USE OF ADVANCED STATISTICAL LEARNING METHODS AND PRINCIPAL COMPONENT ANALYSIS IN QUANTITATIVE STRUCTURE–GENOTOXICITY RELATIONSHIP STUDY OF AMINES

Yueying REN[a1,b,*], Baowei ZHAO[a2,b] and Xiaojun YAO[c]

[a] *School of Environmental and Municipal Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China; e-mail: [1] renry02@st.lzu.edu.cn, [2] baoweizhao@mail.lzjtu.cn*

[b] *Engineering Research Center for Cold and Arid Regions Water Resource Comprehensive Utilization, Ministry of Education, Lanzhou 730070, China*

[c] *Department of Chemistry, Lanzhou University, Lanzhou 730000, China; e-mail: xjyao@lzu.edu.cn*

Received December 9, 2010
Accepted February 7, 2011
Published online March 8, 2011

The paper highlighted the use of advanced nonlinear modeling and subset selection techniques in the construction of a good, predictive model for genotoxicity study of amines. Essentials accounting for a reliable model were all considered carefully. Chemicals were represented by a large number of CODESSA descriptors. Division of a whole sample into the training set and the test set was performed by principal component analysis (PCA). Six descriptors selected by the best multi-linear regression (BMLR) method in CODESSA program were used as inputs to build nonlinear models, using advanced statistical learning methods such as support vector machine (SVM) and projection pursuit regression (PPR). The models were validated through three ways, i.e. internal cross-validation (CV), a test set and an independent validation set. Analysis shows that nonlinear models produced better results than linear models and PPR model outperforms the rest in the following order: PPR > SVM > linear SVM ≥ BMLR. In addition, the relationships between the descriptors and the mutagenic behavior of compounds are well discussed.

Keywords: Quantitative structure–genotoxicity relationship; Amine; Principal component analysis; Support vector machine; Projection pursuit regression.

Mutagenic potency is an important piece of information required by regulatory authorities all over the world, as a part of the safety evaluation process. It is used for screening of substances (potentially hazardous to human health) before their release into the market[1]. This potential is measured by genotoxicity tests based on the fact that chemicals that show adverse effects when interacting with genetic material (DNA) of cells are known as genotoxic[2], and most human carcinogens are genotoxic in nature.

doi:10.1135/cccc2010143

Like other toxicological data, the genotoxicity of the existing chemicals is incomplete because with presently available resources, thorough toxicological testing of all chemicals is neither economically feasible, nor, for ethical reasons (animal protection), justifiable. Therefore, effort has been directed at developing low cost and efficient approaches for the prediction of genotoxicity. Methodologies based on quantitative structure–activity relationships (QSAR) study have the potential to fill above requirements and genotoxicity related (Q)SAR models are currently being integrated into emerging data-gap filling applications, such as the OECD's QSAR Application Toolbox[3]. A good example is the knowledge- and rule-based expert systems (ES), e.g. Deductive Estimation on Risk from Existing Knowledge (DEREK)[4], which uses structural alerts in combination with pattern recognition routines to identify substructures associated with specific toxic effects. This method is characteristic of transparency in the form of structure alerts that have a straightforward interpretation. However, it cannot provide prediction for non-positive structures[5]. Apart from ES, genotoxicity-based (Q)SAR studies have been carried out for a diverse group of chemicals. Aromatic and heteroaromatic amines, as a class of widespread and ubiquitous environmental pollutants, have gained increasing interest in such research. A significant number of studies have been carried out on them alone applying different kinds of techniques ranging from classical multiple regression analysis to machine learning (neural networks[6,7]/pattern recognition/statistical learning methods[8]) to two-dimensional and three-dimensional QSAR methodologies (e.g. comparative molecular field analysis (CoMFA))[9]. Genotoxicity was correlated with different descriptors such as log P, E_{HOMO} and E_{LUMO} by Debnath et al.[10], topological descriptors by Basak et al.[11,12], atomic surface areas, coulombic and electron exchange energies by Katritzky et al.[13], Dragon descriptors by Gramatica et al.[14] and E-state indices by Cash et al.[15], etc. These works were also well reviewed by Benigni[16,17], Vračko[18], and Gramatica[19] very recently.

In the present research, QSAR study was carried out to predict the genotoxicity of 124 aromatic and heteroaromatic amines. The same data was also studied by Cash et al.[15] using E-state indices with the conclusions that (i) it is important to understand the difference between a model's fit and a model's predictive ability and (ii) training set statistics and internal validation techniques alone may be very misleading when trying to evaluate the true predictive accuracy of a model, thus highlighted the need to perform an external validation of a model to assess its true predictive ability. This opinion was recommended by other researchers, e.g. Zefirov et al.[20], Tropsha et al.[21] and Gramatica et al.[22], and requested also by the OECD

principles[3]. In addition to the above arguments, what aroused our curiosity of this work is the big difference between the model's fit ability and predictive ability in terms of squared regression coefficient (R^2) given by the authors, i.e. goodness-of-fit parameters R^2 = 0.78 for model 1 (a seven-descriptor equation) and 0.77 for model 2 (a six-descriptor equation), with the respective predictive R^2 = 0.27 (model 1) and 0.44 (model 2) for test sets. According to the authors, these should be contributed to the possible over-fitting of models[15]. Based on our experience with these, however, what to doubt first should be the reliability of the division of samples (i.e. data splitting into training set and test set) rather than the possible over-fitting supported by the authors[15]. In other words, the training set was not representative of the test set. To support this point of view, a principal component analysis (PCA) was performed in the present study to help us inspect the two models by Cash et al.[15] and further understand the data distribution as well. PCA was also used for subset design in this study, in comparison to another famous subset selection method, i.e. Duplex, to generate a representative training set. The widely used CODESSA descriptors, which encode various aspect of structural information, were calculated to provide information space for these purposes.

In addition, it is well-known that chemicals exhibit toxicity via different mechanisms of toxic action and, as Cronin and Schultz[23] pointed out, that biology and the modeling of biology is a nonlinear phenomenon in essence and that always expecting linear relationships in biological modeling is not realistic. As mentioned previously, neural networks have been used in genotoxicity modeling for amines[6,7]. Apart from neural networks, a number of nonlinear modeling methods have been developed in the field of statistics to handle nonlinearity exhibited in a given data set. In this study, support vector machine (SVM) and project pursuit regression (PPR) were employed for this purpose.

