lymph Node Assay Measures

Yi Li[$], **Yufeng J. Tseng**[%,&], **Dahua Pan**[$], **Jianzhong Liu**[@,&], **Petra S. Kern**[#], **G. Frank Gerberick**[##], and **Anton J. Hopfinger**[@,&,*]

$ *Laboratory of Molecular Modeling and Design (MC 781), College of Pharmacy, University of Illinois at Chicago, 833 South Wood Street, Chicago, IL 60612-7231*

@ *College of Pharmacy, MSC09 5360, 1 University of New Mexico, Albuquerque, NM 87131-0001*

Procter& Gamble Eurocor, Temselaan 100, B-1853 Strombeek-Bever, Belgium

##*The Procter & Gamble Company, Miami Valley Innovation Center, P.O. Box 538707, Cincinnati, OH 45253-8707*

& *The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, IL 60045*

% *Dept. of Computer Science and Information Engineering, National Taiwan University, No.1 Sec. 4, Roosevelt Road, Taipei, Taiwan 106*

Abstract

Currently, the only validated methods to identify skin sensitization effects are *in vivo* models, such as the Local Lymph Node Assay (LLNA) and guinea pig studies. There is a tremendous need, in particular due to novel legislation, to develop animal alternatives, eg. Quantitative Structure-Activity Relationship (QSAR) models. Here, QSAR models for skin sensitization using LLNA data have been constructed. The descriptors used to generate these models are derived from the 4D-molecular similarity paradigm and are referred to as universal 4D-fingerprints. A training set of 132 structurally diverse compounds and a test set of 15 structurally diverse compounds were used in this study. The statistical methodologies used to build the models are logistic regression (LR), and partial least square coupled logistic regression (PLS-LR), which prove to be effective tools for studying skin sensitization measures expressed in the two categorical terms of sensitizer and non-sensitizer. QSAR models with low values of the Hosmer-Lemeshow goodness-of-fit statistic, χ^2_{HL}, are significant and predictive. For the training set, the cross-validated prediction accuracy of the logistic regression models ranges from 77.3% to 78.0%, while that of PLS-logistic regression models ranges from 87.1% to 89.4%. For the test set, the prediction accuracy of logistic regression models ranges from 80.0%-86.7%, while that of PLS-logistic regression models ranges from 73.3%-80.0%. The QSAR models are made up of 4D-fingerprints related to aromatic atoms, hydrogen bond acceptors and negatively partially charged atoms.

Keywords

Skin Sensitization; QSAR Analysis; Logistic Regression; 4D-fingerprints

* Corresponding Author: Voice: 505.272.8474, Fax: 505.272.0704, Email: hopfingr@unm.edu.

Introduction

Allergic contact dermatitis (ACD) results from the T-lymphocyte mediated immune response to a chemical allergen that comes into contact with the skin[1]. The small allergenic molecule (hapten) penetrates the skin and binds to a carrier protein, typically by a covalent bond, to form an antigenic hapten-protein complex. This complex is then processed by antigen presenting cells, which migrate to the draining lymphnodes where they present the haptens to the T-lymphocytes. The ability of a chemical compound to behave as a contact allergen depends on its ability to penetrate the stratum corneum and on its ability to react, either directly or after metabolic activation, with skin proteins.

Currently, the only validated methods to identify skin sensitization effects are *in vivo* models such as the murine Local Lymph Node Assay (LLNA).[2] There is a tremendous need, particularly in Europe due to new legislative initiatives such as REACH and the 7th Amendment to the Cosmetics Directive[3], to develop new approaches as alternatives to animal testing strategies. Quantitative structure activity relationships (QSARs) are increasingly seen as playing a role in compound evaluation and screening, and are considered an important alternative for the estimation of toxicity effects.

Landsteiner and Jacobs led in establishing the occurrence of allergic contact dermatitis with protein reactivity to allergens.[4] These hapten-protein reactions usually involve the allergen, or its metabolite, acting as an electrophile, and a protein acting as a nucleophile. Based on this concept, it was shown that structure activity relationships can be established, relating skin sensitization potential with physicochemical parameters for a variety of chemical classes using a variety of sensitization assay data. Most of the published QSARs have been individually developed for a particular class of chemicals. Structure-skin sensitization correlations have been derived mainly for homologous series of chemicals.[5-7] Roberts and Williams established SARs based on electrophilicity and hydrophobicity parameters using the relative alkylation index (RAI). This model has been used to evaluate data on various sets of skin sensitizing chemicals, including alkylating agents[8], sulfonate esters and acrylates[9], α,β- diketones[10], aldehydes[11] and other classes of chemicals.[12]

Predictive studies of skin sensitization can be divided into two different major categories: mechanism-based and data correlation-based. Most of work carried out until the early 1990's belong to the former category. [13-15] The limitation of these types of predictive QSAR models is that they are restricted to homolog sets of chemicals, which in effect ensures that the chemicals studied will have a common mechanism of action. To develop a predictive QSAR model for a heterogeneous database, categorical statistical methods have been applied since the middle 1990's, in which a variety of descriptors have been computed for the chemicals selected. Discriminant analysis has been the most widely adopted categorical method,[16-18] while logistic regression, LR, has shown increasing popularity. LR requires few assumptions, in theory, and is easier to use and understand than is discriminant analysis.[19]

Early skin sensitization QSAR models are mainly based on databases using the guinea pig maximization test (GPMT).[20,21] However, this test over-predicts the response and, the intensity score measuring the extent of sensitization is usually expressed as an ordinal number. That is, in the GPMT the result relies on subjective evaluation of a group of animals, and the result of a GPMT study is usually a “yes” or “no” answer. A potency catergorization (ie. weak/ moderate/ strong sensitizer) cannot be estimated with great certainty.

In contrast, the development of the LLNA has facilitated the use of QSARs to predict skin sensitization potential. It provides a standardized continuous scale suitable for quantitative assessment of skin sensitization.[22] The extent of sensitization is quantified from a measurement of the cellular proliferation induced in the lymph nodes draining the site of

chemical application. The response data are usually represented by a stimulation index (SI) that compares the proliferation for each test material to that from a vehicle control. From the dose-response curve EC3, the estimated concentration of a test material required to produce a stimulation index of 3, can be calculated and used as a response variable for biological activity in the construction of a continuous QSAR model. In addition, the skin sensitization potency of a chemical can be categorized into potency classes according to its EC3 value.[23] A molecule which is tested at a high concentrations and with SI values of less than three at each tested concentration is treated as a non-sensitizer, that is, no EC3 value can be determined. Consequently, before a continuous QSAR model can be built to predict the EC3 value for an unknown chemical, it is necessary to build a categorical model that can determine if the molecule is a skin sensitizer or a non-sensitizer, or, ideally, discriminate across potency classes, eg. weak/ moderate/ strong sensitizers in contrast to non-sensitizers.

We have employed various “standard” descriptors, like those derived from the electrophilicity and chemical structure topology of a molecule, as well as estimated molecular properties, such as hydrophobicity, as independent variables in the QSAR modeling of the skin sensitization data set used in the work reported in this paper. Various types of statistical tools for both data fitting and for model optimization were also employed this QSAR modeling work. These efforts were viewed as extensions to earlier QSAR studies of skin sensitization using different data sets. However, all of these efforts to model the LLNA EC3 data set reported in this paper were unsuccessful.

Recently, a set of universal descriptors called 4D-fingerprints, have been derived from a methodology called 4D-molecular similarity [MS] analysis,[24] which is based upon the 4D-QSAR paradigm pioneered in our laboratory.[25] Each “finger” of the 4D-fingerprints corresponds to a particular atom/pharmacophore pair type in a molecule. Moreover, the 4D-fingerprints not only capture conformational ensemble, molecular size and chemical structure information of the molecule, but can be determined independent of molecular alignment. In a previous study, it was shown for five ligand-receptor applications that the 4D-fingerprints can be used to build statistically robust QSAR models as good, or better, than 3D-QSAR models from several currently popular QSAR methods.[26] However, in the case of skin sensitization, allergens need to penetrate skin, react with proteins by covalently binding to form hapten-protein complexes and, be recognized by T-cell receptors via non-covalent binding.[27] This is a more challenging application of 4D-fingerprints than non-covalent ligand-receptor QSAR modeling.

Nevertheless, this study applies 4D-fingerprints to build skin sensitization models. In the initial modeling effort discriminant analysis (DA)[18] and partial least square discriminant analysis (PLS-DA)[28] was used to generate categorical QSAR models for two states: sensitizer and non-sensitizer. However, these methods yielded models with values of prediction accuracy for the training and test sets of less than 75% and 67%, respectively. These results suggested that the application of alternate chemometric methods would be needed to obtain QSAR models with acceptably high predictivity.

Hence, two statistical techniques, namely logistic regression (LR) and partial least square coupled logistic regression (PLS-LR)[29] have been applied to generate categorical QSAR models for two states: sensitizer and non-sensitizer. A comparison of LR and PLS-LR methods has also been carried out as part of this study based on the models that have been constructed and employed as virtual screens to a test set.

Material and Methodology

Database

The LLNA training and test sets used in this study are from a master skin sensitization database constructed from data contributions made by a set of interested organizations.[30] Metal ions and mixtures in the master database were excluded in selecting compounds for the study reported here. This left a pruned database of 219 compounds with EC3 values, which are correspondingly categorized as non-, weak-, moderate-, strong- and extreme-skin sensitizers, as described in Table 1. All 3-dimensional (3D) structures of the compounds in the database were built using the Chemlab-II molecular modeling package.[31]

Compounds exhibiting the greatest difference in skin sensitization potency should be expected to have the greatest corresponding difference in their respective values for the key descriptors/properties responsible for this biological endpoint. Thus, to determine whether or not the 4D-fingerprints of a compound have the potential to "explain" its sensitization potency, the pruned database was reclassified into three categories: 101 "non-weak" sensitizers, 72 "moderate" sensitizers and 46 "strong-extreme" sensitizers. The "moderate" sensitizers were then eliminated to create a "separation gap", and the "non-weak" and "strong-extreme" sensitizers were retained to build categorical QSAR models for a dichotomous dependent variable, having a value of "0" for a "non-weak" sensitizer, and a value of "1" for a "strong-extreme" sensitizer. If the significance of 4D-fingerprints as descriptors for skin sensitization can be established by the fit of such dichotomous data to predictive categorical QSAR models, then the construction of models which relate these descriptors to multicategorical data or continuous EC3 values merits exploration in further studies.

Ten "non-weak" and five "strong-extreme" sensitizers were randomly chosen to be the test set for checking the predictive capability of the categorical QSAR models. The remaining 91 "non-weak" and 41 "strong-extreme" sensitizers formed the training set used to build the dichotomous QSAR models. The training set compounds are given as part of Table 2, and the test set is listed as part of Table 3.

Universal 4D-fingerprints

The theory and methodology of the universal 4D-fingerprints has been presented in detail in a previous publication,[26] and, therefore, is only summarized here. The universal 4D-fingerprints are the eigenvalues of the molecular similarity eigenvectors determined for a molecule from its set of absolute molecular similarity main distance-dependant matrices (MDDM). The eigenvectors capture the molecular information of a molecule regarding composition of atom types, size, shape and conformational flexibility. The types of atoms composing a molecule are currently defined as eight interaction pharmacophore elements (IPE's), whose individual definitions, in turn, are given in Table 4. A unique MDDM is constructed for each of the eight distinct and identical IPE pairs. The elements of the MDDM are derived by inductive derivation to be:

$$E(v,d_{ij})=e^{(-v<d_{ij}>)} \quad (1)$$

The constant v in eq. 1, which is set to 0.25, has been selected such that the difference in the sum of eigenvalues for any two arbitrary compounds with the same number, n, of a particular IPE type, m, is maximized. The term $<d_{ij}>$ is the Boltzmann conformational average distance between the atom pair ij of IPE types u and v,

$$<d_{ij}>=\sum_k d_{ij}(k)\, p(k) \quad (2)$$

p(k) in eq. 2 is the thermodynamic probability of conformer state k, and is computed from the ensemble of conformational energies determined for the molecule being studied. $d_{ij}(k)$ is the distance between atom pair i and j of IPE u and v, respectively, for the kth conformer state.

