QSAR Modeling of Carcinogenic Risk Using Discriminant Analysis and Topological Molecular Descriptors

Joseph F. Contrera*, Philip MacLaughlin[a], Lowell H. Hall[b] and Lemont B. Kier[c]

Center for Drug Evaluation and Research, Office of Pharmaceutical Science, U. S. Food and Drug Administration, Rockville, MD 20857; [a]MDL Information Systems, 200 Wheeler Road, Burlington, MA 01803; [b]Department of Chemistry, Eastern Nazarene College, Quincy, MA 02170; [c]Department of Medicinal Chemistry, School of Pharmacy, Virginia, Commonwealth University, Richmond, VA 23298, USA

Abstract: A discriminant analysis model is presented for carcinogenic risk. The data set is obtained from the two-year rodent study FDA/CDER database and was divided into a training set of 1022 organic compounds and an external validation test set of 50 compounds. The model is designed to use as a decision support tool for a defined decision threshold, and is thus a binary discrimination into "high risk" and "low risk" categories. The carcinogenic risk classification is based on the method for estimating human risk from two-year rodent studies developed at the FDA/CDER/ICSAS. The paradigm chosen for this model allows a straightforward risk analysis based on historic information, as well as the computation of coverage, probability and confidence metrics that can further qualify the computed result. The molecular structures were represented as MDL mol files. The molecular structure information was obtained as topological structure descriptors, including atom-type and group-type E-State and hydrogen E-State indices, molecular connectivity chi indices, topological polarity, and counts of molecular features. The MDL®QSAR software computed all these descriptors. Furthermore, the discriminant analyses were all performed with the MDL®QSAR software. The reported model is based on fifty-three descriptors, using the nonparametric normal kernel method and the Mahalanobis distance to determine proximity. The model performed very well on the fifty compounds of the test set, yielding the following statistics: 76% correctly classified "high risk" (carcinogenic) and 84% correctly classified as "low risk" (non-carcinogenic).

Keywords: Carcinogenicity, discriminant analysis, *in silico*, predictive toxicology, topological structure descriptors, QSAR, e-state, chemoinformatics.

I. INTRODUCTION AND BACKGROUND

Rodent carcinogenicity studies are required for the marketing of most chronically administered drugs. These studies are the most costly and time-consuming non-clinical regulatory testing requirement in the development of a drug. The cost is approximately $2 million for a rat and mouse study, requiring 2 years of treatment, and at least an additional 1-2 years for histopathological analysis and report writing. The human carcinogenic potential of a compound is a property that cannot be evaluated in clinical trials and therefore safety decisions are made mainly on the basis of animal study results and risk/benefit considerations. The results of rodent carcinogenicity studies can have considerable impact on drug approvability. Even when rodent carcinogenicity findings do not prevent marketing, they can seriously restrict the marketing of some products or reduce their competitive advantage. Rodent carcinogenicity studies are usually initiated relatively late in drug development when considerable resources and have already been invested in a potential new product. Significant carcinogenic findings at this stage of drug development can have disastrous and costly consequences for both the sponsoring company and the regulatory agency in the form of additional review cycles and time and effort invested in failed applications. Predictive modeling can reduce the likelihood of developing a compound that produces significant rodent tumors and can, therefore, lead to significant savings for both the pharmaceutical industry and the regulatory agency. The rodent carcinogenicity bioassay is also a pivotal component of food safety and environmental regulatory policy.

Ready access to scientific knowledge is critical to support safety-related regulatory and product development decisions, particularly in situations where available experimental information is inadequate or unavailable, to identify information gaps, and to prioritize research. A current challenge is the development of better means to identify useful relationships and insights from large sets of data. Based on the major advances in computer technology, chemoinformatics, and predictive toxicology, the accumulated results of rodent carcinogenicity studies in public databases and FDA files can be more effectively used to improve the scientific basis of regulatory and product development decisions and reduce the use of animals in testing. It is conceivable that over time with increased experience and confidence in carcinogenicity predictive software, it may be possible to reduce carcinogenicity testing for compounds that have molecular structures that are highly represented in the carcinogenicity database. This process

* Address correspondence to this author at the Center for Drug Evaluation and Research, Office of Pharmaceutical Science, U. S. Food and Drug Administration, Rockville, MD 20857, USA; Tel: 301-827-5188; Fax: 301-827-3787; E-mail: contrerajf@cder.fda.gov

1570-1638/05 $50.00+.00

would reduce unnecessary testing and also free resources for testing compounds that are truly new molecular entities and are poorly represented in the carcinogenicity database.

In recent years methods have been developed for grouping together molecules with similar molecular structures, based on the use of topological structure information. Over the past decade there has been a significant growth in the use of similarity-based searching of databases in drug design and these methods have been shown to have a broader application [1-4]. The objective has been to organize a database of molecules according to a set of structure criteria so that compounds can be identified as being similar to a reference or target molecule. These similar compounds become candidates for screening or further analysis in the design process. The rationale is that compounds that are similar to a reference molecule are likely to be related to the behavior of the reference molecule in some sense. With the growth of combinatorial chemistry, the compounds in a database may be entirely or partially virtual; in other words, they are synthesized *in silico*. As a result, there may be no property value information with the molecules; hence, similarity is based entirely on the structural descriptions chosen in a particular study. There is thus no useful way of evaluating similarity based on physical properties except by virtue of the future success of the drug design project employing this general method.