The developed models were validated through three ways, i.e. internal cross-validation (CV), a test set and an independent validation set in terms of R^2 and the root mean squared error (RMSE), which is calculated as the root square of the sum of squared errors in modeling or prediction divided by their corresponding total number. All models were compared with respect to above two statistical parameters in order to determine a reliable predictive model for genotoxicity prediction. The preferred model should have the highest statistical parameter values (i.e. highest R^2/smallest RMSE) and the most balanced results (i.e. very similar R^2/RMSE values) for training and test set chemicals, highlighting the model's generalizability[24]. It was also explored applying n-fold cross-validation procedure and the results

were compared to those by training set/test set procedure to verify its stability.

MATERIAL AND METHODS

Data set: The data set reported in the study of Cash et al.[15] was used, which involves 124 aromatic and heteroaromatic amines. Mutagenic potency was expressed as log reversions per nanomole of compound (LogR) in the strain Salmonella typhimurium TA98+S9 with the addition of an exogenous metabolic activation system. The structures as well as mutagenic potency marked as LogR are shown in Supplementary Material S1.

Descriptors generation and selection: Overall, 618 descriptors classified as constitutional, topological, geometrical, electrostatic and quantum chemical descriptors were generated using CODESSA package[25]. These descriptors encode information about the connections between atoms, shape, branching, symmetry, distribution of charge, and quantum-chemical properties of the molecule. Before calculation of the descriptors, the molecules were optimized with AM1 method in Hyperchem 6.0 [26], with no symmetry constraints imposed and applying a gradient norm limit of 0.1 kcal/mol as a stopping criterion.

Many of the calculated descriptors carry redundant or highly correlated information and their existence would result in a chance correlation during the construction of the model. Therefore, once descriptors were generated, feature selection should be performed to reduce the original pool of descriptors to an appropriate size and choose a subset of descriptors that is significantly correlated with the property of interest. The selection includes objective procedure and subjective procedure. In objective feature selection, the independent variable (i.e. descriptors) alone was used to filter out useless ones employing identical test, pairwise correlation test, and vector space descriptor analysis, etc.[27]. The remaining descriptors were then reduced by subjective feature selection to search for an information-rich subset of descriptors. Here, the best multi-linear regression (BMLR) method in CODESSA was used for this purpose. This method implements the following strategy to search for the multi-parameter regression with the maximum predicting ability. It commences by correlating the given property/activity employing two-parameter regression with pairs of orthogonal descriptors (default value R^2 of the inter-correlation less than 0.1). The descriptors sets with highest correlation coefficients are chosen to perform higher order regression. Further inclusion of non-collinear descriptors (default value $R^2 < 0.6$) in the regression is made, one descriptor after another,

on the basis of the improved Fisher criterion F at a given probability level upon successive addition of descriptors[28]. The correlation of descriptors in the model can thus be efficiently avoided.

Support vector machine for regression (SVR): SVM algorithm was proposed by Vapnik and co-workers in 1995 and detailed description can be found in a tutorial[29]. Briefly, this method was designed around the computation of an optimal separating hyperplane which provides minimum expected generalization error in a multi-dimensional space called "feature space". By the use of a kernel function, the input data is first mapped into the feature space and then linear regression is performed in this space. The elegance of using kernel function lies in the facts that one can deal with feature spaces of arbitrary dimensionality without having to compute the map explicitly and SVM can actually locate the hyperplane without ever representing the feature space explicitly. In the function estimation problems, the radial basis function kernel is most commonly used because of its effectiveness and speed in the training process.

Projection pursuit regression (PPR): For many problems with high-dimensional data, the most common practice is using dimension reducing transformations such as linear projections to project the original data into a lower-dimensional space, line or a plane, etc., to try to find the intrinsic structure for visual inspection. More precisely, given a data set $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_n)$, $X \in IR^k$ are k-dimensional matrix ($k \times n$), where k is the number of observed variables and n is the number of units, and an orthonormal matrix $\alpha(m \times k)$. Then a matrix with a dimension ($m \times n$) is constructed by multiplying matrix $\alpha(m \times k)$ to $X(k \times n)$ and represents the coordinates of the projection data onto the m-dimensional ($m < k$) space spanned by the rows of α. As there are infinitely many projections from a higher dimension to a lower dimension, it is important to have a technique to pursue a finite sequence of projections that can reveal the most interesting structures of the data. "Projection pursuit" (PP) presented by Friedman and Turkey is such a powerful tool that combines both ideas of projection and pursuit[30,31] and, therefore, can overcome the cause of dimensionality. Briefly, its basic idea was to assign a numerical index (named a PP index $I(a)$ to measure the "interestingness" of projection; the larger the index value, the more interesting the projection is) to every projection and then maximize the index, via numerical optimization by a PP algorithm, over all possible α. For an observed pair (X,Y) of random variables, where $X \in IR^k$ is a k-dimensional variable and $Y \in IR$ is a response, projection pursuit for regression (PPR) aims to approximate the regression function $f(x) = E(Y|X = x)$ by a finite sum of ridge functions

$$g^{(p)}(x) = \sum_{i=1}^{p} g_i(\alpha_i^T, x)$$

where α_i are $m \times n$ orthonormal matrices, p is the number of ridge functions. PPR model can be used to approximate a large class of function by suitable choices of α_i and g_i. It was found that the largest of the first several maxima from Friedman's PP algorithm is often very close to the global maximum[32]. Therefore, this algorithm was used in this study, where g_i is found by smoothing operation that entails a back-fitting[30,31]. SVM and PPR are conducted using calculation programs written in *R*-file based on *R* script.

RESULTS AND DISCUSSION

Inspection of two models by Cash et al.[15]: Training set statistics and internal validation results for two linear models by Cash et al.[15] were good and similar to each other, with goodness-of-fit parameters $R^2 = 0.78$ and $R^2_{adj} = 0.76$ for model 1 and 0.77 and 0.75 for model 2, respectively. However, for test set, the respective predictive R^2 was only 0.27 and 0.44. Neither model 1 nor model 2 showed acceptable predictive accuracy when subjected to the external validation though the internal validation statistic (Q^2_{LOO}) indicated that both models would be expected to predict LogR well. Closer examination of test set results revealed that only three of the eight descriptors enclosed in model 1 were represented in the test set, indicating that the test set may not have been adequately representative of the training set. It means that many compounds in the test set may have been outside the valid prediction space of the training set. On the contrary, an increase in the predictive R^2 for test set (from 0.27 in model 1 to 0.44 in model 2) demonstrates that the improvement of the valid prediction space can result in corresponding improvements of the model's predictive ability, as subsets in model 2 were generated based on the adjustment of subsets in model 1 considering the relationship between the descriptors and the structure of chemicals in the test set. Nevertheless, the improvement was not so significant and a predictive R^2 of 0.44 for the test set indicates that it did not adequately predict LogR yet.