Diagonalization of the MDDM yields its eigenvector and constituent eigenvalues. If the members of an IPE pair are the same, i.e. u=v, MDDM is a square upper/lower triangular matrix and can be directly diagonalized. The resulting eigenvalues of the diagonalization are normalized, ranked in numerically descending order and represented as an eigenvector. The n^{th} normalized eigenvalue for IPE type m of a compound α, $\varepsilon_{mn}(\alpha)$, can be obtained by scaling the non-normalized eigenvalue $\varepsilon_{mn}'(\alpha)$ relative to the rank of its MDDM,

$$\varepsilon_{mn}(\alpha)=\frac{\varepsilon_{mn}'(\alpha)}{\mathrm{rank}(\alpha)_m} \tag{3}$$

If the members of an IPE pair are not the same, i.e. u≠v, the number of u and v IPE elements can be different ($n_u \neq n_v$). MDDM will be rectangular in this case, but the following two square MDDMs can be constructed:

$$\mathrm{MDDM}(u,u)=\mathrm{MDDM}(n_u n_v) * \mathrm{MDDM}(n_u,n_v)^T \tag{4}$$

$$\mathrm{MDDM}(v,v)=\mathrm{MDDM}(n_v n_u) * \mathrm{MDDM}(n_v,n_u)^T \tag{5}$$

MDDM(u,u) and MDDM(v,v) have the same set of eigenvalues since they have the same rank and trace. As a result, for each IPE pair u≠v,

$$\varepsilon(\alpha)_{u,v}=\left\{[\varepsilon(\alpha)]_{\mathrm{MDDM}(u,u)}\right\}^{\frac{1}{2}} \tag{6}$$

Each IPE pair corresponds to one MDDM from which one molecular similarity eigenvector can be formed. Since there are totally 36 distinct combinations of the currently defined eight IPE types, 36 eigenvectors, that is measures of molecular similarity, can be obtained for each molecule α. Dissimilarity between molecules α and β is given by

$$D_{\alpha\beta}=\sum_i |\varepsilon(\alpha)_i - \varepsilon(\beta)_i| \tag{7}$$

Where i refers to the i^{th} eigenvalue in the corresponding eigenvetor of a specific IPE pair. Molecular similarity is then defined as

$$S_{\alpha\beta}=(1 - D_{\alpha\beta})(1 - \varphi) \tag{8}$$

Where φ=|rank (α)-rank (β)|/(rank(α)+rank(β)). Since the rank of a MDDM matrix is the number of atoms of the specific IPE type present, φ serves to reincorporate molecular size information into $S_{\alpha\beta}$. The normalized eigenvalues limit the range of both $D_{\alpha\beta}$ and $S_{\alpha\beta}$ to be between 0 and 1. A $D_{\alpha\beta}$ value closer to 1 means a higher degree of dissimilarity while a $S_{\alpha\beta}$ value closer to 1 means higher molecular similarity.

The universal 4D-fingerprint descriptor set for molecule α is composed of all the eigenvalues of all the eigenvectors derived from all the MDDM for α. Operationally, a threshold cutoff value 0.002 is applied so that normalized eigenvalues less than the threshold value are disregarded.

Construction of the trial descriptor matrix for all of the training set compounds is based upon maximizing its information content. For each compound in the training set, the number of significant eigenvalues in the eigenvector for a particular IPE pair (u, v) is first computed. Then the maximum number of significant eigenvalues, $n_{max}(u,v)$, across the training set is determined. Finally, all the molecules of the training set are assigned $n_{max}(u,v)$ eigenvalues from their corresponding eigenvectors for the IPE pair (u,v). Eigenvectors that contain less than $n_{max}(u,v)$ significant eigenvalues have these "missing" eigenvalues set to zero. For instance, if $n_{max}(3,5)$ is 10, and the eigenvector for IPE pair (u,v) of compound α has only eight significant eigenvalues, the ninth and tenth eigenvalues for IPE pair(u,v) of α are set to zero.

The total number of universal descriptors, n_{total}, for each compound in the training set will be the sum of the $n_{max}(u,v)$ values for the 36 eigenvectors which is 720 for the training set in this study. The entire descriptor matrix is referred to as UMAX. A descriptor $\varepsilon_i(u,v)$ in UMAX represents the i^{th} eigenvalue in the eigenvector for the IPE pair (u,v). The method of creating UMAX introduces some degree of "noise" with the need to add zero eigenvalues. However, if the "noise" present in a particular descriptor column results in the descriptor being unfit to aid in describing the variance of the dependant categorical variable, this descriptor will not be present in the optimized QSAR model.

Partial lease square (PLS) regression was employed to reduce collinearity and multilinearity between the descriptors in UMAX. As discussed previously[32], the quality of a model produced in PLS-type regression is highly dependent on the manner in which the training set data is preprocessed. The relative importance of descriptors can be determined from their regression coefficients based on a common mean and variance. Hence, the auto-scaled form of UMAX, called USMAX, was also constructed and used in this study. Here, auto-scaling involves calibrating the column data to zero mean and unit variance by dividing each column by its respective standard deviation.

For each test set compound the number of eigenvalues in an eigenvector for a specific IPE type (u,v) must be the same as that of a training set compound, which is $n_{max}(u,v)$. If the number of eigenvalues of a test compound exceeds $n_{max}(u,v)$, the extra eigenvalues are discarded. If the number is less than $n_{max}(u,v)$, the "missing" eigenvalues are set to zero. In addition, after a categorical QSAR model is constructed from the auto-scaled descriptor matrix USMAX of the training set, the descriptor values of each test compound also need to be auto-scaled in the same manner as the corresponding descriptor values of the training set. For example, a descriptor ε3(1,3) is the third eigenvalue in the eigenvector of the IPE pair (1,3) and this descriptor corresponds to one column in UMAX. If the ε3(1,3) values of the training set compounds for this column have a mean of 0.3 and a variance of 0.64, the corresponding auto-scaled ε3(1,3) value of a test compound is obtained as $\left[\frac{\varepsilon 3(1,3) - 0.3}{0.8}\right]$.

Logistic Regression (LR)

For binary response models, the response, Y, of an individual, or and experimental endpoint, can take on one of two possible values, which are denoted for convenience by 1 and 0. For example, Y=1 if a compound is a strong or extreme skin sensitizer and, Y=0 if the compound is a non- or weak- sensitizer. Assuming x is a vector of explanatory independent variables, and P=Pr(Y=1|x) is the response probability to be modeled, then the linear logistic model has the form,

$$\text{Logit}(P)=\log\frac{P}{1-P}=\delta+x'\rho \tag{9}$$

$$\text{or } P=\frac{\exp(\delta+x'\rho)}{1+\exp(\delta+x'\rho)} \tag{10}$$

δ is the intercept parameter, and ρ is the vector of slope parameters in eqs. (9) and (10). The linear logistic regression models are fit to binary response data by the method of maximum likelihood. The maximum likelihood estimation is carried out using either Fisher-scoring or a Newton-Raphson algorithm. The predicted response probabilities are obtained by replacing the ρ parameter with its maximum likelihood estimate (MLE), $\hat{\rho}$.

Partial Least Square (PLS)

PLS extracts one factor (component) at a time in the building of a model. If $X=X_0$ is the (mean-) centered and (auto-) scaled matrix of predictors, and $Y=Y_0$ the (mean-) centered and (auto-) scaled matrix of response values, then the PLS method starts with a linear combination, $t=X_0w$, of the predictors, where t is called a score vector (extracted factor) and w is its associated weight vector. The PLS algorithm predicts both X_0 and Y_0 by regression on t:

$$\widehat{X_0}=tp', \text{ where } p'=(t't)-t'X_0 \tag{11}$$

$$\widehat{Y_0}=tc', \text{ where } c'=(t't)-t'X_0 \tag{12}$$

The vectors p and c are called the X- and Y-loadings, respectively.

The specific linear combination, $t=X_0w$, has a maximum covariance t′u with some response linear combination $u=Y_0q$. Another property of PLS is that the X- and Y-weights w and q, respectively, are proportional to the first left and right singular vectors of the covariance matrix, $X_0'Y$, or, equivalently, the first eigenvectors of the $X_0'Y_0Y_0'X_0$ and the $Y_0'X_0X_0'Y_0$ matrices, respectively.

These relationships define how the first PLS factor is extracted. The second factor is extracted in the same way by replacing X_0 and Y_0 with the X- and Y-residuals from the first factor,

$$X_1=X_0-\widehat{X_0} \tag{13}$$

$$Y_1=Y_0-\widehat{Y_0} \tag{14}$$

These residuals are also called the deflated X and Y blocks. The process of extracting a score vector, and then deflating the data matrices is repeated for as many extracted factors as are desired.

Partial Least Square-Logistic Regression (PLS-LR)

In logistic regression the maximum likelihood estimated parameter, $\hat{\rho}$, is approximately equals to $(X'WY)(X'WX)^{-1}$, where Y is the vector of responsible variables, X is the matrix of explanatory variables and W is a weighting matrix changing with $\hat{\rho}$ in each iteration to achieve convergence. Collinearity in X can cause $(X'WX)^{-1}$ to approach zero and $\hat{\rho}$ then becomes very unstable.

To diminish collinearity, and to reduce the dimensionality of the explanatory variables in the matrix of predictors, X, PLS is first used to acquire the sequential extracted factors (score vectors ts') termed as xscr1, xscr2, xscr3, etc. The linear logistic model is then built for the binary response variable Y, and the extracted factors from the PLS.

Building, goodness of fit and predictivity of a model

Whether to select variables from the initial descriptor set before logistic regression according to the univariable Wald statistic, or to two-sample t-test at the univariable level, has been a disputed issue for years.[33] Univariable analysis of each variable can exclude statistically insignificant variables. However, the possibility that a collection of variables, each of which is weakly associated with the outcome, can become an important predictor of outcome, when taken together, or combined with other existing variables in the model, is ignored. As a result, the logistic regressions in this study were first carried out for the whole set of auto-scaled universal descriptors, USMAX. Next, a single-descriptor logistic regression model was constructed, and the p-value of the univariable Wald statistic was evaluated for each descriptor in the USMAX. Descriptors with p-values less than 0.25 were selected to form an auto-scaled universal 4D-fingerprint matrix called the VS_USMAX. Subsequently, a comparison of the QSAR models based on USMAX and on VS_USMAX was performed.

A stepwise procedure was used to build multivariable categorical QSAR models with an appropriate number of significant descriptors. To obtain models with less numbers of descriptors for comparison, a backward procedure was subsequently applied.