Lajiness has shown quite clearly that a random search through a list of molecules is inferior to a search through an organized database, based on its ability to generate similarity or diversity in a study [3]. Some form of encoding structure information should be present for meaningful exploitation of a database. The code of structure information thus becomes the metric to evaluate similarity or its complement, diversity. This approach is not an exercise in multi-parameter QSAR modeling. With virtual molecules, many or all of the property values are unknown. The search is conducted by selecting a set of descriptors deemed important and finding the relation of molecules relative to a reference molecule using a metric such as distance or a grouping such as nearest neighbors. The objective is to create a cluster of molecules of potential interest based on several structure indices. Interesting compounds may appear that can be selected for screening or for further applications in the database search process.

The encoding and subsequent searching can be a browsing process, using electrotopological state indices (E-State) values or other information-rich indices, such as molecular connectivity, removing the need for carefully delineated structural features which may be unknown or which can severely limit diversity. The choice of limiting distance values among molecules in the database makes it possible to reduce the number of output molecules. A qualitative advantage of this process is the stimulation of the chemist's imagination [5].

A large number of descriptors are available to be employed in the organization of a database. It is not our intention here to create a list of these or to make comparisons, each method being suitable for different circumstances. Our intention however is to build on the use of atom-type E-State descriptors along with molecular connectivity indices as a systematic organizer of a database, imparting a rich information source that can produce potentially useful structure patterns [6-8]. This present study requires that the structure representation employed be able to organize the data set molecules in such a way that those with high potential for a particular property (i.e. carcinogenic risk) be more closely associated with each other than with those that are associated with another property such as low carcinogenic risk. Previous work appears to indicate such a possibility [6-9]. The ability of the simple molecular connectivity indices to organize a set of skeletal structures has been demonstrated [9]. Molecular skeletons are grouped in a meaningful manner. Furthermore directions within the representation space have meaning in terms of significant chemical information such as degree of skeletal branching, adjacency of branch points, and number of rings and types of fused ring systems. The atom-type E-State structure descriptors have also been shown to organize molecular structures in a chemically meaningful manner, emphasizing electronic information [10,11]. Based on the structure space provided by the atom type E-State descriptors, excellent similarity searches through a chemical database have been reported [7,8,10]. This combination of structure information representations provided the basis for the use of structure similarity methods together with topological descriptors that have recently been applied to QSAR modeling of rodent carcinogenicity [12].

The result of these investigations indicates that the use of the atom-type E-State descriptors together with the molecular connectivity chi indices provides a structure space in which molecular structures are organized in chemically meaningful ways so that carcinogenic properties associated with those structures can also be expected to be usefully organized. As a result, statistical methods of analysis can be successfully applied to a data set based on E-State and molecular connectivity descriptors, as is demonstrated in this present work.

II. EXPERIMENTAL DATA AND METHODS

The FDA/CDER Rodent Carcinogenicity Database

The FDA/CDER Rodent Carcinogenicity database was created from summary rat and mouse carcinogenicity study findings for over 1300 compounds that include both industrial chemicals and pharmaceuticals. Rodent carcinogenicity study results in the FDA database were obtained from the National Toxicology Program (NTP) rodent carcinogenicity database, the Lois Gold Carcinogen Potency (CPD) Database [13], FDA/CDER archives, and the scientific literature. The database includes the name and identification codes, the chemical structure represented as an MDL MOL file, and numeric carcinogenic activity units assigned (discussed below) to each compound.

Acceptance Criteria for Carcinogenicity Studies

Most of the carcinogenicity study results for pharmaceuticals in the FDA database were derived from pharmacology/toxicology and biostatistics reviews and reports in FDA files. The results of carcinogenicity studies in FDA new drug application (NDA) regulatory reviews for marketed products are available under the Freedom of

Information Act and are considered non-proprietary. The identity of pharmaceuticals currently under regulatory review as an investigational new drug application (IND) or new drug application (NDA) or drugs that have never been marketed are proprietary and cannot be disclosed without the consent of the sponsor. Proprietary compounds represent approximately 8 % of the total number of carcinogenicity studies in the FDA/CDER database and are coded in this report.

Carcinogenicity Study Design and Analysis

The design of rodent carcinogenicity studies for pharmaceuticals is essentially the same as the design employed for industrial and environmental chemicals and U.S. National Toxicology Program (NTP) rodent carcinogenicity studies. Male and female rats and mice are divided randomly into one or two control and three treatment groups of 50-70 animals per group per species. Historically, the highest dose in the studies analyzed generally approximates the maximum tolerated dose (MTD) in the test species, and is administered daily, usually in the feed or by oral gavage for 2 years. The rodent strains most often used in NTP studies is the inbred Fisher 344 rat and the hybrid B6C3F1 (C3H x C57B16) mouse. In pharmaceutical studies submitted to the FDA, the predominant rodent strains are the Sprague-Dawley derived CD rat, and the CD-1 Swiss-Webster derived mouse. Despite the long experience in the FDA with these assays, the significance of tumors from life-time exposure at the maximum tolerated dose, the dose response extrapolation and the relevance of rodent tumors to humans continue to be highly controversial issues.