Based on above observation, there is reason to believe that the big difference between fit ability and predictive ability of models by Cash et al.[15] should be, or at least to a great extent, attributed to data separation amongst the chemicals. In other words, compounds in the test set are

underrepresented in these models, the chemical 'space' – or extent of chemical diversity represented by the molecules used to construct the model – is enriched with rather different structures. With the objective to prove this hypothesis, a principal components analysis (PCA) was performed within the calculated CODESSA descriptors space for the whole data set. The calculated principal components (PCs) can be used to derive scores to display most of the original variations in a smaller number of dimensions. These scores can also allow us to recognize groups of samples with similar behavior. Details about PCA can be found in ref.[33].

Here, PCA gives 34 PCs with eigenvalues > 1. Of them, the first three significant PCs explain 56.73% of the variation in the data (26.94, 15.85 and 14.94%, respectively). The distribution of compounds over the first three PCs information space is shown in Fig. 1. As can be seen in this figure, all 124 chemicals are basically formed into two groups over the PCs space; one group is composed of Anilines and Quinolines and another one consists of the rest chemicals, i.e. Biphenyls, Fluoranthenes, Naphthalanes and Phenazine etc., with only few anilines fall into the later.

Spatial distribution of compounds in model 1 (denoted as A for test set compounds in Supplementary Material S1) by Cash et al.[15] over the first three PCs space is shown in Fig. 2. As can be seen, the training set is very different from those in the test set and cannot represent the test set at all. Combined with Fig. 1, it is obvious that all Anilines came into the training

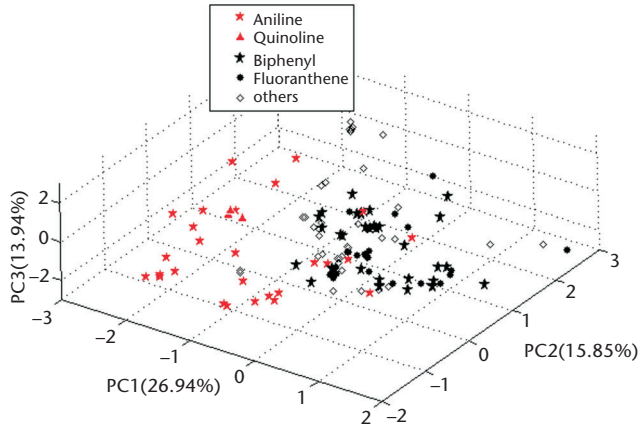

FIG. 1
Distribution of whole data set over the first two PCs space

set accounting for more than one third of the parts that compose the training set while the test set contains only biphenyls, fluoranes and naphthalanes. it is without the question that the unbalance of the components in two subsets contributed much to the big difference between fit ability and predictive ability of model 1.

As for compounds in model 2 (denoted as B for test set compounds in Supplementary Material S1) by Cash et al.[15], each set seem to be relatively balanced over the space of the PCs (not shown). The difference between the training set and the test set are not so obvious as in Fig. 2 but it is still not

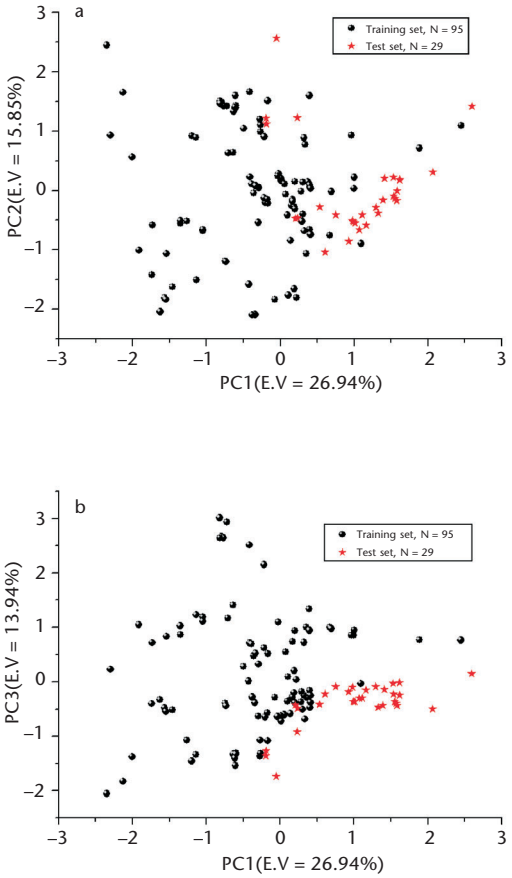

FIG. 2
Distribution of subsets in model1 over the first three PCs space

well enough as the training set does not consist of representatives of the most dissimilar structures (e.g. No. 48 3,5-diisopropyl-4-aminobiphenyl, No. 76 4,4′-methylenebis(*o*-isopropylaniline) and No. 93 7-adamantyl-2-aminofluorene), thus affecting the applicability of subset splitting. On the other hand, it can be used to explain the difference between the model's fit ability and predictive ability and, to some extent, also the improvements of the predictive ability of model 2 over model 1.

To demonstrate it more clearly, in the present study new QSAR models were built for same subsets by Cash et al.[15] using CODESSA descriptors. The results were very similar to those by Cash et al.[15]. All linear models demonstrated high ability in predicting genotoxicity of the training set compounds, with R^2 ranging from 0.7344 (two-descriptor equation) to 0.8672 (seven-descriptor equation) for data set 1 and 0.7639 (two-descriptor equation) to 0.8189 (five-descriptor equation) for data set 2, respectively. With respect to the test sets, the predictive results are unacceptably low with a highest predictive R^2 value of 0.0604 and 0.3021, respectively. Inspection of the detailed predicted results showed that compound No. 93 (7-adamantyl-2-aminofluorene) affects the test set statistics significantly. The highest predictive R^2 for two test sets would be increased to 0.4493 and 0.609, respectively, after discarding this compound from the test set. Nevertheless, the unbalanced results are still unacceptable.