Two goodness-of-fit tests can be used to evaluate a categorical model. These are the model deviance and Hosmer-Lemeshow tests. Model deviance is defined as the log of the likelihood ratio of the actual fitted model to the saturated model. The saturated model has the number of parameters equal to the number of observations under the circumstance that the explanatory variables are continuous. This statistic follows an asymptotic chi-square distribution when subpopulations of both Y=0 and Y=1, at any observed covariate pattern of explanatory variables X, are sufficiently large, e.g., larger than five. However, these subpopulations will be either 0 or 1 when the explanatory variables are continuous, and the resulting model deviance will follow an unknown distribution and make no sense. Therefore, the Hosmer-Lemeshow test was used in this study since all the 4D-fingerprints are continuous variables. This test involves dividing the observations into g groups of approximately the same size and with similar estimated probabilities within one group. The Hosmer-Lemeshow goodness-of-fit statistic χ^2_{HL} is defined as,

$$\chi^2_{HL}=\sum_{j=1}^{g}\sum_{k=0}^{1}\frac{(O_{jk}-E_{jk})^2}{E_{jk}} \tag{15}$$

where O_{jk} and E_{jk} are the observed and estimated number of subjects in the j$^{\underline{th}}$ group with Y=0 or Y=1, respectively. The distribution of χ^2_{HL} is well approximated by the chi-square distribution with [g−2] degrees of freedom when the total number of observations is large. Large values of χ^2_{HL} (and small p-values) indicate a lack of fit of the model.

Since goodness-of-fit depends on the number of descriptors in a model, a comparison of goodness-of-fit between models having different numbers of descriptors for the same data should take this difference into consideration. This scaling is accomplished by employing the “penalized log likelihood” Akaike information criterion (AIC) statistic.[34]

$$\mathrm{AIC}=-2Log_eL+2(p+1) \tag{16}$$

where L is the estimated likelihood, and p denotes the number of explanatory descriptors in a fitted model. Lower values of the AIC statistic indicate a more desirable model.

Model predictivity was evaluated using both leave-one-out cross-validation and a test set of compounds. The higher the accuracy of classification for the training set, based upon the leave-one-out measure, and for the test set, the better the predictivity of a model. Classification accuracy, sensitivity and specificity are defined as:

$$\text{Accuracy} = \frac{(tp+tn)}{tp+fn+tn+fp} \tag{17}$$

$$\text{Sensitivity} = \frac{tp}{tp+fn} \tag{18}$$

$$\text{Specificity} = \frac{tn}{tn+fp} \tag{19}$$

where in this study for a certain cutoff value c, tp and fn are the numbers of strong-extreme sensitizers for which the predicted probabilities are larger than, and less than, c, respectively. In like fashion, tn and fp are the numbers of non-weak sensitizers for which the predicted probabilities are less than, and larger than, c, respectively.

It should be noted that classification accuracy and goodness-of-fit have no direct interrelationship. Accurate, or inaccurate, classification does not indicate goodness-of-fit, or vice versa. However, use of classification accuracy is most appropriate when classification is a stated goal of the analysis, which is the case in this study.[33]

The statistical methodology reported above can be implemented using a combination of standard statistical methods available in SAS.[35]

Results

Logistic Regression (LR)

Statistical features of the LR QSAR models—The best QSAR model derived by application of logistic regression to the data is first presented, and then followed by an explanation why it is deemed “best” based on comparisons to other models.

The best model was obtained by applying stepwise logistic regression on both the non-scaled universal descriptor matrix (UMAX) and the auto-scaled universal descriptor matrix (USMAX) each of which are composed of 720 4D-fingerprints and 132 training set compounds. The significance level for a descriptor to enter, or leave, a model was set at a default value 0.05. Two QSAR models composed of nine 4D-fingerprints were built for the UMAX, eq. (20), and the USMAX, eq. (21), respectively:

$$\begin{aligned}\text{Logit}(P) = 4.693 &+ 61.70\varepsilon 3\,(\text{hbd,aro}) + 754.1\varepsilon 12\,(\text{hs,aro}) - 231.6\varepsilon 18\,(\text{all,all}) \\ &+ 22.33\varepsilon 2\,(\text{np,p}-) - 47.99\varepsilon 2\,(\text{all,all}) - 2.092\varepsilon 1\,(\text{p} - ,\text{hbd}) \\ &- 1014\varepsilon 19\,(\text{hs,all}) + 1753\varepsilon 54\,(\text{all,all}) + 803.1\varepsilon 39\,(\text{all,all})\end{aligned} \tag{20}$$

$$\begin{aligned}\text{Logit}(P) = -1.745 &+ 1.119\varepsilon 3\,(\text{hbd,aro}) + 2.686\varepsilon 12\,(\text{hs,aro}) - 0.9722\varepsilon 18\,(\text{all,all}) \\ &+ 1.654\varepsilon 2\,(\text{np,p}-) - 0.9936\varepsilon 2\,(\text{all,all}) - 0.8841\varepsilon 1\,(\text{p} - ,\text{hbd}) \\ &- 3.966\varepsilon 19\,(\text{hs,all}) + 1.466\varepsilon 54\,(\text{all,all}) + 1.388\varepsilon 39\,(\text{all,all})\end{aligned} \tag{21}$$

Where, for example, ε3(hbd,aro) represents the third largest eigenvalue from the MDDM of u=(hbd) and v=(aro). The eigenvector of (hbd, aro) contains the 4D-fingerprints (eigenvalues) from which the relative molecular similarity of any pair of skin-sensitizer(s), and/or non-skin-sensitizer(s), with respect to hydrogen bond donor and aromatic groups can be extracted. P

denotes the probability of a compound being a strong-extreme sensitizer and Logit(P) is defined by eq. (9). Descriptors in these models can be interpreted. For example, using the model from USMAX, a one unit increase of the auto-scaled 4D-fingerprint ε3(hbd,aro) descriptor of a compound can lead to a 1.119 unit increase in the likelihood [logit(P)] of this compound being a strong-extreme skin sensitizer as opposed to a non-weak sensitizer.

The Hosmer-Lemeshow goodness-of-fit statistic, χ^2_{HL}, for each of the two models is 3.634, and the p-values are both 0.889, which strongly support the null hypothesis that these models fit the training set data very well. Standard errors, Wald chi-square statistics and p-values for the descriptor terms in these two models are listed in Table 3. The p-values are all less than 0.05, which corresponds to rejecting the null hypothesis that the regression coefficient of each 4D-fingerprint descriptor term is zero. In other words, all the 4D-fingerprint terms in the models are significant. The estimates and standard errors of the regression coefficients for a corresponding 4D-fingerprint term in these two models are different. Actually, the models for UMAX and USMAX are exactly the same except for the non-scaled versus auto-scaled descriptors. For the non-scaled descriptor εi(u,v) in model (20), a large "i" value means this eigenvalue is in the middle or toward the end of a eigenvector for IPE type (u,v). Due to the eigenvector normalization process, the larger this "i" value, the smaller is both this eigenvalue and its variance across the training set. Consequently, such eigenvalues will have larger regression coefficients and the corresponding standard errors. The auto-scaled descriptors all have zero means and unit variances. Hence, their regression coefficients, and corresponding standard errors, are all of comparable values.

The predicted probabilities and classifications for both the training and test sets are shown in Tables 2 and 3, respectively, based upon eqs (20) and (21). A summary of classification accuracy, sensitivity and specificity for the training set, based on leave-one-out cross-validation and test set predictions, are given in Table 6. Since the results for UMAX and USMAX are exactly the same, these measures are listed only for USMAX.

Two different cutoff values for classification have been used. A cutoff of 0.57 for the training set has the highest classification accuracy, 85.6%. The sensitivity and specificity measures for the models of the training set are 68.3% and 93.4%, respectively. The reason for the higher specificity than sensitivity is that there are more non-weak sensitizers than strong-extreme sensitizers in the training set, and the corresponding cutoff is larger than 0.50. Classification accuracy for cross-validation of the training set is 78.0%. The test set for the same cutoff value of 0.57 has classification accuracy, sensitivity and specificity values of 86.7%, 60.0% and 100.0%, respectively.

The classification accuracy of the test set is actually higher than that of the training set, which indicates the model fits the data from the test set somewhat better than it does the training set. Although this behavior is not usually the case for most QSAR models, it can occur if the chemical structures and functional groups comprising the test set compounds are similar to those of the training set. Although the test set was chosen randomly from the parent dataset, the large chemical diversity of the overall dataset leads to a high probability that most test compounds have either analogues in the training set, or have relatively similar structures to some compounds in training set. The QSAR model may fit the analogues, or compounds with similar structures, better than other compounds in the training set, and, as a result, this model fits the test set better than the training set. For example, in the training set 1-bromooctadecane and methyl salicylate are both non-weak sensitizers. Their predicted probabilities of being strong-extreme sensitizers are very low based upon LR QSAR model, eq. (21), which is indicated by the predicted values of 0.023 and 0.015, respectively, in Table 2. In the test set, 1-bromododecane is the analogue of 1-bromooctadecane, and methyl 4-hydroxybenzoate shares similar structural features to methyl salicylate. Thus, the predicted probabilities of these

test compounds being strong-extreme sensitizers are also very low using the model, as demonstrated by the predicted values of 0.004 and 0.032, respectively, in Table 3. These predicted probabilities are much closer to zero than those values of some non-weak sensitizers in the training set, which might account for the better fit of the test set by the LR QSAR model, eq. (21), than the corresponding fit of the training set.

Two types of errors can be made as part of a two-state classification prediction. One error, in this application, is that a compound is predicted as a non-weak sensitizer but it is actually a strong-extreme sensitizer. Another is that a compound is predicted as a strong-extreme sensitizer but it's actually a non-weak sensitizer. In some situations a balanced error rate between these two types of errors is more desirable than a high predicative accuracy. The specificity and sensitivity of a model are enhanced by a balanced error rate. For a cutoff value of 0.32, balanced error rates for the training set in Table 2 can be obtained. Classification accuracy for this cutoff value is 79.5%, which is lower than the highest accuracy that can be achieved using a cutoff of 0.57. However, sensitivity and specificity for the training set are 78.0% and 80.2%, respectively, which are closer to each other than those respective values using a cutoff value of 0.57. The classification accuracy of cross-validation is 77.3%. The test set classification accuracy is 80.0%, which is slightly higher than that of the training set. However, test set sensitivity and specificity are 60.0% and 90.0%, respectively, which are more balanced than those respective measures using a cutoff 0.57. In addition, when a cutoff value is changed, the accuracy, sensitivity and specificity values usually change significantly for the data set of this study. That is, classification using a QSAR model from logistic regression for this dataset is very sensitive to the cutoff value. This behavior is due to more than one third of the predicted probabilities of being a strong-extreme sensitizer lying in the range of 0.100 to 0.900 as seen in Tables 2 and 3.

Determination of the “best” LR QSAR model—Before concluding the model given by eq. (21) is the “best” categorical QSAR model, it is necessary to compare this model to other models with different numbers of descriptors and their corresponding goodness-of-fit and predictive capabilities. Removal of 4D-fingerprints one by one from eq. (21), using the backward procedure along with increasing the significance level to retain a descriptor in a model, yields models with less descriptors than eq. (21). Using this procedure, ε39(all,all), ε19 (hs,all), ε54(all,all) are successively removed from eq. (21) leading to models with 8, 7 and 6 descriptors. The AIC values for models with 9, 8, 7, and 6 descriptors are 109.3, 112.6, 117.6 and 122.3, respectively, while the corresponding highest cross-validation accuracy values of each of these models are 81.1%, 78.0%, 78.0% and 77.3%, respectively, as reported in Table 7. Note that the model with 9 descriptors, eq. (21), has the lowest AIC value and highest cross-validation accuracy indicating it better fits the data, and performs more accurate classification, than the models with lesser numbers of descriptors.

A model containing additional descriptors and derived from eq. (21), using a forward procedure and increasing the significance level for the entry of a descriptor into a model, leads to a model with ε5(p+, aro) as an additional descriptor. This descriptor has the highest scoring chi-square statistic among descriptors not already in eq. (21). However, the p-value for the Wald chi-square statistic of this descriptor is 0.9833, an extremely high value. Hence, the null hypothesis holds that the regression coefficient of this descriptor is effectively zero due to its insignificance, or its high correlation to more significant descriptors already in the model. Overall, a more complex model than eq. (21) cannot be readily built due to the lack of significance of the descriptors in the remaining descriptor pool. Thus, variable selection must be done using univariable analysis, before logistic regression is carried out, in order to identify the set of significant descriptors.