Classification and Stratification of Rodent Tumor Findings

In studies reviewed by FDA/CDER, tumor findings are classified as positive if either benign and/or malignant findings are statistically significant in pair-wise comparison to concurrent controls by Fisher's Exact Test or equivalent statistical analysis. An adjustment for rare and common events is also applied to tumor findings [14]. Tumors are considered significant if they attained a level of $p \leq 0.01$ for common tumors and $p \leq 0.05$ for rare tumor types. Rare tumors are those with a spontaneous background incidence rate equal to or less than a 1%. The incidence of benign and malignant tumors (adenomas and carcinomas) are combined and statistically evaluated where appropriate [15].

Data Transformation: The Numeric Representation of Carcinogenic Activity

Carcinogenicity studies are generally carried out in male and female rats and mice. Each sex/species is considered an individual study cell and therefore a complete battery of carcinogenicity studies for a compound is comprised of 4 study cells. A simplified numerical activity scale was used to quantify and stratify the results of rodent carcinogenicity studies. Compounds that produce statistically significant (by pair-wise comparison) tumors at multiple organ/tissue sites in a study cell were assigned the highest activity value of 50. Compounds that produce statistically significant single site tumors received an activity value of 40 and weaker or equivocal single site responses were assigned an activity value of 30. Studies with no statistically significant treatment-related tumor findings were assigned an activity value of 10. Compounds with 30 or more activity units in 2 or more study cells (2Plus) , that is, having activity that crossed the biological barrier of gender or species, were classified as high risk carcinogens. Compounds with less than 30 activity units in 3 or more study cells were considered not to be high risk carcinogens. Compounds that were tested only in the rat or mouse may also be considered positive if there were significant tumor findings in both males and females. Compounds tested only in one species that have no tumor findings cannot be considered negative without additional information from at least one other study cell. Applying these rules, a training carcinogenicity database was created containing 1022 compounds with 4 cell or equivalent data of which 649 compounds were classified as carcinogenic (High Risk), having tumor findings in at least 2 study cells, and 373 compounds were non-carcinogenic (Low Risk) with negative findings in 3 or more study cells. The greater number of positive compounds is partly a function of the scoring method employed. This scoring method is the same as that used to predict rodent carcinogenicity based on molecular similarity [12] and is a simplification of the multi-cell method used for MCASE-ES rodent carcinogenicity predictions [16].

The name and structure of proprietary compounds were coded and kept confidential by the FDA. Electrotopological descriptors derived from proprietary molecules were included in the training data set. Although electrotopological state and other topological descriptors employed contain sufficient information for successful modeling they are insufficient to unambiguously recreate a proprietary molecular structure.

A validation experiment employing a total of 50 test compounds that were not part of the MDL®QSAR (see below) control or training data set were used in this investigation. The 50 test compounds were randomly removed from the 1072 compound rodent carcinogenicity training set. The carcinogenicity model was based on the remaining 1022 training set compounds. The 50 randomly selected test compounds included 38 pharmaceuticals of which 9 (18%) were newer pharmaceuticals currently under regulatory review that are not yet marketed (structures and identity not disclosed) and 12 industrial chemicals. The 50 validation test compounds contained 25 “High Risk” compounds with tumor findings in two or more study cells (2Plus) and 25 “Low Risk” compounds with either no tumor findings or findings in only 1 study cell. Table **3** lists the compounds, their assigned risk level from the FDA/CDER Rodent Carcinogenicity database and the risk level as predicted from the model presented in this work.

III. COMPUTATIONAL METHODS

Descriptors and Descriptor Selection

The MDL®QSAR module implements molecular topological descriptors available within the Molconn-Z program [17a, 17b]. (A list of publications that illustrate the nature of topological descriptors and their applications is available [17c].) An initial set of 195 topological descriptors

was computed by the MDL®QSAR module for the entire training set of 1022 compounds that were tested for rodent carcinogenicity. The descriptors included atom-type, group-type, and individual atom E-State and hydrogen E-State indices, molecular connectivity chi indices, kappa shape indices, topological polarity, counts of molecular features (number of rings, number of H-bond donors and acceptors, etc), and others. This initial set was reduced using the following criteria: first, descriptors were only considered that had non-zero value for at least 95% of all compounds and, second, the variance of the descriptor values had to be no less than a certain threshold, set equal to 1.

Model Development

The compounds in the FDA rodent carcinogenicity dataset are characterized either as carcinogenic or non-carcinogenic; therefore, this dataset presents a typical example of a binary classification problem. For the analysis of such datasets, MDL®QSAR employs methods discriminant analysis. The complete description of this method analysis can be found in a number of textbooks and monographs [18,19]. Herein, we provide a short description of the method pertinent to its implementation within MDL® QSAR.