The above observation demonstrates that data separation (subset selection) is of crucial importance in the development and validation of reliable QSARs. The quality of the prediction depends highly on the data set used to develop the model and similarity between the training set and the test set can affect the predictive ability of the models dramatically. The rational division of the subsets should satisfy that, on one side, the diversity of the training set, which is a necessary condition for the construction of a QSAR model applicable to further compounds of interest in the same chemical domain and, on the other side, the closeness of the representative points of both the training set and the test set in the descriptor space that ensures a proper validation of the model[7]. This can be shown intuitively over PCs space occupied by the entire data set to some extent. Therefore, what need us to do is to try different splitting methodologies to generate a representative training set and test set. Before that, 20 chemicals were selected randomly to form an independent validation set to check the generalization ability of the developed model: by selection of every sixth point starting from compounds No. 3 (4-chloroaniline).

General approaches for selecting representative training set samples including random selection, *D*-optimal concept, Kenstone algorithm, Næs'

cluster analysis and Duplex algorithm, etc. Amongst them, Duplex seems to be the best way to select representative training and test sets in a validation context and its principle as well as the treating procedure are described in literature[34]. In addition to above approaches, PCA was another very useful method proved by our previous works to assist the data splitting[35,36]. Design of the training set is performed with respect to the PCs by selecting a subset of substances that are most efficient in spanning the substance (or PCA model) space. When the number of PCs is less than three, it is often sufficient to select samples manually from visual inspection of score plots. This method, together with Duplex algorithm, was used in the present work to generate different QSAR subsets with two aims in mind: first, to further explore the importance of data splitting on the performance of the QSAR model and, second, to compare the efficiency of two methods for this specific case. According to general separation ratio[37], the rest 104 compounds were split into 80% for the training set (80 compounds) to develop models and 20% for the test set (24 compounds) to evaluate the model performance.

BMLR method in CODESSA program was used to select most relevant descriptors and construct QSAR linear models. For each data set, series of linear models containing different number of descriptors were generated. To avoid the "over-parametrization" of the model, an increase of the R^2 value of less than 0.02 was chosen as the breakpoint criterion[38].

Linear models based on Duplex method: Predictive R^2 for test set by all models are less than 0.26, indicating that the models have rather poor predictive accuracy though fit for the training set are very satisfactory. The optimum model is a five-descriptor linear equation given $R^2 = 0.8128$ and 0.244, and RMSE = 0.7789 and 1.8878 for the training set and the test set, respectively. In this sense, it means that the descriptors in the models, which are structural features selected out for compounds in the training set and therefore can reflect important structural information related to the studied property (in this case, it is LogR of the chemicals) in the training set, cannot cover the whole range of the compounds in the test set. With respect to the two models in literature15, there is reason to doubt the subset selection. Comparison of spatial distribution over PCs space revealed that the training set does not consist of representatives of the most dissimilar structures (as in model 2 in ref.[15]), thus affecting the applicability of data set splitting.

Compound No. 93 (7-adamantyl-2-aminofluorene) affects the test set statistics significantly again. Discarding this compound from the test set resulted in improvements on the predictive R^2, with the highest $R^2 = 0.617$. If

we transfer this compound into the training set and develop new models using same procedure, the predictive statistics for new test set would increase. However, the results are still not good; the highest predictive R^2 = 0.463 corresponding to a six-descriptor linear equation (training set R^2 = 0.8170). In addition, in both cases, models were not stable especially for the test set statistics. From above observation, it is concluded that models based on Duplex subset selection algorithm were not satisfactory in this case.

Linear models based on PCA subset design method: Subset selection based on PCA can be checked by observing their spatial distribution over the first three PCs information space (Fig. 3). It is seen that the training set covers evenly the PCs space and that the compounds of the test set are close to those of the training set with the most dissimilar compounds enclosed into the training set; thus allows predictions to be made by interpolation and not extrapolation out of the domain of the particular QSAR model[39].

Best multi-linear regression models including up to eight descriptors were obtained; details are described in Supplementary Material S2. According to 0.02 break criterion[38] during the construction of the model, a six-desriptor linear equation was considered as the optimum model to correlate logR to structural features.

The model is presented in details in Table I. Analysis of the descriptors correlation matrix indicates that no obvious correlation exists between these descriptors and that the obtained model has statistic significance.

TABLE I
Linear model based on data splitting and BMLR feature selection

Descriptor	Meaning of descriptor	B	Std. error	Beta	t-test
	constant	213.6157	58.5056		3.6512
RN_R	realtive number of rings	31.4574	6.2179	0.4874	5.0592
N_{BR}	number of benzene rings	1.6038	0.2846	0.6253	5.6360
^2CIC	complementary information content (order 2)	–0.0647	0.0128	–0.6782	–5.0412
$ER_{max,C-H}$	maximal resonance energy for a C–H bond	–18.4986	5.1885	–0.2789	–3.5653
#TMEI	total molecular electrostatic interaction/# of atoms	–1.6448	0.3941	–0.4102	–4.1731
I_C	principal moment of inertia C	–137.1430	22.3205	–0.5706	–6.1443

Using this model, the genotoxicity of each compound was predicted; shown in Supplementary Material S1. No outliers were found according to three times standard deviation criterion. Correlations between predicted and observed genotoxicity reveal that both fit ability and predictive ability were good with R^2 = 0.749 and 0.75, and RMSE = 0.8949 and 0.9854 for the training set and test set, respectively. For the whole data, it can correctly predict 75% of the variance with RMSE = 0.9166. In addition, a scrambling procedure was applied to check the chance correlation during the model construction[40]. Nine-trivial randomizations resulted in an average R^2 value