VS_USMAX is formed by picking out 266 significant descriptors from the set of 720 descriptors in USMAX. A LR QSAR model with ten 4D-fingerprint terms is obtained using the stepwise procedure on VS_USMAX:

$$\begin{aligned}\text{Logit}(P) = &-1.576 + 2.877\varepsilon 3\,(\text{hbd,aro}) + 1.387\varepsilon 12\,(\text{hs,aro}) - 4.911\varepsilon 18\,(\text{all,all}) \\ &+ 1.017\varepsilon 2\,(\text{np,p}-) - 0.7601\varepsilon 2\,(\text{all,all}) - 0.9919\varepsilon 12\,(\text{hs,hs}) \\ &+ 4.589\varepsilon 15\,(\text{np,np}) - 1.890\varepsilon 2\,(\text{np,hbd}) + 0.8954\varepsilon 2\,(\text{hs,p}+) \\ &+ 0.6735\varepsilon 3\,(\text{np,hba})\end{aligned} \quad (22)$$

Although the cross-validation accuracy of eq. (22) is 81.8%, its Hosmer-Lemeshow goodness-of-fit statistic χ^2_{HL} is 24.003, and the corresponding p-value is 0.002 indicating this model doesn't fit the data very well. Also, the AIC for this model is 117.3, which is higher than that of eq. (21) developed for USMAX. However, since the lack of goodness-of-fit does not necessarily imply poor classification accuracy, and correct classification is one the goals of this study, the predicted probabilities and classification information using eq. (22) are given in Tables 2 and 3. The summary features of eq. (22) are shown in Table 6. The cutoff value for highest classification accuracy of the training set for eq. (22) is 0.43, and that for balanced error rates is 0.31. An inspection of Table 6 indicates that all corresponding classification accuracies, sensitivities and specificities for eq. (22) are comparable to those of eq. (21). Interestingly, the error rates of the test set are more balanced using eq. (22) than eq. (21). The test set sensitivity and specificity for eq. (22) are both 80.0% for a cutoff value 0.31. However, although the classification accuracies are similar, eq. (21) is still better than eq. (22) because it better fits the training set according to the criteria of goodness-of-fit and, more importantly, it has less descriptors.

The comparison study of eq. (21) to eq. (22) also supports the argument that performing variable selection analysis may not always be appropriate. Such an analysis ignores the information contained by those descriptors that are insignificant in a single-descriptor logistic regression model, but significant when they enter into a multivariable model as a part of a unique combination with other descriptors.

The fully comprehensive strategy to choose the “best” LR model is to evaluate all possible models. Such an evaluation would include all possible model sizes composed of all combinations and interactions of descriptors from a trial set. Identification of the best LR model would be made in terms of the individual goodness-of-fit statistic and the predictivity of each model. However, it is impractical to do this exhausting search owing to practical time limits and computational effort. Equation (21) has both the best goodness-of-fit, as shown by its low AIC value, and the highest predictive power among all the LR-QSAR models identified using stepwise, backward and forward procedures, and variable selection. Thus, the model given by eq. (21) is at least one of the “best” LR QSAR models, and its descriptors present significant molecular structure-activity information related to skin sensitization. Overall, eq. (21) has a meaningful chance of correctly predicting a molecule being a strong-extreme, or a non-weak, skin sensitizer.

Interpretation of the 4D-fingerprints in the “best” LR-QSAR model—The 4D-fingerprints contain information regarding the conformaitonal flexibility, molecular shape, size, bonding topology and inherent 3D-pharmacophores of a molecule.[31] The high quality of the LR QSAR model given by eq. (21) using 4D-fingerprints to model skin-sensitization data suggests these descriptors capture information about the mechanism of skin sensitization.

Before trying to understand what particular molecular information can be gleaned from the 4D-fingerprints of the “best” LR-QSAR model, and how this information might relate to the process of skin sensitization, it is necessary to determine the most important descriptors in the

model, eq. (21). The “first five” descriptors, which are ε3(hbd,aro), ε12(hs,aro), ε18(all,all), ε2(np,p-) and ε2(all,all), are exactly the same for both eq. (21) and eq. (22). Equation (22) is constructed using VS_UMAX with significant univariable selection. Hence, it can be concluded that these five descriptors are significant in univariable LR QSAR models since the p-values for their Wald chi-square statistics are all less than 0.25. Furthermore, as can be seen in Table 3, the Wald chi-square statistics of these five descriptors are among the highest six values indicating they are also significant in a multivariable LR QSAR model. Consequently, these five 4D-fingerprints can be taken as the most significant descriptors in the entire descriptor pool. The other four descriptors in eq. (21), which are ε1(p-,hbd), ε19(hs,all), ε54 (all,all) and ε39(all,all), are not significant as single variables to explain the variance of the logit. However, these descriptors become significant in the multivariable LR QSAR model, eq. (21), as shown by their large Wald chi-square statistics, because they significantly impact logit as a descriptor combination with up to several of the five most significant descriptors. Thus, these four 4D-fingerprints can be considered as correction and refinement terms to the five significant descriptors of eq. (21).

Skin sensitization is a complicated process that very likely involves several steps necessary to occur. First, a potential sensitizer, the antigen, needs to be absorbed into and penetrate through the skin. Among the five most significant descriptors, ε2 (np,p-), ε3(hbd,aro) and ε12(hs,aro) all have positive regression coefficients suggesting non-polar and aromatic atoms of a molecule, which increase its lipophilicity, may promote absorption into the skin. Therefore, the corresponding probability of the molecule to be a strong-extreme skin sensitizer also increases.

Subsequent to skin absorption, and possibly after some form of modification due to specific skin metabolism activities, the resultant molecule is then recognized by Langerhans cells (LC) which are present in the epidermis and are responsible for initiating primary immune responses. This recognition process usually involves a covalent link to a specific amino acid on the surface of the LC. 4D-fingerprints ε2(np,p-) and ε3(hbd,aro) may capture the requisite reactivity specificity of a molecule since atom types p- and hbd are indicative of the polarity and hydrogen extraction behavior of a molecule. It is generally accepted that most skin sensitizers are electrophiles. Interestingly, the p+ atom type of electrophiles does not appear in the model, given by eq. (21), but rather its counterpart, the p- atom type, is found in eq. (21). Both negative polar and positive polar atoms exist in an electrophile, and it is the net polarity that most likely accounts for overall reactivity. Moreover, not all skin sensitizers are electrophiles since some need to be metabolized to be eletrophiles. Regardless, a sensitizer either covalently binds directly to peptide, or through a metabolized product. The resulting reactive adduct is a polar structure. The positive regression coefficients of np and p- descriptors in eq. (21) may indicate that an increase in a molecular polarity corresponds to an increase in reactivity, and the probability that the molecule may be a strong-extreme skin sensitizer. The ε1(p-,hbd) descriptor in eq. (21) may indicate a need for a particular spatial distribution p- and hbd atom types over a molecule in the process of skin sensitization. Moreover, this descriptor has the smallest regression coefficient in absolute value among all descriptors. This suggests ε1(p-,hbd) may be a correction term to the other significant descriptors in eq. (21).

Finally, the LCs migrate from the epidermis, via afferent lymphatics, to draining lymph nodes where they present the antigen bound at the binding cleft of a major histocompatibility protein (MHC) to responsive T lymphocytes[36]. Antigen-specific T lymphocytes are activated and are stimulated to divide and differentiate. Cell division results in the clonal expansion of allergen-responsive cells such that if the now-sensitized subject is subsequently exposed to the inducing allergen, then an accelerated and more aggressive secondary response will be provoked causing allergic contact dermatitis.

Few structural studies have been performed at the molecular level to investigate the precise contacts, or interactions, involved in TCR recognition of the MHC-peptide-hapten complex. However, since the immune response is mainly directed against the hapten, and shows little dependence on the structure of the peptide, a common proposal is that the TCR recognizes the hapten structure by requisite complementary geometries between them, which depend on molecular size, shape, and steric interactions.[37] The significant 4D-fingerprints ε2(all,all), ε18 (all,all), and the correction terms ε39(all,all), ε54(all,all), ε19(hs,all) may, in composite, capture the important role molecular shape, size and steric interactions play in the process of skin sensitization. Moreover, since the regression coefficients of ε2(all,all) and ε18(all,all) are negative, while those of ε39(all,all) and ε54(all,all) are positive, "medium" size molecule with 18 to 38 atoms may have the highest probability of being a non-weak sensitizer, and a "large" molecule with 54 or more atoms, having the highest probability of being a strong-extreme sensitizer.

Partial-Least-Square Coupled Logistic Regression (PLS-LR)

Statistical features of the PLS-LR QSAR models—UMAX and USMAX, rather than VS_USMAX, are used to construct the PLS-LR QSAR model. There are three reasons not to perform any variable selection process before the PLS-LR regression study. First, it is necessary to extract as much information as possible from the entire descriptor pool by PLS. Second, since PLS extracts components which account for the variances both in the explanatory variables as well as in the response variables, significant descriptors will possess much larger weights than insignificant descriptors in the first several important components. Third, it has been shown that that VS_USMAX does not fit the training set very well because it contains less information than USMAX.

PLS-LR QSAR models have been constructed using a stepwise procedure with the same entry, and significance levels of 0.05 as used in the previously constructed LR QSAR models.

The best PLS-LR QSAR model for UMAX is given by eq. (23),

$$\begin{aligned}\text{Logit}(P) = &-3.361+2.740\text{xscr2}+7.050\text{xscr3}+13.88\text{xscr5}+11.49\text{xscr7}\\ &+13.41\text{xscr9}+35.28\text{xscr12}+24.09\text{xscr14}+44.64\text{xscr18}\\ &+47.18\text{xscr20}+45.72\text{xscr24}\end{aligned} \quad (23)$$

In eq. (23), xscri denotes the i^{th} PLS component extracted from the 720 non-scaled 4D-fingerprints in UMAX. Although 132 components, which is the smaller value between the number of compounds and the number of descriptors, can be extracted by PLS, only the first several components are selected. The reason for this limited selection is due to the first several components accounting for most of the variances in both the explanatory variables and the response logit. The Hosmer-Lemeshow goodness-of-fit statistic, χ^2_{HL}, of the model given by eq. (23) is 1.006, and the corresponding p-value is 0.985, indicating this model fits UMAX very well. The AIC value for this model is 45.903, which is much less than that of the QSAR model given by eq. (21). Thus, eq. (23) fits the data better than eq. (21). Standard errors, Wald chi-square statistics and p-values for each of the regression coefficients are listed in Table 8 for eq. (23). All p-values are less than 0.05 indicating all the descriptor terms in eq. (23) are significant. The xscri are extracted by PLS in order of decreasing magnitude of explained variance. Thus, for the set of xscri, the smaller the "i", the larger the variance in xscri, and, correspondingly, the smaller the standard error of the corresponding regression coefficient.

The predicted probabilities and corresponding classification assignments of the training set and test set compounds are listed in Tables 10 and 11, respectively. A summary of the classification performance of eq. (23) is given in Table 12.

Using a cutoff value of 0.28, the training set has the highest classification accuracy of 97.0%, a sensitivity of 95.1%, specificity of 97.8% and cross-validation accuracy of 89.4%. All these performance measures are much higher than the corresponding values of eq (21). However, the classification accuracy for the test set is 73.3%, which is less than that of eq. (21), suggesting eq. (23) may be "overfit" to some extent. The training set has balanced error rates using eq. (23) for a cutoff value of 0.27.