MDL® QSAR incorporates the algorithms to develop discriminant models and the graphics interface that allows users to input data sets, initiate calculations, analyze and manipulate resulting models. Each model is characterized by rich statistics available to the user. MDL®QSAR implements the entire range of discriminant analysis methods such as parametric, nonparametric kernel, and nearest-neighbor approaches. The classic parametric method of discriminant analysis is applicable in the case of approximately normal within-class distributions. The method generates either a linear discriminant function (the within-class covariance matrices are assumed to be equal) or a quadratic discriminant function (the within-class covariance matrices are assumed to be unequal). Our initial chemometric analysis of the FDA data set demonstrated that the distribution of the descriptor values did not follow the Gaussian law, which was indicated by the normal distribution hypothesis testing with the confidence level of 0.01. When the distribution is not assumed to follow a particular law or is assumed to be other than the multivariate normal distribution, nonparametric methods can be used to derive classification criteria. The nonparametric methods available within the MDL® QSAR include the kernel and k-nearest-neighbor (kNN) methods. The main types of kernels implemented in MDL® QSAR include uniform, normal, Epanechnikov, biweight, or tri-weight kernels, which are used to estimate the group specific density at each observation.

In general, either Mahalanobis or Euclidean distances can be used to determine proximity between compound-vectors in multidimensional descriptor space. When the k-nearest-neighbor method is used, the Mahalanobis distances are based on the pooled covariance matrix. When a kernel method is used, the Mahalanobis distances are based on either the individual within-group covariance matrices or the pooled covariance matrix. Either the full covariance matrix or the diagonal matrix of variances can be used to calculate the Mahalanobis distances. In our studies, various combinations of model building preferences have been explored to achieve a model with the highest accuracy. The performance of each candidate model was assessed by making use of the prediction error rate in the training set (i.e., probabilities of misclassification). For each model, the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates were studied both for re-substitution analysis as well as those resulting from leave-one-out cross-validation. The computation of each model studied took several seconds on a Pentium 4 processor with 2.8 MHz and 1 Gb memory.

IV. COMPUTATIONAL STUDIES

The task of finding the best model falls into two interconnected parts: the search for the best subset of descriptors and the selection of the type and optimal parameters of the model. Both these subtasks admit no formal algorithmic solution and require some experimentation to achieve the best solution. In our studies, more than 3000 discriminant analysis models were built in total, using different criteria for descriptor reduction and various parameters of discriminant analysis as described in the Methods Section. The best model included 53 variables (Table **1**) and was characterized by the following parameters: normal kernel; smoothing parameter of 2; Mahalanobis distance to determine proximity; distance calculations based on the full individual within-group covariance matrices; prior probabilities set to 0.5.

The correct classification rates for the best discriminant model, which is supplied with MDL® QSAR, are shown in Table **2**. Total accuracy of the model, both for C re-substitution and LOO cross-validation analyses is shown as well as separate data for the test set prediction of carcinogenic (high risk carcinogens) as well as non-carcinogenic (low risk carcinogen) compounds.

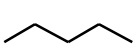
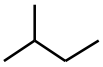
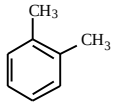
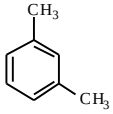
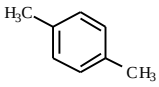
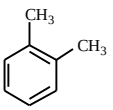
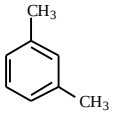
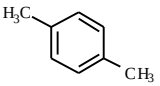
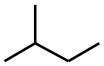
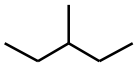
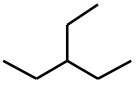
V. DISCUSSION

Risk Mitigation

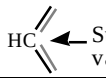
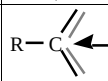
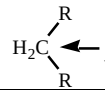
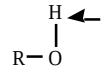
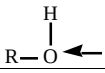
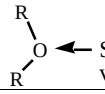
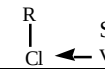
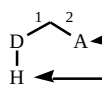
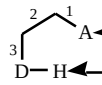
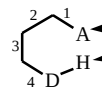
The modeling approach in this study was chosen specifically for a risk analysis approach. Unexpected results in long-term carcinogenicity bioassays on a new drug candidate can be extremely costly in time, money, and market viability. With today's technology, applicants must carry the risk of long-term carcinogenicity well into phase 2-3 clinical development with little or no mitigation. At this stage of development, even a single failure can result in a huge loss: the late-stage non-approval of a drug can mean the loss of $700 million or more [20], and represent the loss of six to eight years of development effort.