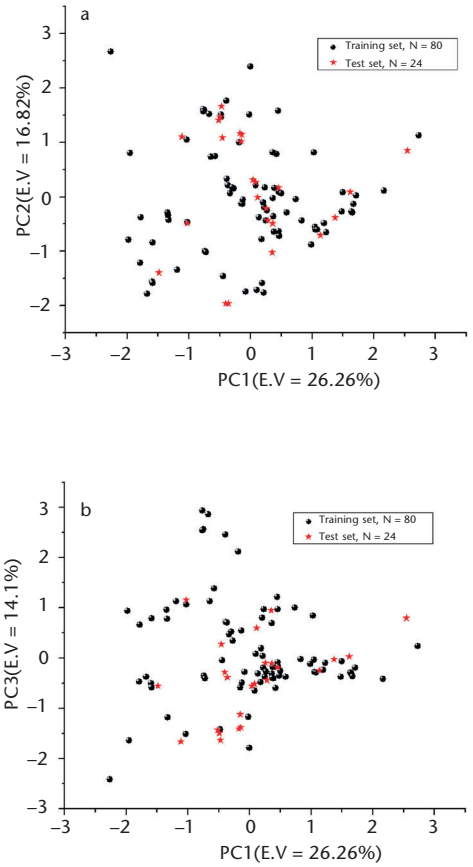

FIG. 3
Distribution of subsets by PCA algorithm over the first three PCs space

of 0.112 (shown in Supplementary Material S2), which is far less than 0.75 by BMLR model and thus showed the breakdown of predictive power of the model. This was accepted as a proof of validity of the BMLR model we developed.

Compared to linear model based on Duplex algorithm, both fit ability and predictive ability of this model were satisfactory. It indicates that descriptors enclosed into the model can describe the whole data set relatively well. However, none of the descriptors showed individual strong linear relationship with logR. In this sense, it means that these descriptors work together in a complex way to correlate with the mutagen potency of compounds.

For the purpose to investigate possible nonlinearities and to avoid the limitations imposed by the multi-linear method, nonlinear methods such as SVM and PPR were used to perform further QSAR study. The same descriptors selected by BMLR method were used. In addition, considering the differences between groups in the training set, a six-fold cross-validation procedure was performed on the whole data set[41]. The averaged results were compared to those predicted by each nonlinear model.

SVM model based on PCA subset design method: To evaluate whether the relationship is really nonlinear, linear SVM utilizing linear kernel function as well as nonlinear SVM utilizing RBF kernel function were used to develop linear SVM model and nonlinear SVM model, respectively. The quality of SVM depends on a good choice of parameters combination, i.e. the regularization parameter (cost, C), the nature and the parameters of the kernel function. As parameters influence each other, a systematic grid search method was utilized to determine the best set, using the minimum mean squared error (MSE) of leave-one-out (LOO) cross-validation of the training set as the optimal condition.

For nonlinear SVM modeling, the final optimal model was determined as $C = 800$, $\gamma = 0.02$ and $\varepsilon = 0.24$. This model gives $R^2 = 0.8692$ with RMSE = 0.6521 for the training set and predicted $R^2 = 0.8390$ with RMSE = 0.7270 for the test set, respectively. In the case of linear SVM modeling, systematic grid search process determined $C = 50$ and $\varepsilon = 0.0032$ as final optimal model parameters resulting in $R^2 = 0.7595$ and 0.7874, and RMSE = 0.8951 and 0.8962 for the training and test set, respectively.

Direct comparison of the statistical parameters and predictive power of three models we developed above (BMLR, linear SVM and nonlinear SVM models) is feasible and it appears that the model quality of linear SVM model is comparable to that of BMLR model; both are poorer than results by nonlinear SVM model. The detailed predicted results are listed in

Supplementary Material S1. A closer examination of the residuals reveals that, generally, the data points from the nonlinear SVM have smaller deviations from the regression line than the linear models. In this sense, it implies that the factors influencing the genotoxicity were complex and not all of them relate to genotoxicity in a linear fashion. On the contrary, the relationships between the structure and genotoxicity are, in principle, nonlinear and nonlinear methods are more capable of recognizing this nonlinearity.

PPR modeling based on PCA subset design method: The PP algorithm proposed by Friedman was used to construct PPR model and the quality of PPR modeling depends on the choice of parameters, i.e. "nterms", "optlevel" and "span". "nterms" amounts to the number of variables in the model; in this case, nterms = 7. "optlevel" is an integer from 0 to 3. It means the levels of optimization which differ in how thoroughly the models are refitted during this process. "span" defines the fraction of the observations in the span of the running lines smoother. Thus, the parameters need to be determined in this study is "optlevel" and "span". As there are no clear guidelines for selecting the optimum set of theoretical parameters, the only practical way of finding the above two terms is through extensive experiments optimizing the PP index.

The final optimal set was determined as "optlevel" = 1 and "span" = 0.22, respectively. Compared with the nonlinear SVM model developed above, for the training set, the PPR model gave better fit with an increased R^2 of 0.8848 and decreased RMSE of 0.6078. The improvement indicates that the training set is described more accurately and the PPR model is expected to be a better predictor for genotoxicity than nonlinear SVM model. As expected, the predictive ability of this model was also satisfactory with the predictive R^2 value increased to 0.8440 and RMSE value dropped to 0.7423, indicating the good generalization capability of the PPR model. This is also demonstrated in Fig. 4, which shows the graphic presentation of the relationship between the experimental and predicted logR by BMLR, nonlinear SVM and PPR modeling. As can be clearly seen, the data points from the nonlinear models show smaller deviations from the regression line than the BMLR model; the PPR model performs best. Predictions for following compounds, i.e. No. 27 (4-chloro-1,2-phenylenediamine), No. 59 (3,5-diethyl-4-aminobiphenyl), No. 81 (7-aminofluoranthene), No. 93 (7-adamantyl-2-aminofluorene), No. 96 (1-*n*-butyl-2-aminofluorene), No. 115 (2,7-diaminophenazine), No. 120 (1-aminopyrene) by PPR model are much better than those by BMLR and SVM. For compound No. 106 (3-aminophenanthrene) all models give similar prediction with deviations more

than 1.6 log unit. Besides, the largest deviation by PPR model is 1.74 log unit (i.e. compound No. 6 (4-bromoaniline)); whereas the largest deviation by BMLR and SVM are 2.82 and 2.28 log units (for compound No. 115 (2,7-diaminophenazine)), respectively.

Generally, nonlinear models produced better results than linear models and PPR model outperforms the rest in the following order: PPR > SVM > linear SVM ≥ BMLR (also shown in Table II). Detailed analysis reveals that improvements on test set are more significant, indicating best generalization ability of PPR model. In particular, within an absolute error of 1 log unit, PPR can correctly predict genotoxicity for 87.5% compounds in the test sets; while for the BMLR and SVM, the respective proportion were 66.67 and 79.17%, respectively (shown in Fig. 5).