The optimum PLS-LR QSAR model for USMAX is given by eq. (24),

$$\text{Logit}(P) = -3.564 + 0.498\text{xscr1} + 0.407\text{xscr4} + 0.636\text{xscr5} + 0.513\text{xscr8} + 0.415\text{xscr10} + 0.700\text{xscr13} + 0.882\text{xscr20} \quad (24)$$

In eq. (24), xscri again denotes the $i^{\underline{th}}$ PLS component extracted from 720 auto-scaled 4D-fingerprints in USMAX. χ^2_{HL} and the corresponding p-value of this model, eq. (24), are 0.906 and 0.996, respectively, which strongly supports the null hypothesis that eq. (24) fits the data in USMAX. The AIC value for eq. (24) is 58.202, which is smaller than that of eq. (21), indicating eq. (24) fits data better than eq. (21). Standard errors, Wald chi-square statistics and p-values for each of the regression coefficients are listed in Table 9, and, overall, indicate the significance of all seven descriptors in eq. (24). All the regression coefficients and their corresponding standard errors of fit are on the same scale because each PLS component has a zero mean and unit variance.

Table 10 contains the predicted probabilities and corresponding classifications of the training set using eq. (24). Table 11 contains the probabilities and classifications for the test set. A summary performance of eq. (24) is given in Table 12.

The training set has the highest classification accuracy of 93.2%, sensitivity of 85.4%, and specificity of 96.7% using a cutoff value of 0.51. Moreover, the cross-validation accuracy is 87.9%, and the test set has a classification accuracy of 80.0%, corresponding to a sensitivity of 60.0% and a specificity of 90.0%. These values of the performance measures for the training set, and cross-validation accuracy all decrease slightly when the cutoff changes to 0.36 in order to realize balanced error rates. The performance measures for the test set remain the same.

Model comparisons—Compared to the PLS-LR QSAR model, eq. (23) for UMAX, the model given by eq. (24) has lower classification accuracy, sensitivity, specificity and cross-validation accuracy for the training set. Thus, eq. (23) fits the data somewhat better, and has more predictive ability, than eq.(24). However, eq. (24) has a higher classification accuracy for the test set which suggests it's not an "overfit" model. And more important, eq. (24) has three less descriptors than eq. (23). Hence, eq. (24) is judged a better overall model than eq. (23). This conclusion is consistent with the argument that in most cases data auto-scaling is required to construct a high-quality QSAR model when using PLS.[38]

A comparison between the PLS-LR model, eq. (24), and the best logistic regression model, eq (21), is important to carry out. First, eq. (24) has a lower AIC value than eq. (21). Moreover, training set classification accuracy, sensitivity, specificity and cross-validation accuracy for eq. (24) are all higher than those respective measures for eq. (21). Thus, eq. (24) has a better goodness-of-fit and higher predictive power than eq. (21). Second, an inspection of Table 10 for eq. (24), reveals only two values of the 15 predicted probabilities for the test set compounds lie in the range of 0.100-0.900. Consequently, the test set classification accuracy remains at the same value of 80% when the cutoff changes from 0.51 to 0.36 for the probabilities values reported in Table 11. However, there are seven predicted probabilities that lie in the same range in Table 5 for eq. (21) which leads to a decrease in test set classification accuracy from 87.6% to 80.0% when the cutoff value changes from 0.57 to 0.32. This finding indicates that PLS-LR

gives more “stable” predicted probabilities, most of which are larger than 0.100, or less than 0.900, and the corresponding classification accuracies are less sensitive to selection of the cutoff value. Third, eq. (24) has less descriptors than eq. (21), which is always desirable in QSAR model. Hence, from a statistical point of view, the PLS-LR model, eq. (24), is superior to any other logistic regression model. Partial least square fitting diminishes collinearity and multilinearity by extracting orthogonal components, and a logistic regression model based on those orthogonal variables will have higher quality, and less components, than other logistic regression models using non-orthogonal variables.

The most important application of a PLS-LR QSAR model is to make predictions. However, the PLS components are difficult to interpret since they are linear combinations of all the original 4D-fingerprints in the descriptor pool. Hence, it is difficult make any conclusion or inference about the mechanism of action from a PLS model.

Discussion

The results of this study indicate that the eigenvalues, or 4D-fingerprints, derived from the main distance-dependant matrices (MDDM) from a 4D-MS absolute similarity analysis can be used to develop significant dichotomous logistic regression QSAR models for skin sensitization based on LLNA potency measures. The capability of distinguishing non-weak sensitizers from strong-extreme sensitizers shows the 4D-fingerprints of a molecule capture key molecular features responsible for its sensitization potency. Therefore, these descriptors may also be useful in constructing multicategorical, and even continuous, QSAR models for skin sensitization. Such studies will be undertaken in the near future.

The most appealing aspect of the 4D-fingerprints of a molecule is that they are a set of ‘universal’ descriptors, derived independent of any molecular alignment. Moreover, these “universal” QSAR descriptors embed thermodynamically relevant 3D and conformational information, and are independent of any external constraints. Thus, the 4D-fingerprints can be used in virtually any application, including ADMET property predictions, molecular similarity clustering, and developing toxicity profile libraries, to name only some possible applications.

A limitation, or trade-off, of the generality of the 4D-fingerprints is that they are somewhat abstract compared to many other descriptors that are used to develop QSAR models. Although the 4D-fingerprints are derived from the 3-dimensional structures of molecules, visualizing them in Euclidean space is not possible. These descriptors are latent variables, or principle components, of a distance matrix which permits their definition only in a mathematical sense, but not in physical space. Perhaps descriptors that are universal in applicability must necessarily be of a form which limits their physical representation and interpretation. As a result, it is difficult to ascertain what change of values in the 4D-fingerprints occurs when the structure of a molecule is modified. However, the 4D-fingerprints can always be applied in a QSAR analysis, and, more significantly, have led to significant models when other descriptors have failed including earlier studies on the database investigated in this study.

The two-state categorical skin sensitization models developed in this study can be used as limited virtual skin sensitization screens. The limitation is that, strictly speaking, only non- and weak- sensitizers can be differentiated from strong- and extreme- skin sensitizers using these models. There is no way of identifying/classifying moderate skin sensitizers. We did divide the moderate skin sensitizers whose potency EC3 values were available into two sub-populations, namely those having the lowest 50% potency [analogous to non and slight skin sensitization] and those having the highest 50% potency [corresponding to strong and extreme skin sensitization]. Then we used the model given by eq. 24 to predict the two-state category assignment of these compounds. Sixty-two percent of the moderate compounds were correctly

predicted, but very often the prediction probabilities were near the two-state cutoff values. While these findings are not too encouraging from a prediction accuracy point of view, they do support the expectation that compounds whose prediction probabilities are near the two-state-separation cutoff values would be expected to likely be either moderate sensitizers and/or have 4D-fingerprint descriptor values which make their predicted assignments uncertain. Thus, compounds virtually screened whose probabilities of prediction are near the cutoff values can be flagged as compounds most likely to be moderate sensitizers or outliers.

Acknowledgements

This work was funded by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 R21 GM075775-01. Information on Novel Preclinical Tools for Predictive ADME-Toxicology can be found at http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-04-023.html. Links to nine initiatives are found here http://nihroadmap.nih.gov/initiatives.asp. This work was also supported, in part, by The Procter & Gamble Company. Resources of the Laboratory of Molecular Modeling and Design at UIC and The Chem21 Group, Inc. were used in performing these studies.

References

1. Engelhard VH. How cells process antigens. Sci Am 1994;271:54–61. [PubMed: 8066431]
2. Kimber I, Basketter DA. The murine Local Lymph Node assay; collaborative studies and directions: A commentary. Food, Chem Toxicol 1992;30:165–169. [PubMed: 1555798]Basketter DA, et al. Use of the local lymph node assay for the estimation of relative contact allergy potency. Contact Dermatitis 2000;42:344–348. [PubMed: 10871098]
3. EC2003. 2003/15/EC. Commission Directive of 27 February **2003** amending Council Directive 76/768/EEC on the approximation of laws of the Meber States relating to cosmetic products, Off. J. Eur. Union L66, 26-35. http://www.europa.eu.int/comm/environment/chemicals/index.html
4. Landsteiner K, Jacobs J. Studies on the sensitization of animals with simple chemical compounds. J Exp Med 1936;64:625–639.
5. Roberts DW. Structure-activity relationships for skin sensitisation potential of diacrylates and dimethacrylates. Contact Dermatitis 1987;17:281–289. [PubMed: 3436133]
6. Roberts DW, Fraginals R, Lepoittevin JP, Benezra C. Refinement of the relative alkylation index (RAI) model for skin sensitization and application to mouse and guinea-pig test data for alkyl alkanesulphonates. Arch Dermatol Res 1991;283:387–394. [PubMed: 1665682]
7. Basketter DA, Roberts DW, Cronin M, Scholes EW. The value of the local lymph node assay in quantitative structure-activity investigations. Contact Dermatitis 1992;27:137–142. [PubMed: 1451456]
8. Roberts DW, Basketter DA. QSAR: Sulfonate esters in the LLNA. Contact Dermatitis 2000;42:154–161. [PubMed: 10727166]
9. Roberts DW. Structure-activity relationships for skin sensitization potential of diacrylates and dimethacrylates. Contact Dermatitis 1987;17:281–289. [PubMed: 3436133]
10. Roberts DW, York M, Basketter DA. Structure-activity relationships in the murine local lymph node assay for skin sensitization: α,β- diketones. Contact Dermatitis 1999;41:264–271. [PubMed: 10554060]
11. Patlewicz G, Basketter DA, Smith CK, Hotchkiss SAM, Roberts DA. Skin-sensitization structure-activity relationships for aldehydes. Contact Dermatitis 2001;44:331–336. [PubMed: 11380542]
12. Roberts DW, Williams DL. The derivation of quantitative correlations between skin sensitisation and physio-chemical parameters for alkylating agents, and their application to experimental data for sultones. J Theor Biol 1982;99:807–825. [PubMed: 6191155]
13. Roberts DW, Basketter DA. A quantitative structure activity/dose response relationship for contact allergic potential of alkyl group transfer agents. Contact Dermatitis 1990;23:331–335. [PubMed: 1965716]
14. Basketter DA, Roberts DW. A quantitative structure activity/dose relationship for contact allergic potential of alkyl group transfer agents. Toxicol In Vitro 1990;4:686–687.