The chosen modeling paradigm, in its selection of chemical structure variable type, its identification of a real world risk threshold in endpoint definition, and its straightforward statistical method, can be called 'actuarial' in approach. This allows the user of the model to view various confidence and applicability measures and restrict acceptable ranges as desired. Two such calculated metrics are the Distance and the Probability of Membership in Class. The

Table 1. Descriptors selected in the best model for the prediction of carcinogenicity. The descriptors appear in groupings that relate to their structure information content. A ranking is also given, based on descending order of F-value for inclusion of the descriptor in the model. A brief definition is given for each descriptor along with an illustration for a selected few descriptors. For more specific information on structure interpretation see the appropriate references [6-11,17d].

Index	Rank	Description
Low Order Chi Indices		Encode degree of skeletal branching and molecular size. Low order indices x0 and x1 increase with molecular size and decrease with increased branching. Chi 2 (x2) shows the greatest sensivity to differences in branching and increases with branching. Valence indices add information about heteroatoms and valence state. Simple illustration for x2 given below.
x2 =  = 1.353		 = 1.802  = 3.000  = 1.707
x0	5	Simple chi 0 index decreases minimally with increased branching, insensitive to adjacency.
x1	2	Simple chi 1 index encodes degree of branching and decreases with increased branching.
x2	9	Simple chi 2 index gives high sensitivity and increases with increasing branching.
xv0	10	Valence chi 0 index is highly intercorrelated with molecular surface area and volume.
xv1	6	Valence chi 1 is similar to x1 but also includes heteroatom and valence state information.
xv2	11	Valence chi 2 index includes heteroatom and valence state information with high sensitivity.
Path Chi Indices		Encode complexities and specifics of overall skeletal variation, including degree of branching and molecular size. Each higher order path index encodes different aspects of skeletal variation. Valence indices add information about heteroatoms and valence state. A simple illustration for xp3 is given below.
xvp3 =		 = 1.426  = 1.174  = 1.218
xp3	12	Simple chi path 3 is sensitive to adjacent branch points in the molecular skeleton.
xp4	13	Simple chi path 4 is sensitive to branch points separated by one atom in the skeleton.
xp5-8	15,17,23,25	Simple chi path 5-8 encode specific skeletal information to disciminate among skeletal classes.
xvp3	24	Valence chi path 3 is similar to xp3 with aditional heteroatom and valence state information.
xvp4	29	Valence chi path 4 is similar to xp4 with aditional heteroatom and valence state information.
xvp5	31	Valence chi path 5 is similar to xp5 with aditional heteroatom and valence state information.
Cluster & Path/Cluster Chi Indices		Encode structure information specifically based on a branch point, emphasizing the immediate branch point environment. Simple illustration for xvpc4 given below.
xvc4 =		 = 0.622  = 0.359  = 0.385
xc3	30	Simple chi cluster 3 index is defined for a single branch point and encodes the number and branching environment of such points.
xpc4	35	Simple chi path-cluster 4 index is defined for the isobutane skeleton and is espeically sensitive to adjacency of skeletal branch points.
xvpc4	48	Valence path-cluster 4 index encodes information similar to xpc4 but with heteroatom and valence state information added.
knotp	42	Knotp gives the difference between chi cluster-3 and chi path/cluster-4 descriptor. Knotp is largest where an xc3 subgaph is not associated with an xpc4 subgaph. Each path cluster-4 (xpc4) subgraph contains a cluster-3 (xc3) subgraph and one additional atom. Each xc3 subgraph may be associated with up to three of these additional atoms and thus be contained within up to 3 xpc4 subgraphs, as shown in the table below. The knotp descriptor helps to separate this overlapping structure information into distinct numerical values.
knotp =  = 0.577		 = 0.000  = - 0.289  = - 0.408

(Table 1) contd.….

Index	Rank	Description	
Atom Type E-state		Encode for a specified atom type a combination of electron accessiblity, presence/absence, and the count of the atom type in the molecule based on the electrotoplogical state indices. The atom type E-State has been shown to be very useful for similarity analysis and structure classification and modeling of various properties [6,10,11].	
SaaCH	16	Sum of the atom level E-state values for all non-substituted aromatic carbon atoms.	 Sum of atom level E-state values in molecule
SaasC	26	Sum of the atom level E-state values for all the substituted aromatic carbon atoms.	 Sum of atom level E-state values in molecule
SssCH2	27	Sum of the atom level E-state values for all the methylenes in the molecule.	 Sum of atom level E-state values in molecule
SsCH3	36	Sum of the atom level E-state values for all the carbon atoms in methyl groups in the molecule.	 Sum of atom level E-state values in molecule
SHsOH	37	Sum of the hydrogen atom level E-state values for all the hydrogen atoms in OH groups.	 Sum of hydrogen atom level E-state values
SsOH	38	Sum of the atom level E-state values for all the oxygen atoms in OH groups in the molecule.	 Sum of atom level E-state values in molecule
SssO	46	Sum of the atom level E-state values for all the ether oxygen atoms in the molecule.	 Sum of atom level E-state values in molecule
SsNH2	47	Sum of the atom level E-state values for all the nitrogen atoms in primary amines.	 Sum of atom level E-state values in molecule
SsCl	49	Sum of the atom level E-state values for all the chlorine atoms in the molecule.	 Sum of atom level E-state values in molecule
SdO	50	Sum of the atom level E-state values for all the double bonded oxygen atoms in the molecule.	 Sum of atom level E-state values in molecule
Atom Type Count			
SaaCH_acnt	18	Number (count) of all non-substituted aromatic carbon atoms in the molecule.	
SssCH2_acnt	20	Number (count) of methylene groups in the molecule	
SaasC_acnt	22	Number (count) of substituted aromatic carbon atoms in the molecule	
SdssC_acnt	34	Number (count) of =C< groups in the molecule	
SsCH3_acnt	40	Number (count) of methyl groups in the molecule	
SdO_acnt	53	Number (count) of double bonded oxygen molecules in the molecule	
Internal Hydrogen Bonding E-state		The largest single product of E-state and H E-state values from all acceptor and donor pairs separated by 4 skeletal bonds and not part of a rigid skeletal structure.	
SHBint2	41	Donor acceptor pair do not form an internal hydrogen bond. This group is associated with acids, amides, etc.	 Prduct of the E-state and HE-state values
SHBint3	44	Forms 5-membered ring for potential internal H bond.	 Prduct of the E-state and HE-state values
SHBint4	51	Forms 6-membered ring for potential internal H bond.	 Prduct of the E-state and HE-state values