Based on above observations, we have such conclusions that (i) PPR model is clearly superior both in fitness and in prediction performance for this end point of interest and, (ii) PPR model can simulate the nonlinear relationship within the data set investigated more accurately. Besides, PPR is computationally quite feasible and simple and it takes much less time to convergence compared to SVM model. Since the averaged R^2 value by six-fold cross-validation procedure was very similar to that obtained based on the training set/test set (shown in Supplementary Material S2), it can be concluded that the models we developed are stable. Finally, the independent validation set of 20 compounds was utilized to evaluate the model's generalization ability. Predicted R^2 and RMSE for independent validation

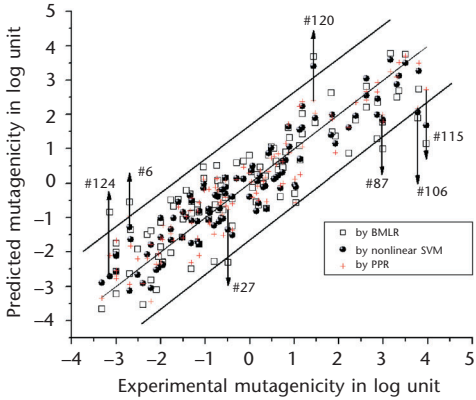

FIG. 4
Comparison of the predicted results for whole dataset by BMLR, nonlinear SVM and PPR

set are 0.8350 and 0.6440, respectively, indicating the good generalization ability of PPR mode we developed (shown in Table II). For the convenience of comparative analysis, the external validation results for other three mod-

TABLE II
Statistical parameters for all models we developed in the present study

Statistical terms	Data set	Number of compounds	BMLR	Linear SVM	SVM	PPR
R^2	test set	24	0.7500	0.7874	0.8390	0.8440
	training set	80	0.7490	0.7595	0.8692	0.8848
	whole set (train+test)	104	0.7471	0.7640	0.8590	0.8730
	independent validation test	20	0.7800	0.7893	0.8155	0.8350
RMSE	test set	24	0.9854	0.8962	0.7758	0.7423
	training set	80	0.8949	0.8951	0.6521	0.6078
	whole set (train+test)	104	0.9166	0.8953	0.6826	0.6300
	independent validation test	20	0.7054	0.6708	0.6521	0.6440

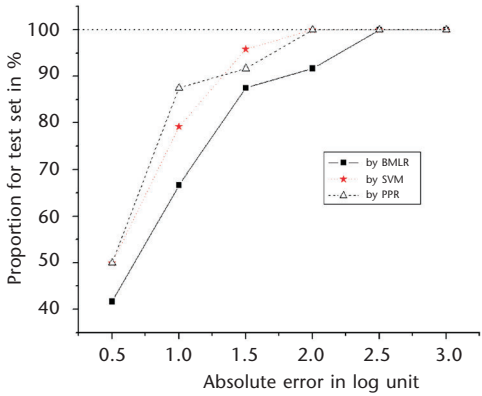

FIG. 5
Proportion of compounds in test set within a given deviation from the experimental logR by threer models

els, i.e. BMLR, linear SVM and nonlinear SVM, were also listed in this table. It is obvious that utilizing PPR as modeling technique to predict genotoxicity is a good choice.

Interpretation of the descriptors: Apart from the acceptable predictive ability of the studied property/activity, physicochemical interpretation of the selected descriptors is another prerequisite of a good predictive model. In other words, the constructed model should provide good prediction but also help investigate the underlying physical phenomenon and identify the key molecular features associated with the property/activity of interest as well. For toxicity topics, it might help provide some chemical clues for the identification of risk chemicals from a large group of structurally related molecules with less cost compared to the animal test. Nevertheless, the mechanistic interpretation of descriptors is seldom easy, especially for a biochemical mechanism that always involves toxicity topics and for high heterogeneous data set.

As described previously, the descriptors were selected into the BMLR model through a systematic way and the scrambling procedure does not indicate presence of chance correlation during the construction of models. On the basis of such observations, there is reason to believe that by interpreting the descriptors in linear model, it is possible to gain some insight into factors that are likely to affect the genotoxicity. In this study six descriptors were found to be important to genotoxicity. Of them, two are constitutional (RN_R and N_{BR}), one is topological (^2CIC), and the rest are quantum-chemical ($ER_{Max,C\text{-}H}$, #TMEI and I_C), respectively. These descriptors encode different structure information affecting the mutagenic interaction mechanism and the significance of each descriptor in the model can be checked with corresponding t-test values.

RN_R (relative number of rings) is the proportion of ring to all atoms in a molecule. Inclusion of this descriptor into the model is not surprising at all as the compounds under investigation are aromatic and heteroaromatic amines. It plays important role on genotoxicity in a similar way to another constitutional descriptor, N_{BR} (number of benzene rings). N_{BR} is approximately proportional to the area of hydrophobic aromatic hydrocarbon art of the molecule and can be therefore related to the hydrophobicity of the (poly)cyclic compounds as well[42]. This descriptor has also been found of significant importance to genotoxicity of aromatic and heterocyclic amines in other publications and the authors concluded that the size of the ring system can affect the genotoxicity in various steps in genotoxicity mechanism, but most probably it affects the penetration through the biomembraness[13,43].

^{2}CIC (secondary complementary information content) is defined in ref.[44]. It encodes the information on size and degree of branching of a molecule. In other words, it represents the difference between the maximum possible complexity of a graph and the realized topological information of the chemical species as defined by the information content. The molecular polarizability strongly depends on the size of molecule. The polarizability values have also been shown to be related to the hydrophobicity and the higher order polarizability terms are also known to characterize the electrophilic properties of the molecule[45]. Therefore, the specific information on the complexity of a topological graph and skeletal variations of the chemical species may lead to difference of the steric property and the hydrophobic of the compounds. This descriptor shows the largest negative influence on the genotoxicity of the amines and it is shown in this work that the genotoxicity of amines decrease with growing steric demand of substituents. For example, derivatives of compound No. 84 (2-aminofluorene) substituted by different size of alkyl group in the ortho position of the amino funtionality with growing steric demand, i.e. compound No. 91 (1-ethyl-2-aminofluorene), compound No. 97 (1-*i*-propyl-2-aminofluorene), compound No. 96 (1-*n*-butyl-2-aminofluorene) and compound No. 94 (1-t-butyl-2-aminofluorene). Similar trend can also be seen for ortho substituted 4-aminobiphenyl derivatives. For derivatives with substituents "far away" from the amino functionality, this is also the case, e.g. compound No. 95 (7-methyl-2-aminofluorene), compound No. 94 (7-*tert*-butyl-2-aminofluorene) and compound No. 93 (7-adamantyl-2-aminofluorene), etc. Similar observations can be found in literatures[46–48].