15. Roberts DW, Benezra C. Quantitative structure-activity relationships for skin sensitization potential of urushiol analogues. Contact Dermatitis 1993;29:78–83. [PubMed: 8365181]
16. Cronin MT, Basketter DA. Multivariate QSAR analysis of a skin sensitization database. SAR QSAR Environ Res 1994;2:159–179. [PubMed: 8790644]
17. Magee PS, Hostynek JJ, Maibach HI. A classification model for allergic contact dermatitis. Quantitative Structure-Activity Relationship 1994;13:22–33.
18. Enslein K, Gombar VK, Blake BW, Maibach HI, Hostynek JJ, et al. A quantitative structure-toxicity relationships model for the dermal sensitization guinea pig maximization assay. Food Chem Toxicol 1997;35:1091–1098. [PubMed: 9463544]
19. Fedorowicz A, Zheng L, Singh H, Demchuk E. QSAR study of skin sensitization using local lymph node assay data. Int J Mol Sci 2004;5:56–66.
20. Magnusson B, Kligman AM. The identification of contact allergens by animal assay. The guinea pig maximization test. J Invest Dermatol 1969;52:268–276. [PubMed: 5774356]
21. Magnusson, B.; Kligman, AM. Allergic Contact Dermatitis in the Guinea Pig Identification of Contact Allergens. Charles A Thomas; Springfield, IL: 1970.
22. Basketter DA, Lea LJ, Cooper K, Stocks J, Dickens A, et al. Threshold for classification as a skin sensitizer in the local lymph node assay: a statistical evaluation. Food Chem Toxicol 1999;37:1167–1174. [PubMed: 10654593]
23. European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC). Technical Report No 87, Contact Sensitisation: Classification According to Potency. Brussels, Belgium: Apr. 2003
24. Duca JS, Hopfinger AJ. Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. J Chem Inf Comput Sci 2001;41:1367–1387. [PubMed: 11604039]
25. Hopfinger AJ, Wang S, Tokarski JS, Jin B, Albuquerque M, et al. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. J Am Chem Soc 1997;119:10509–10524.
26. Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ. 4D-fingerprints, universal QSAR and QSPR descriptors. J Chem Inf Comput Sci 2004;44:1526–1539. [PubMed: 15446810]
27. Martin S, Weltzien HU. T cell recognition of haptens, a molecular view. Int Arch Allergy Immunol 1994;104:10–16. [PubMed: 7524836]
28. Nouwen J, Lindgren F, Hansen B, Karcher W. Classification of Environmentally Occurring Chemicals Using Structural Fragments and PLS Discriminant Analysis. Environ Sci Technol 1997;31:2313–2318.
29. Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002;18:39–50. [PubMed: 11836210]
30. Gerberick GF, Ryan CA, Kern PS, Schlatter H, Dearman RJ, Kimber I, Patlewicz GY, Basketter DA. Compilation of historical local lymph node assay data for evaluation of skin sensitization alternative methods. Dermatitis 2005;16(4):157–202. [PubMed: 16536334]
31. Pearlstein, RA. CHEMLAB-II Users Guide. CHEMLAB Inc.; Lake Forest, IL: 1998.
32. Glen WG, Dunn WJ III, Scott DR. Principle component analysis and partial least squares regression. Tetrahedron Computer Methodology 1989;2:349–376.
33. Hosmer, DW.; Lemeshow, S. Applied Logistic Regression. second. John Wiley & Sons, Inc.; New York: 2000.
34. Agresti, A. Categorical Data Analysis. 2nd. Wiley-Interscience; New York: 2002.
35. Institute, S.. SAS/STAT User's Guide. Version 8. SAS Institute Inc.; Cary, NC: 1999.
36. Kimber I, Cumberbatch M, Dearman RJ, Bhushan M, Griffiths CE. Cytokines and chemokines in the initiation and regulation of epidermal Langerhans cell mobilization. British Journal of Dermatology 2000;142:410–41.
37. Lepoittevin, JP.; Basketter, DA.; Goossens, A.; T, KA. Allergic Contact Dermatitis: The Molecular Basis. Springer-Verlag; Berlin: 1998.
38. Wold, S.; Johansson, E.; Cocchi, M. 3D QSAR in Drug Design, Theory, Methods, and Applications. ESCOM Science Publishers; Leiden: 1993. PLS - Partial Least Squares Projections to Latent Structures.

Table 1

Numbers of compounds in each skin sensitization potency category for the local lymph node assay (LLNA) database, training set and test set.

Dataset	Skin sensitization potency				
	Non sensitizers	Weak sensitizers	Moderate sensitizer*	Strong sensitizers	Extreme sensitizers
LLNA database	41	60	72	32	14
Training set	37	54	——	29	12
Test set	4	6	——	3	2

* Moderate sensitizers have been left out of the analysis reported in this paper.

Table 2

Predicted probabilities and corresponding classifications of training set compounds using logistic regression QSAR models for the auto-scaled universal matrix (USMAX) and for the auto-scaled universal matrix after variable selection (VS_UMAX).

Chemical Name	Observed Class	USMAX			VS_USMAX		
		Predicted Probablity	Predicted Class (cutoff=0.57)	Predicted Class (cutoff=0.32)	Predicted Probablity	Predicted Class (cutoff=0.43)	Predicted Class (cutoff=0.31)
1-(2′,3′,4′5′-Tetramethylphenyl)-3-(4′-tertbutylphenyl) propane-1,3-dione	0	0.312	0	0	0.894	1	1
1-(2′,5′-Dimethylphenyl)butane-1,3-dione	0	0.569	0	1	0.182	0	0
1-(3′,4′,5′-Trimethoxyphenyl)-4-dimethylpentane-1,3,-dione	0	0.000	0	0	0.428	0	1
1-Bromobutane	0	0.148	0	0	0.174	0	0
1-Bromohexane	0	0.008	0	0	0.035	0	0
1-Bromononane	0	0.003	0	0	0.009	0	0
1-Bromooctadecane	0	0.023	0	0	0.002	0	0
1-Bromotridecane	0	0.067	0	0	0.001	0	0
1-Bromoundecane	0	0.003	0	0	0.001	0	0
1-Butanol	0	0.014	0	0	0.134	0	0
1-Chloromethylpyrene	1	1.000	1	1	0.927	1	1
1-Chlorononane	0	0.003	0	0	0.009	0	0
1-Chlorooctadecane	0	0.039	0	0	0.002	0	0
1-Chlorotetradecane	0	0.052	0	0	0.001	0	0
1-Iodododecane	0	0.003	0	0	0.003	0	0
1-Iodohexadecane	0	0.064	0	0	0.001	0	0
1-Iodononane	0	0.003	0	0	0.009	0	0
1-Iodooctadecane	0	0.056	0	0	0.002	0	0
1-Iodotetradecane	0	0.050	0	0	0.004	0	0
1-Methyl-3-nitro-1-nitrosoguanidine	1	0.317	0	0	0.825	1	1
1-Phenyl-2-methylbutane-1,3-dione	0	0.130	0	0	0.035	0	0
12-Bromododecanoic acid	0	0.047	0	0	0.208	0	0
1,2-Dibromo-2,4-dicyanobutane	1	0.665	1	1	0.510	1	1
2-Acetylcyclohexanenone	0	0.124	0	0	0.467	1	1
2-Aminophenol	1	0.985	1	1	1.000	1	1
2-Ethyl butaldehyde	0	0.034	0	0	0.134	0	0
2-Hydroxypropyl methacrylate	0	0.012	0	0	0.240	0	0
2-Mercaptobenzothiazole	1	0.238	0	0	0.885	1	1
2-Methyl-5-hydroxyethylaminophenol	1	0.749	1	1	0.996	1	1
2-Methylundecanal	0	0.002	0	0	0.001	0	0
2-Nitro-p-phenylenediamine	1	0.931	1	1	0.821	1	1
2,2,6,6-Tetramethyl-heptane-3,5-dione	0	0.080	0	0	0.018	0	0
2,3-Butanedione	0	0.314	0	0	0.245	0	0
2,4,6-Trichloro-1,3,5-triazine (cyanuric chloride)	1	0.928	1	1	0.932	1	1
3-Ethoxy-1-(2′,3′,4′,5′-tetramethylphenyl)propane-1,3-dione	0	0.209	0	0	0.072	0	0
3-Methyl catechol	1	0.313	0	0	0.194	0	0
3-Methyleugenol	0	0.071	0	0	0.304	0	0
3-Phenylenediamine	1	0.990	1	1	1.000	1	1
4-Allylanisole	0	0.007	0	0	0.034	0	0
4-Amino-m-cresol	1	0.973	1	1	0.952	1	1
4-Hydrobenzoic acid	0	0.443	0	1	0.212	0	0
4-Nitrobenzyl bromide	1	0.725	1	1	0.517	1	1
4-Nitroso-N,N-dimethylaniline	1	0.224	0	0	0.571	1	1
4,4,4-Trifluro-1-phenylbutane-1,3-dione	0	0.532	0	1	0.259	0	0
5-Chloro-2-methyl-4-isothiazolin-3-one	1	0.602	1	1	0.430	1	1
5-Methyl-2,3-hexanedione	0	0.113	0	0	0.142	0	0
5-Methyleugenol	0	0.090	0	0	0.230	0	0
5-Methylisoeugenol	1	0.135	0	0	0.661	1	1
5,5-Dimethyl-3-thiocyanatomethyl-2(3H)-furanone	1	0.279	0	0	0.151	0	0

Chemical Name	Observed Class	USMAX			VS_USMAX		
		Predicted Probablity	Predicted Class (cutoff=0.57)	Predicted Class (cutoff=0.32)	Predicted Probablity	Predicted Class (cutoff=0.43)	Predicted Class (cutoff=0.31)
6-Methylcoumarin	0	0.370	0	1	0.060	0	0
6-Methyleugenol	0	0.067	0	0	0.224	0	0
7-Bromotetradecane	0	0.062	0	0	0.003	0	0
7,12-Dimethylbenz[a]anthrace ne	1	0.961	1	1	0.434	1	1
a-Butyl cinnamic aldehyde	0	0.013	0	0	0.004	0	0
Abietic acid	0	0.000	0	0	0.167	0	0
Aniline	0	0.023	0	0	0.018	0	0
b-Propriolactone	1	0.714	1	1	0.578	1	1
Benzaldehyde	0	0.174	0	0	0.197	0	0
Benzo[a]pyrene	1	0.938	1	1	0.768	1	1
Benzocaine	0	0.027	0	0	0.007	0	0
Benzoyl peroxide	1	0.999	1	1	0.927	1	1
Benzoyl chloride	1	0.809	1	1	0.611	1	1
Benzyl benzoate	0	0.771	1	1	0.069	0	0
Bis-1,3-(2′,5′-dimethylphenyl)-propane-1,3-dione	0	0.061	0	0	0.553	1	1
Butyl glycidyl ether	0	0.114	0	0	0.126	0	0
C11-Azlactone	0	0.002	0	0	0.089	0	0
C15 Azlactone	0	0.209	0	0	0.053	0	0
C19-Azlactone	0	0.082	0	0	0.067	0	0
Camphorquinone	0	0.825	1	1	0.420	0	1
Chlorobenzene	0	0.338	0	1	0.326	0	1
Chlorpromazine hydrochloride	1	0.746	1	1	0.845	1	1
Cinnamic alcohol	0	0.001	0	0	0.022	0	0
cis-6-Nonenal	0	0.004	0	0	0.207	0	0
Coumarin	0	0.681	1	1	0.481	1	1
Diethylphthalate	0	0.225	0	0	0.241	0	0
Dimethyl sulfate	1	0.348	0	1	0.538	1	1
Dimethylsulfoxide	0	0.070	0	0	0.101	0	0
Diphenylcyclopropenone	1	0.777	1	1	0.470	1	1
Dodecylthiosulphonate	1	0.673	1	1	0.441	1	1
Ethyl acrylate	0	0.498	0	1	0.337	0	1
Ethyl benzoylacetate	0	0.174	0	0	0.223	0	0
Ethyl vanillin	0	0.120	0	0	0.394	0	1
Ethylene glycol dimethacrylate	0	0.087	0	0	0.473	1	1
Eugenol	0	0.035	0	0	0.071	0	0
Farnesal	0	0.133	0	0	0.001	0	0
Fluorescein-5-isothiocyanate	1	0.825	1	1	0.990	1	1
Formaldehyde	1	0.486	0	1	0.436	1	1
Furil	0	0.126	0	0	0.000	0	0
Glutaraldehyde	1	0.668	1	1	0.805	1	1
Glycerol	0	0.430	0	1	0.446	1	1
Glyoxal	1	0.529	0	1	0.746	1	1
Hexadecyl methyl sulphonate	1	0.595	1	1	0.270	0	0
Hexahydrophthalic anhydride	1	0.489	0	1	0.708	1	1
Hexane	0	0.007	0	0	0.067	0	0
Hydroquinone	1	0.577	1	1	0.184	0	0
Hydroxycitronellal	0	0.012	0	0	0.008	0	0
Imidazolidinyl urea	0	0.000	0	0	0.012	0	0
Isopropanol	0	0.061	0	0	0.330	0	1
Isopropyl eugenol	0	0.011	0	0	0.010	0	0
Isopropyl isoeugenol	1	0.183	0	0	0.011	0	0
Isopropyl myristate	0	0.005	0	0	0.028	0	0
Kanamycin	0	0.000	0	0	0.048	0	0