(Table 1) contd.…..

Index	Rank	Description
Molecular Properties		A group of structure descriptors that encode a general aspect of structure information for the whole molecule.
Qs	1	A whole molecule polarity index that decreases in value as the polarity increases and more sensitive to molecular size.
nvx	3	Number of graph verticies (non-hydrogen atoms) in the moleclue.
SHother	4	Sum of the Hydrogen E-state values for hydrogens on carbon atoms.
Hmax	7	The maximum hydrogen atom level E-state value in a molecule.
Hmaxpos	8	The maximum positive hydrogen atom level E-state value in a molecule.
fw	14	Formula weight.
Nrings	19	Number (count) of (independent) rings in the molecule.
Phia	21	A kappa shape molecular flexibility descriptor that increases with homologation and decreases with increased branching or cyclicity. Larger Phia values indicate greater molecular flexibility.
Gmax	28	The maximum atom level E-state value in a molecule.
SHHBd	32	Sum of the hydrogen atom level E-state values for all hydrogens bonded to donating atoms.
numHBd	33	Number (count) of hydrogen bond donors in the molecule.
numHBa	39	Number (count) of hydrogen bond acceptors in the molecule.
nelem	43	Number (count) of chemical elements in the molecule.
ncirc	45	Number (count) of graph circuits in the molecule.
Qv	52	A whole molecule E-state polarity index that decreases in value as the polarity increases.

Table 2. Carcinogenic Risk Prediction Accuracy

Training set, 1022 compounds			
Resubstitution	Carcinogenic	Non-Carcinogenic	Total
	0.96	0.97	0.96
LOO Cross-validation	0.73	0.68	0.70
Test set, 50 compounds			
Prediction	Carcinogenic	Non-Carcinogenic	Total
	0.76	0.84	0.80

calculated Distance shows whether the subject compound vector is adequately represented within the historic variance of chemical structure descriptors. The Probability of Membership in Class is a measure of how well the historic knowledge is able to discriminate high risk compounds from low risk compounds within the nearest space of the subject compound vector.

Probability of Membership in Class

The results for prediction are given in Table **3** along with the original rodent data. The prediction rates are 84% correct for low risk and 76 % for high risk with an overall rate of 80% correct. Incorrect predictions are marked in bold. In this present study, we found that by placing limits on probability in class we could trade overall coverage for accuracy. By placing a minimum of 60% Probability-in-Class on our 50 compound test set, overall coverage was reduced to 76%, Sensitivity rose from 76% to 87%, Specificity remained essentially constant at 84% (83%), and Concordance improved from 80% to 84% (See Table **4**). By placing a minimum of 65% Probability in Class, Coverage was 70%, Sensitivity 93%, Specificity 86% and Concordance 89% (See Table **5**). Exercising this option allows a flexibility in how the model is employed, perhaps allowing a wider range of acceptable probability in screening large compound libraries to glean general characteristics, while restricting this range when assessing safety risks in lead compounds for better confidence.

Distance Measure

MDL®QSAR evaluates two quantitative measures of applicability of data models to new observations: 1. regression

Table 3. List of Compounds used in the Validation Test Set along with the rodent data. Compounds showing only a number for "Name" are currently confidential at the FDA because they are under regulatory review. Predicted values are shown in bold for incorrect prediction. Predictions shown here are made without regard to probability level. See Tables 4 and 5 for tabulations based on selected probability ranges. (See text for details). See Experimental Data and Methods in text, pages 5-8, for detailed description of the data fields in Tables 3, 4 & 5.