$ER_{Max,C\text{-}H}$ (max resonance energy for a C–H bond) is an energy partition term related to the site in the molecule where the resonance between the carbon and hydrogen is the strongest. It also represents the kinetic energy and electronuclear attraction energy associated with a charge distribution that lies between two atoms. The presence of this descriptor in the model may be related to the formation of highly reactive radical centres in the aromatic systems that affect the reproductory system of cell[49]. Another energy partition term related descriptor is #TMEI (total molecular electrostatic interaction/# of atoms), which characterizes the total energy of the molecule in electrostatic energy scales and describes the electrostatic feature of the molecule.

The principal moment of inertia C, I_C, characterizes the mass distribution along the longest rotational axe of the molecule[50] and thus provides information on rigidity of a molecule. It also shows the lowest diameter of the molecule, which has effect to the penetration through the bio-membranes.

I_C is useful for distinguishing between the isomers[50], especially the positional isomers.

In conclusion, these structural factors can be classified into three groups, related to bulk properties of the compound described through size and shape (RN_R, N_{BR}, and I_C) and skeletal variations and complexity (^{2}CIC) and the reactivity described through the energy partition term ($ER_{Max,C-H}$, #TMEI). A closer inspection shows that descriptors with highest R^2 are related to the bulk properties of the compounds and they, together with the hydrophobicity property, are major contributors in the whole data set's genotoxicity. They can influence the transportation process of mutagenic compounds to the active site, particularly the penetration through the bio-membranes. This is in well accordance with the scientific conclusion of other papers treating same topics[42].

The mutagenic action refers to the interactions of the molecule with the cell reproductory system. According to general mechanistic concepts, the interactions include the non-specific interactions and specific interactions. The former usually determines the solubility of a compound in the cell environment, penetration through the bio-membranes and hydrophobicity, while the latter involves specific interactions with certain mutagenic sites (electrophilic, nucleophilic or alkylation) of the molecule (parent mutagenic compound or its metabolite)[51]. Therefore, several possible mechanisms of toxic action can be involved even in the case of small group of similar compounds. In our present study, the non-specific interactions are represented mainly by the bulk properties of the compound whereas the specific interactions are modeled by the electronic and energetic characteristics described by the charge distribution, hydrogen bonding and energy partition of the molecule. Based on above observations, it is significant that our models show the general trends in the data set and also present key molecular features that have effect to genotoxicity.

CONCLUSIONS

In this study, the widely used CODESSA program was used to calculate the structural descriptors for the representation of chemicals. Using these descriptors, we performed series of computation, i.e. principal component analysis for the whole data, subset selection, construction and validation of QSAR models.

Inspection of two models mentioned in ref.[15] and linear models based on subsets generated by Duplex algorithm and PCA procedure in our present study leads to following conclusions. First, and what the most important is,

the results highlighted the use of subset selection in constructing a good, general QSAR model. The quality of the prediction depends on the data set used to develop the model and similarity between the training set and the test set can affect the predictive ability of the models dramatically. Second, the unbalance of the components in two subsets (i.e. the training set and the test set) contributed much to the big difference between fit ability and predictive ability of two models in literature, rather than the over-fitting supported by the authors[15]. In other words, compounds in the test set are underrepresented in these models, the chemical 'space' – or extent of chemical diversity represented by the molecules used to construct model - is enriched with rather different structures, as shown intuitively in PCs score plot. Finally, comparison of linear models based on two subsets generated using Duplex algorithm and PCA method indicates that the latter seems to be a better way to select the representative training set and test set in this work.

With respect to the subsets designed by PCA process, the optimum linear model contains six descriptors which were selected by BMLR method in CODESSA program. Mechanistic interpretation shows that these descriptors have specific physico-chemical meaning and have effect to the mutagenic action of aromatic and heteroaromatic amines. They mainly encode structural information related to the size and shape, hydrogen bonding, and the energy partition ability of the compounds, respectively. The size and shape of a molecule and the hydrophobicity property are major contributors in the whole data set's genotoxicity. The same six descriptors were also used as inputs to perform nonlinear QSAR studies using SVM and PPR. All models were cross-validated, and their predictive powers were evaluated on a test set and an independent validation set. After investigating the results of models constructed by using BMLR, linear SVM, radial basis function SVM and PPR method, we concluded that: (i) the six descriptors work together in a complex way to correlate with the mutation potency of compounds, (ii) there do exist the nonlinearity among the data set, as is very common in toxicity topics, and (iii) generally, nonlinear models produced better results than linear models and PPR model outperforms the rest in the following order: PPR > SVM > linear SVM ≥ BMLR, especially for the test set. Prediction for the independent validation set of 20 compounds was good and thus proved the best generalization ability of PPR model. In addition, PPR is computationally quite feasible and simple and it takes much less time to convergence compared to SVM model. Considering these, there is little doubt that utilizing PPR as modeling technique to predict genotoxicity is a good choice. In addition, this work also provides a new idea

and an alternative method to investigate the genotoxicity of the similar structures with aromatic amines, and can be extended to other toxicity studies.

This work was supported by Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT0966). The authors thank the R Development Core Team for affording the free R 2.0 software. We also thank Dr. Yuming Song for literature support. Our thanks also due to the reviewers for the careful reading and highly constructive comments on the manuscript.