Chemical Name	Observed Class	USMAX			VS_USMAX		
		Predicted Probablity	Predicted Class (cutoff=0.57)	Predicted Class (cutoff=0.32)	Predicted Probablity	Predicted Class (cutoff=0.43)	Predicted Class (cutoff=0.31)
Lactic acid	0	0.365	0	1	0.261	0	0
Lauryl gallate (dodecyl gallate)	1	0.750	1	1	0.944	1	1
Lilial(p-*tert*-butyl-a-ethyl hydrocinnamal	0	0.005	0	0	0.001	0	0
R(+)-Limonene	0	0.012	0	0	0.03	0	0
Linalool	0	0.001	0	0	0.014	0	0
Lyral	0	0.046	0	0	0.025	0	0
Methyl dodecane sulphonate	1	0.867	1	1	0.149	0	0
Methyl salicylate	0	0.015	0	0	0.353	0	1
Methyl hexadecene sulphonate	1	0.644	1	1	0.718	1	1
Methyl hexadecyl sulphonate	0	0.529	0	1	0.135	0	0
Octanoic acid	0	0.044	0	0	0.339	0	1
Oleyl methane sulphonate	0	0.056	0	0	0.147	0	0
Oxalic acid	0	0.026	0	0	0.365	0	1
Oxazolone	1	0.065	0	0	0.205	0	0
p-Benzoquinone	1	0.967	1	1	0.965	1	1
p-Methylhydrocinnamic aldehyde	0	0.007	0	0	0.030	0	0
Pentachlorophenol	0	0.378	0	1	0.221	0	0
Phenyl Benzoate	0	0.262	0	0	0.023	0	0
Phthalic anhydride	1	0.898	1	1	0.924	1	1
Piperonyl butoxide	0	0.028	0	0	0.328	0	1
Product 2040 (2-methyl-4H,3,1-benzoxazin-4-one)	1	0.180	0	0	0.125	0	0
Propyl paraben	0	0.016	0	0	0.088	0	0
Propyl gallate	1	0.769	1	1	0.996	1	1
QRM 2113 (2-(4-tert-amylcyclohexyl)acetaldehyde)	0	0.022	0	0	0.003	0	0
Resorcinol	0	0.590	1	1	0.185	0	0
Saccharin	0	0.373	0	1	0.827	1	1
Salicylic acid	0	0.772	1	1	0.359	0	1
Sulphanilamide	0	0.764	1	1	0.307	0	0
Sulphanilic acid	0	0.178	0	0	0.000	0	0
Vinylidene dichloride	0	0.559	0	1	0.427	0	1

Table 3

Predicted probabilities and corresponding classifications of the test set using logistic regression QSAR models for the auto-scaled universal matrix (USMAX) and for the auto-scaled universal matrix after variable selection (VS_UMAX).

Chemical Name	Observed Class	USMAX			VS_USMAX		
		Predicted Probablity	Predicted Class (cutoff=0.57)	Predicted Class (cutoff=0.32)	Predicted Probablity	Predicted Class (cutoff=0.43)	Predicted Class (cutoff=0.31)
1-Bromododecane	0	0.004	0	0	0.003	0	0
1-Chloro-2,4-dinitrobenzene	1	0.830	1	1	0.748	1	1
1-Iodohexane	0	0.011	0	0	0.083	0	0
1-Phenyloctane-1,3-dione	0	0.254	0	0	0.027	0	0
1, 4-Phenylenediamine	1	0.985	1	1	1.000	1	1
a-Amyl cinnamic aldehyde	0	0.016	0	0	0.008	0	0
Benzyl bromide	1	0.187	0	0	0.212	0	0
C17 Azlactone	0	0.006	0	0	0.067	0	0
Cyclamen aldehyde	0	0.006	0	0	0.004	0	0
Maleic anhydride	1	0.939	1	1	0.959	1	1
Methyl 4-hydroxybenzoate(methylparaben)	0	0.032	0	0	0.393	0	1
N-Methyl-*N*-nitrosourea	1	0.176	0	0	0.608	1	1
Propylene glycol	0	0.436	0	1	0.154	0	0
Pyridine	0	0.331	0	1	0.293	0	0
Vanillin	0	0.144	0	0	0.812	1	1

Chem Res Toxicol. Author manuscript; available in PMC 2008 September 24.

Table 4

Interaction Pharmacophore Elements, IPEs, currently used in the 4D-QSAR paradigm

IPE Code	IPE abbreviation	IPE Description
0	any	All atoms in the molecule
1	np	Nonpolar atoms
2	p+	Polar (+) atoms
3	p-	Polar (-) atoms*
4	hba	Hydrogen bond acceptor atoms
5	hbd	Hydrogen bond donor atoms
6	aro	Aromatic atoms
7	hs	Non-hydrogen atoms

*Sometimes a carbon or sulfur will end of having a relatively large negative partial atomic charge[< -0.150] owing to the bonding topology in which it is involved. In these cases the 4D-QSAR paradigm considers the carbon or sulfur to be a polar negative IPE as opposed to a nonpolar IPE.

Table 5

Statistical measures for the model parameters of the logistic regression QSAR models for the non-scaled universal matrix (UMAX) and for the auto-scaled universal matrix (USMAX).

Parameter	UMAX		USMAX		Wald Chi-Square	P-value
	Estimate	Standard Error	Estimate	Standard Error		
Intercept	4.693	2.143	-1.745	0.3835	20.70	<0.0001
ε3(hbd,aro)	61.70	17.76	1.119	-0.3219	12.0733	0.0005
ε12(hs,aro)	754.1	201.0	2.686	0.7159	14.07	0.0002
ε18(all,all)	-231.6	76.10	-0.9722	0.3194	9.262	0.0023
ε2(np,p-)	22.33	5.809	1.654	0.4302	14.78	0.0001
ε2(all,all)	-47.99	16.59	-0.9936	0.3435	8.369	0.0038
ε1(p-,hbd)	-2.092	0.7979	-0.8841	0.3371	6.877	0.0087
ε19(hs,all)	-1014	291.3	-3.966	1.139	12.11	0.0005
ε54(all,all)	1753	745.6	1.467	0.6237	5.527	0.0187
ε39(all,all)	803.1	348.9	1.388	0.6029	5.299	0.0213

Table 6

Summaries of the performance measures of the logistic regression QSAR models for the auto-scaled universal matrix (USMAX) and for the auto-scaled universal matrix after variable selection (VS_UMAX).

Universal Matrix	Cutoff*	Date Set	Accuracy	Sensitivity	Specificity	Cross-validation
USMAX	0.57	Training	85.6%	68.3%	93.4%	78.0%
		Test	86.7%	60.0%	100.0%	——
	0.32	Training	79.5%	78.0%	80.2%	77.3%
		Test	80.0%	60.0%	90.0%	——
VS_USMAX	0.43	Training	88.6%	80.5%	92.3%	80.3%
		Test	86.7%	80.0%	90.0%	——
	0.31	Training	79.5%	80.5%	79.1%	75.8%
		Test	80.0%	80.0%	80.0%	——

*The two cutoffs for each universal matrix are the values under which the training set has a) the highest predictive accuracy, and b) a balanced error rate, respectively.

Table 7

The Akaike information criterion (AIC) statistics, highest leave-one-out cross-validation accuracies and Hosmer-Lemeshow goodness-of-fit statistic, χ^2_{HL}, of the logistic regression QSAR models with different numbers of descriptors for the auto-scaled universal matrix (USMAX) of the training set.

Statistics	Number of Descriptors in the model				
	6	7	8	9	10*
AIC	122.335	117.6	112.619	109.271	117.328
Cross-Validation Accuracy	77.3%	78.0%	78.0%	81.1%	81.8%
HL χ^2_{HL}	8.452	4.485	4.368	3.634	19.478
(p-value)	(0.391)	(0.811)	(0.823)	(0.8896)	(0.013)

* Model with 10 descriptors is obtained from the auto-scaled universal matrix after variable selection (VS_USMAX)

Table 8

Statistical metrics of the partial least square coupled logistic regression (PLS-LR) QSAR models for the non-scaled universal matrix (UMAX) of the training set.

Parameter	Estimate	Standard Error	Wald Chi-Square	P-value
Intercept	-3.361	1.136	8.754	0.003
xscr2	2.740	1.164	5.544	0.019
xscr3	7.050	2.700	6.840	0.009
xscr5	13.886	4.974	7.783	0.005
xscr7	11.487	3.911	8.628	0.003
xscr9	13.413	5.412	6.144	0.013
xscr12	35.282	13.266	7.067	0.008
xscr14	24.091	9.470	6.469	0.011
xscr18	44.636	16.003	7.787	0.005
xscr20	47.177	20.188	5.461	0.019
xscr24	45.722	19.112	5.727	0.017

Table 9

Statistical metrics of the partial least square coupled logistic regression (PLS-LR) QSAR models for the auto-scaled universal matrix (USMAX) of the training set.

Parameter	Estimate	Standard Error	Wald Chi-Square	P-value
Intercept	-3.564	0.890	16.054	<0.001
xscr1	0.498	0.129	14.878	<0.001
xscr4	0.407	0.111	13.358	<0.001
xscr5	0.636	0.169	14.150	<0.001
xscr8	0.513	0.172	8.879	0.003
xscr10	0.415	0.133	9.792	0.002
xscr13	0.700	0.228	9.396	0.002
xscr20	0.882	0.323	7.441	0.006

Table 10

The predicted probabilities and corresponding classification assignments of the training set using the partial least square coupled logistic regression (PLS-LR) QSAR models for the auto-scaled universal matrix (USMAX) and the non-autoscaled universal matrix (UMAX).