Name	Male Rat	Female Rat	Male Mouse	Female Mouse	2Plus	Class	Predicted
Bupropion	10	10	10	10	neg	'LOW'	'LOW'
C.I. Orange10	10	10	10	10	neg	'LOW'	**'HIGH'**
Citral	10	10	10	10	neg	'LOW'	**'HIGH'**
Dimethylaminino 4,4	10	10	10	10	neg	'LOW'	'LOW'
Etoricoxib	30	10	10	10	neg	'LOW'	'LOW'
Ezetimbe (Zeita)	10	10	10	10	neg	'LOW'	'LOW'
Famciclovir	10	40	10	10	neg	'LOW'	'LOW'
Glipizide	10	10	10	10	neg	'LOW'	'LOW'
Glyburide	10	10	10	10	neg	'LOW'	'LOW'
Indapamine	10	10	10	10	neg	'LOW'	'LOW'
Ketoconazole	10	10	10	10	neg	'LOW'	'LOW'
Nadolol	10	10	10	10	neg	'LOW'	'LOW'
Nuviva	10	10	10	10	neg	'LOW'	'LOW'
Oxaprocin	10	10	40	10	neg	'LOW'	'LOW'
p-nitrotoluene	10	40	10	10	neg	'LOW'	**'HIGH'**
Ranitidine	10	10	10	10	neg	'LOW'	'LOW'
Rifampin	10	10	10	40	neg	'LOW'	'LOW'
Scopolamine	10	10	10	10	neg	'LOW'	**'HIGH'**
Spirapril	10	10	10	10	neg	'LOW'	'LOW'
212904	10	10	10	10	neg	'LOW'	'LOW'
915	10	30	10	10	neg	'LOW'	'LOW'
215558	10	10	10	10	neg	'LOW'	'LOW'
938	10	10	10	10	neg	'LOW'	'LOW'
931	10	10	10	10	neg	'LOW'	'LOW'
212156	10	10	10	30	neg	'LOW'	'LOW'
Adapalene	40	30	10	10	pos	'HIGH	'HIGH'
Anthroquinone	10	40	40	40	pos	'HIGH	'HIGH'
Benzyl acetate	40	10	50	50	pos	'HIGH'	**'LOW'**
Cabergoline	40	30	10	10	pos	'HIGH'	**'LOW'**
Carmustine	50	50	50	50	pos	'HIGH'	'HIGH'
Chlorendic acid	50	40	40	10	pos	'HIGH'	'HIGH'
Chlorodibromomethane	10	10	30	40	pos	'HIGH'	'HIGH'
Cimetidine	50	40	40	40	pos	'HIGH'	**'LOW'**
Cytembena	40	50	10	10	pos	'HIGH'	'HIGH'
Diethanolamine	10	10	40	40	pos	'HIGH'	**'LOW'**
Isoniazid	50	40	50	50	pos	'HIGH'	'HIGH'
Mestranol	40	40	50	50	pos	'HIGH'	'HIGH'
Methyleugenol	50	50	50	50	pos	'HIGH'	'HIGH'
Methythiouracil	40	50	50	50	pos	'HIGH'	'HIGH'
Metronidazole	50	50	50	50	pos	'HIGH'	'HIGH'
Nitrophenylenediamine	40	30	40	30	pos	'HIGH'	'HIGH'
Nitrososarcosine, N-	40	40	50	50	pos	'HIGH'	'HIGH'
Phenobarbital	40	40	50	50	pos	'HIGH'	**'LOW'**
Riddelline	50	40	40	40	pos	'HIGH'	'HIGH'
989	10	30	30	30	pos	'HIGH'	'HIGH'
987	50	50	50	50	pos	'HIGH'	'HIGH'
Streptozocin	50	50	50	50	pos	'HIGH'	'HIGH'
936	30	10	10	30	pos	'HIGH'	'HIGH'
935	40	40	40	40	pos	'HIGH'	**'LOW'**
711	30	30	10	10	pos	'HIGH'	'HIGH'

Table 4. List of test compounds together with the posterior probability for classification based on a 60% probabilit-in-class dividing line (See text). Compounds showing only a number for "molecule" are currently confidential at the FDA because they are under regulatory review. Compounds 'not covered' using this threshold are outlined in grey.