REFERENCES

1. Purves D., Harvey C., Tweats D., Lumley C. E.: *Mutagenesis* **1995**, *10*, 297.
2. Brusick D.: *Principles of Genetic Toxicology*, 2nd ed.. Plenum Press, New York 1987.
3. http://www.oecd.org/document/23/0,2340,en_2649_34365_33957015_1_1_1_1,00.html, 2009.
4. Cariello N. F., Wilson J. D., Britt B. H., Wedd D. J., Burlinson B., Gombar V.: *Mutagenesis* **2002**, *17*, 321.
5. Sanderson D. M., Earnshaw C. G.: *Hum. Exp. Toxicol.* **1991**, *10*, 261.
6. Karelson M., Sild S., Maran U.: *Mol. Simul.* **2000**, *24*, 229.
7. Valkova I., Vračko M., Basak S. C.: *Anal. Chim. Acta* **2004**, *509*, 179.
8. Ren S. J.: *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1679.
9. Kulkarni S. A., Zhu J. P. in: *Ecotoxicology Modeling, Emerging Topics in Ecotoxicology: Principles, Approaches and Perspectives* (J. Devillers, Ed), pp. 245–261. L. R. Shugart and Associates, Oak Ridge (TN) 2009.
10. Debnath A. K., Debnath G., Shusterman A. J., Hansch C.: *Environ. Mol. Mutagen.* **1992**, *19*, 37.
11. Basak S. C., Gute B. D., Grunwald G. D. in: *Quantitative Structure–Activity Relationships in Environmental Sciences VII.* (F. Chen and G. Schuurmann, Eds), pp. 245–261. SETAC Press, Pensacola (FL) 1998.
12. Basak S. C., Mills D., Balaban A. T., Gute B. D.: *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 671.
13. Maran U., Karleson M., Katritzky A. R.: *Quant. Struct.–Act. Relat.* **1999**, *18*, 3.
14. Gramatica P., Connsonni V., Pavan M.: *SAR QSAR Environ. Res.* **2003**, *14*, 237.
15. Cash G. G., Anderson B., Mayo K., Bogaczyk S., Tunkel J.: *Mutat. Res.* **2005**, *585*, 170.
16. Benigni R., Giuliani A., Franke R., Gruska A.: *Chem. Rev.* **2000**, *100*, 3697.
17. Benigni R.: *Chem. Rev.* **2005**, *105*, 1767.
18. Vračko M.: *Top Heterocycl. Chem.* **2006**, *4*, 85.
19. Gramatica P. in: *Recent Advances in QSAR Studies* (T. Puzyn, J. Leszczynski and M. T. Cronin, Eds), pp. 327–366. Springer, Berlin 2010.
20. Zefirov N. S., Palyulin V. A.: *J. Chem. Inf. Comput Sci.* **2001**, *41*, 1022.
21. Tropsha A., Gramatica P., Gombar V. K.: *QSAR Comb. Sci.* **2003**, *22*, 69.
22. Gramatica P., Papa E.: *Environ. Sci. Technol.* **2007**, *41*, 2833.
23. Cronin M. T. D., Schultz T. W.: *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 39.
24. Guha R., Serra J. R., Jurs P. C.: *J. Mol. Graph. Model* **2004**, *23*, 1.
25. Katritzky A. R., Lobanov V. S., Karelson M.: *CODESSA*, Version 2.0, Reference Manual, 1995–1997.

26. *HyperChem*, Release 6.0 for Windows. Hypercube, Inc. 2000.
27. Russell C. J., Dixon S. L., Jurs P. C.: *Anal. Chem.* **1992**, *64*, 1350.
28. Katritzky A. R., Perumal S., Petrukhin R., Kleinpeter E.: *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 569.
29. Smola A. J., Schölkopf B.: *A Tutorial on Support Vector Regression.* NeuroCOLT2 Technical Report Series, NC2-TR-1998-030.
30. Friedman J. H., Tukey J. W.: *IEEE Trans. Comput.* **1974**, *C-23*, 881.
31. Friedman J. H.: *J. Am. Stat. Assoc.* **1987**, *82*, 249.
32. Sun J. Y.: *SIAM J. Sci. Statist. Comput.* **1993**, *14*, 68.
33. Sharma S.: *Applied Multivariate Techniques*. John Wiley & Sons, Singapore 1996.
34. De Maesschalck R., Estienne F., Verdú-Andrés J., Candolfi A., Centner V., Despagne F., Jouan-Rimbaud D., Walczak B., Massart D. L., de Jong S., de Noord O. E., Puel C., Vandeginste B. M. G.: *Internet J. Chem.* **1999**, *2*, 19.
35. Ren Y. Y., Liu H. X., Li S. Y., Yao X. J., Liu M. C.: *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2474.
36. Ren Y. Y., Qin J., Liu H. X., Yao X. J., Liu M. C.: *QSAR Comb. Sci.* **2009**, *28*, 1237.
37. Golbraikh A., Shen M., Xiao Z., Xiao Y. D., Lee K. H., Tropsha A.: *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241.
38. Katritzky A. R., Kuanar M., Fara D. C., Karelson M., Acree W. E.: *J. Bioorg. Med. Chem.* **2004**, *12*, 4735.
39. Schultz T. W., Cronin M. T. D.: *Environ. Toxicol. Chem.* **2003**, *22*, 599.
40. Eriksson L., Jaworska J., Worth A. P., Cronin M. T. D., Mcdowell R. M., Gramatica P.: *Environ. Health Perspect.* **2003**, *111*, 1361.
41. Geman S., Bienenstock E., Doursat R.: *Neural Comput.* **1992**, *4*, 1.
42. Maran U., Sild S.: *Artif. Intell. Rev.* **2003**, *20*, 13.
43. Hatch F. T., Colvin M. E.: *Mutat. Res.* **1997**, *376*, 87.
44. Basak S. C., Harriss D. K., Magnuson V. R.: *J. Pharm. Sci.* **1984**, *73*, 429.
45. Karelson M., Lobanov V. S., Katritzky A. R.: *Chem. Rev.* **1996**, *96*, 1027.
46. Klein M., Voigtmann U., Haack T., Erdinger L., Boche G.: *Mutat. Res.* **2000**, *467*, 55.
47. Glende C., Schmitt H., Erdinger L., Engelhardt G., Boche G.: *Mutat. Res.* **2001**, *498*, 19.
48. Glende C., Klein M., Schmitt H., Erdinger L., Boche G.: *Mutat. Res.* **2002**, *515*, 15.
49. Pullman A., Pullman B.: *Adv. Cancer Res.* **1955**, *3*, 117.
50. Weast R. C. (Ed.): *Handbook of Chemistry and Physics*, p. F-112. CRC Press, Cleveland (OH) 1974.
51. Miller J. A., Miller E. C. in: *Origins of Human Concerns* (H. H. Hiatt, J. D. Watson and J. A. Winsten, Eds), pp. 605–627. Laboratory Press, Cold Spring Harbor (NY) 1997.