Chemical Name	Observed Class	USMAX			UMAX		
		Predicted Probablity	Predicted Class (cutoff=0.51)	Predicted Class (cutoff=0.36)	Predicted Probablity	Predicted Class (cutoff=0.28)	Predicted Class (cutoff=0.27)
1-(2′,3′,4′5′-Tetramethylphenyl)-3-(4′-tertbutylphenyl) propane-1,3-dione	0	0.000	0	0	0.000	0	0
1-(2′,5′-Dimethylphenyl)butane-1,3-dione	0	0.001	0	0	0.000	0	0
1-(3′,4′,5′-Trimethoxyphenyl)-4-dimethylpentane-1,3,-dione	0	0.000	0	0	0.000	0	0
1-Bromobutane	0	0.002	0	0	0.000	0	0
1-Bromohexane	0	0.002	0	0	0.000	0	0
1-Bromononane	0	0.000	0	0	0.000	0	0
1-Bromooctadecane	0	0.001	0	0	0.000	0	0
1-Bromotridecane	0	0.000	0	0	0.000	0	0
1-Bromoundecane	0	0.000	0	0	0.000	0	0
1-Butanol	0	0.000	0	0	0.000	0	0
1-Chloromethylpyrene	1	0.843	1	1	0.998	1	1
1-Chlorononane	0	0.000	0	0	0.000	0	0
1-Chlorooctadecane	0	0.000	0	0	0.000	0	0
1-Chlorotetradecane	0	0.000	0	0	0.000	0	0
1-Iodododecane	0	0.000	0	0	0.000	0	0
1-Iodohexadecane	0	0.000	0	0	0.000	0	0
1-Iodononane	0	0.000	0	0	0.000	0	0
1-Iodooctadecane	0	0.001	0	0	0.000	0	0
1-Iodotetradecane	0	0.000	0	0	0.000	0	0
1-Methyl-3-nitro-1-nitrosoguanidine	1	1.000	1	1	1.000	1	1
1-Phenyl-2-methylbutane-1,3-dione	0	0.009	0	0	0.000	0	0
12-Bromododecanoic acid	0	0.002	0	0	0.000	0	0
1,2-Dibromo-2,4-dicyanobutane	1	0.999	1	1	0.795	1	1
2-Acetylcyclohexanenone	0	0.131	0	0	0.266	0	0
2-Aminophenol	1	1.000	1	1	1.000	1	1
2-Ethyl butaldehyde	0	0.010	0	0	0.189	0	0
2-Hydroxypropyl methacrylate	0	0.008	0	0	0.000	0	0
2-Mercaptobenzothiazole	1	0.994	1	1	1.000	1	1
2-Methyl-5-hydroxyethylaminophenol	1	1.000	1	1	1.000	1	1
2-Methylundecanal	0	0.000	0	0	0.000	0	0
2-Nitro-p-phenylenediamine	1	1.000	1	1	1.000	1	1
2,2,6,6-Tetramethyl-heptane-3,5-dione	0	0.000	0	0	0.000	0	0
2,3-Butanedione	0	0.001	0	0	0.060	0	0
2,4,6-Trichloro-1,3,5-triazine (cyanuric chloride)	1	0.628	1	1	1.000	1	1
3-Ethoxy-1-(2′,3′,4′,5′-tetramethylphenyl)propane-1,3-dione	0	0.000	0	0	0.000	0	0
3-Methyl catechol	1	0.945	1	1	0.999	1	1
3-Methyleugenol	0	0.009	0	0	0.000	0	0
3-Phenylenediamine	1	1.000	1	1	1.000	1	1
4-Allylanisole	0	0.021	0	0	0.015	0	0
4-Amino-m-cresol	1	1.000	1	1	1.000	1	1
4-Hydrobenzoic acid	0	0.776	1	1	0.001	0	0
4-Nitrobenzyl bromide	1	0.999	1	1	1.000	1	1
4-Nitroso-N,N-dimethylaniline	1	0.996	1	1	1.000	1	1
4,4,4-Trifluro-1-phenylbutane-1,3-dione	0	0.111	0	0	0.000	0	0
5-Chloro-2-methyl-4-isothiazolin-3-one	1	0.930	1	1	1.000	1	1
5-Methyl-2,3-hexanedione	0	0.106	0	0	0.016	0	0
5-Methyleugenol	0	0.015	0	0	0.021	0	0
5-Methylisoeugenol	1	0.356	0	0	0.741	1	1
5,5-Dimethyl-3-thiocyanatomethyl-2(3H)-furanone	1	0.900	1	1	0.999	1	1

Chemical Name	Observed Class	USMAX			UMAX		
		Predicted Probablity	Predicted Class (cutoff=0.51)	Predicted Class (cutoff=0.36)	Predicted Probablity	Predicted Class (cutoff=0.28)	Predicted Class (cutoff=0.27)
6-Methylcoumarin	0	0.438	0	1	0.000	0	0
6-Methyleugenol	0	0.013	0	0	0.001	0	0
7-Bromotetradecane	0	0.000	0	0	0.000	0	0
7,12-Dimethylbenz[a]anthracene	1	0.999	1	1	0.522	1	1
a-Butyl cinnamic aldehyde	0	0.000	0	0	0.000	0	0
Abietic acid	0	0.000	0	0	0.000	0	0
Aniline	0	0.000	0	0	0.003	0	0
b-Propriolactone	1	0.407	0	1	1.000	1	1
Benzaldehyde	0	0.011	0	0	0.272	0	1
Benzo[a]pyrene	1	1.000	1	1	0.990	1	1
Benzocaine	0	0.427	0	1	0.013	0	0
Benzoyl peroxide	1	0.988	1	1	0.796	1	1
Benzoyl chloride	1	0.959	1	1	0.999	1	1
Benzyl benzoate	0	0.021	0	0	0.009	0	0
Bis-1,3-(2′,5′-dimethylphenyl)-propane-1,3-dione	0	0.000	0	0	0.000	0	0
Butyl glycidyl ether	0	0.127	0	0	0.268	0	0
C11-Azlactone	0	0.000	0	0	0.159	0	0
C15 Azlactone	0	0.000	0	0	0.024	0	0
C19-Azlactone	0	0.000	0	0	0.000	0	0
Camphorquinone	0	0.004	0	0	0.001	0	0
Chlorobenzene	0	0.001	0	0	0.000	0	0
Chlorpromazine hydrochloride	1	0.453	0	1	1.000	1	1
Cinnamic alcohol	0	0.001	0	0	0.000	0	0
cis-6-Nonenal	0	0.020	0	0	0.001	0	0
Coumarin	0	0.503	0	1	0.003	0	0
Diethylphthalate	0	0.001	0	0	0.592	1	1
Dimethyl sulfate	1	0.981	1	1	0.999	1	1
Dimethylsulfoxide	0	0.000	0	0	0.000	0	0
Diphenylcyclopropenone	1	0.993	1	1	0.999	1	1
Dodecylthiosulphonate	1	0.809	1	1	0.991	1	1
Ethyl acrylate	0	0.022	0	0	0.007	0	0
Ethyl benzoylacetate	0	0.037	0	0	0.005	0	0
Ethyl vanillin	0	0.063	0	0	0.004	0	0
Ethylene glycol dimethacrylate	0	0.002	0	0	0.000	0	0
Eugenol	0	0.005	0	0	0.053	0	0
Farnesal	0	0.000	0	0	0.000	0	0
Fluorescein-5-isothiocyanate	1	0.053	0	0	0.953	1	1
Formaldehyde	1	0.520	1	1	1.000	1	1
Furil	0	0.071	0	0	0.000	0	0
Glutaraldehyde	1	0.546	1	1	0.997	1	1
Glycerol	0	0.430	0	1	0.007	0	0
Glyoxal	1	0.996	1	1	1.000	1	1
Hexadecyl methyl sulphonate	1	0.243	0	0	0.054	0	0
Hexahydrophthalic anhydride	1	0.995	1	1	1.000	1	1
Hexane	0	0.017	0	0	0.000	0	0
Hydroquinone	1	0.743	1	1	0.208	0	0
Hydroxycitronellal	0	0.000	0	0	0.000	0	0
Imidazolidinyl urea	0	0.483	0	1	0.000	0	0
Isopropanol	0	0.000	0	0	0.000	0	0
Isopropyl eugenol	0	0.000	0	0	0.000	0	0
Isopropyl isoeugenol	1	0.885	1	1	1.000	1	1
Isopropyl myristate	0	0.001	0	0	0.000	0	0
Kanamycin	0	0.000	0	0	0.000	0	0

Chemical Name	Observed Class	USMAX			UMAX		
		Predicted Probablity	Predicted Class (cutoff=0.51)	Predicted Class (cutoff=0.36)	Predicted Probablity	Predicted Class (cutoff=0.28)	Predicted Class (cutoff=0.27)
Lactic acid	0	0.024	0	0	0.016	0	0
Lauryl gallate (dodecyl gallate)	1	0.947	1	1	1.000	1	1
Lilial(p-*t ert*-butyl-a-ethyl hydrocinnamal	0	0.000	0	0	0.002	0	0
R(+)-Limonene	0	0.028	0	0	0.000	0	0
Linalool	0	0.000	0	0	0.000	0	0
Lyral	0	0.000	0	0	0.000	0	0
Methyl dodecane sulphonate	1	0.173	0	0	0.991	1	1
Methyl salicylate	0	0.237	0	0	0.142	0	0
Methyl hexadecene sulphonate	1	0.988	1	1	0.949	1	1
Methyl hexadecyl sulphonate	0	0.106	0	0	0.003	0	0
Octanoic acid	0	0.000	0	0	0.000	0	0
Oleyl methane sulphonate	0	0.000	0	0	0.272	0	1
Oxalic acid	0	0.003	0	0	0.000	0	0
Oxazolone	1	0.999	1	1	0.998	1	1
p-Benzoquinone	1	0.992	1	1	1.000	1	1
p-Methylhydrocinnamic aldehyde	0	0.011	0	0	0.000	0	0
Pentachlorophenol	0	0.012	0	0	0.117	0	0
Phenyl Benzoate	0	0.067	0	0	0.000	0	0
Phthalic anhydride	1	0.994	1	1	1.000	1	1
Piperonyl butoxide	0	0.000	0	0	0.001	0	0
Product 2040 (2-methyl-4H,3,1-benzoxazin-4-one)	1	0.996	1	1	0.368	1	1
Propyl paraben	0	0.420	0	1	0.122	0	0
Propyl gallate	1	1.000	1	1	1.000	1	1
QRM 2113 (2-(4-tert-amylcyclohexyl)acet aldehyde)	0	0.000	0	0	0.027	0	0
Resorcinol	0	0.124	0	0	0.164	0	0
Saccharin	0	0.520	1	1	0.765	1	1
Salicylic acid	0	0.285	0	0	0.022	0	0
Sulphanilamide	0	0.273	0	0	0.000	0	0
Sulphanilic acid	0	0.726	1	1	0.002	0	0
Vinylidene dichloride	0	0.012	0	0	0.013	0	0

Table 11

The predicted probabilities and the corresponding classifications for the test set using the partial least square coupled logistic regression (PLS-LR) QSAR models for the auto-scaled universal matrix (USMAX) and the non-autoscaled universal matrix (UMAX).

Chemical Name	Observed Class	USMAX			UMAX		
		Predicted Probablity	Predicted Class (cutoff=0.51)	Predicted Class (cutoff=0.36)	Predicted Probablity	Predicted Class (cutoff=0.28)	Predicted Class (cutoff=0.27)
1-Bromododecane	0	0.000	0	0	0.000	0	0
1-Chloro-2,4-dinitrobenzene	1	0.821	1	1	1.000	1	1
1-Iodohexane	0	0.012	0	0	0.000	0	0
1-Phenyloctane-1,3-dione	0	0.020	0	0	0.018	0	0
1, 4-Phenylenediamine	1	1.000	1	1	1.000	1	1
a-Amyl cinnamic aldehyde	0	0.000	0	0	0.000	0	0
Benzyl bromide	1	0.007	0	0	0.000	0	0
C17 Azlactone	0	0.000	0	0	0.003	0	0
Cyclamen aldehyde	0	0.011	0	0	0.925	1	1
Maleic anhydride	1	0.983	1	1	1.000	1	1
Methyl 4-hydroxybenzoate(methylparaben)	0	0.259	0	0	0.001	0	0
N-Methyl-*N*-nitrosourea	1	0.000	0	0	1.000	1	1
Propylene glycol	0	0.033	0	0	0.920	1	1
Pyridine	0	0.001	0	0	0.265	0	0
Vanillin	0	0.989	1	1	0.971	1	1

Table 12

Summaries of the performance measures of the partial least square coupled logistic regression (PLS-LR) QSAR models for the auto-scaled universal matrix (USMAX) and for the non-autoscaled universal matrix (UMAX).

Universal Matrix	Cutoff*	Date Set	Accuracy	Sensitivity	Specificity	Cross-validation
USMAX	0.51	Training	93.2%	85.4%	96.7%	87.9%
		Test	80.0%	60.0%	90.0%	——
	0.36	Training	90.2%	90.2%	90.2%	87.1%
		Test	80.0%	60.0%	90.0%	——
UNAX	0.28	Training	97.0%	95.1%	97.8%	89.4%
		Test	73.3%	80.0%	70.0%	——
	0.27	Training	95.5%	95.1%	95.6%	89.4%
		Test	73.3%	80.0%	70.0%	——

*The two cutoffs for each universal matrix are the values under which the training set has a) the highest predictive accuracy, and b) a balanced error rate, respectively.