Posterior Probability of Membership in Class				
Molecule	**Experimental**	**Predicted**	**High**	**Low**
Rifampin	Low	Low	0	1
935	**High**	Low	0	1
Nuviva	Low	Low	0.0048	0.9952
Glyburide	Low	Low	0.0546	0.9454
Ketoconazole	Low	Low	0.0592	0.9408
Glipizide	Low	Low	0.0663	0.9337
Spirapril	Low	Low	0.0797	0.9203
Ranitidine	Low	Low	0.1158	0.8842
Dimethylaminino 4,4	Low	Low	0.1704	0.8296
Nadolol	Low	Low	0.2151	0.7849
215558	Low	Low	0.2330	0.7670
Indapamine	Low	Low	0.2435	0.7565
931	Low	Low	0.2711	0.7289
Oxaprocin	Low	Low	0.2917	0.7084
915	Low	Low	0.3022	0.6978
Ezetimbe (Zeita)	Low	Low	0.3031	0.6969
Buspirone	Low	Low	0.3048	0.6952
212904	Low	Low	0.3330	0.6670
Famciclovir	Low	Low	0.3439	0.6561
938	Low	Low	0.3832	0.6168
Cabergoline	**High**	Low	0.3964	0.6036
Cimetidine	NA	NA	0.4311	0.5689
Diethanolamine	NA	NA	0.4341	0.5659
Etoricoxib	NA	NA	0.4440	0.5561
212156	NA	NA	0.4611	0.5389
Benzyl acetate	NA	NA	0.4941	0.5059
Phenobarbital	NA	NA	0.4948	0.5052
Cytembena	NA	NA	0.5095	0.4905
Anthroquinone	NA	NA	0.5357	0.4643
Nitrososarcosine, N-	NA	NA	0.5658	0.4343
Adapalene	NA	NA	0.5809	0.4191
Metronidazole	NA	NA	0.5841	0.4159
Nitrophenylenediamine	NA	NA	0.5931	0.4069
p-nitrotoluene	**Low**	High	0.6056	0.3944
936	High	High	0.6561	0.3439
Isoniazid	High	High	0.6963	0.3037
Citral	**Low**	High	0.7102	0.2899
Methyleugenol	High	High	0.7781	0.2213
711	High	High	0.7923	0.2077
Streptozocin	High	High	0.8142	0.1859
C.I. Orange 10	**Low**	High	0.8288	0.1712
989	High	High	0.8848	0.1152
Chlorodibromomethane	High	High	0.9262	0.0738
Methythiouracil	High	High	0.9375	0.0625
Carmustine	High	High	0.9528	0.0472
Scopolamine	**Low**	High	0.9641	0.0359
Mestranol	High	High	0.9652	0.0348

(Table 4) contd.….

Molecule	Experimental	Predicted	High	Low
987	High	High	0.9745	0.0255
Riddelline	High	High	0.9991	0.0009
Chlorendic acid	High	High	0.99997	2.76E-05
60% Minimum Probability				
Coverage	76%			
Concordance	84%			
Sensitivity	87%			
Specificity	83%			

Table 5. List of test compounds together with the posterior probability for classification based on a 65% probability-in-class dividing line (See text). Compounds showing only a number for "molecule" are currently confidential at the FDA because they are under regulatory review. Compounds 'not covered' using this threshold are outlined in grey.

Posterior Probability of Membership in Class				
Molecule	**Experimental**	**Predicted**	**High**	**Low**
Rifampin	Low	Low	0	1
935	**High**	Low	0	1
Nuviva	Low	Low	0.00482	0.99518
Glyburide	Low	Low	0.05463	0.94537
Ketoconazole	Low	Low	0.05923	0.94077
Glipizide	Low	Low	0.06634	0.93367
Spirapril	Low	Low	0.07967	0.92033
Ranitidine	Low	Low	0.11577	0.88423
Dimethylaminino 4,4	Low	Low	0.17043	0.82957
Nadolol	Low	Low	0.21509	0.78491
215558	Low	Low	0.23299	0.76701
Indapamine	Low	Low	0.24346	0.75654
931	Low	Low	0.27109	0.72891
Oxaprocin	Low	Low	0.29165	0.70835
915	Low	Low	0.30219	0.69781
Ezetimbe (Zeita)	Low	Low	0.30311	0.69689
Buspirone	Low	Low	0.3048	0.6952
212904	Low	Low	0.33303	0.66697
Famciclovir	Low	Low	0.3439	0.6561
938	NA	NA	0.3832	0.6168
Cabergoline	NA	NA	0.3964	0.6036
Cimetidine	NA	NA	0.4311	0.5689
Diethanolamine	NA	NA	0.4341	0.5659
Etoricoxib	NA	NA	0.4440	0.5561
212156	NA	NA	0.4611	0.5389
Benzyl acetate	NA	NA	0.4941	0.5059
Phenobarbital	NA	NA	0.4948	0.5052
Cytembena	NA	NA	0.5095	0.4905
Anthroquinone	NA	NA	0.5357	0.4643
Nitrososarcosine, N-	NA	NA	0.5658	0.4343
Adapalene	NA	NA	0.5809	0.4191
Metronidazole	NA	NA	0.5841	0.4159
Nitrophenylenediamine	NA	NA	0.5931	0.4069
p-nitrotoluene	NA	NA	0.6056	0.3944

(Table 5) contd…..

Molecule	Experimental	Predicted	High	Low
936	High	High	0.6561	0.3439
Isoniazid	High	High	0.69634	0.30366
Citral	**Low**	High	0.71015	0.28985
Methyleugenol	High	High	0.7787	0.2213
711	High	High	0.79231	0.20769
Streptozocin	High	High	0.81415	0.18585
C.I. Orange 10	**Low**	High	0.82884	0.17116
989	High	High	0.8848	0.1152
Chlorodibromomethane	High	High	0.92624	0.07376
Methythiouracil	High	High	0.93749	0.06251
Carmustine	High	High	0.95283	0.04717
Scopolamine	**Low**	High	0.96414	0.03586
Mestranol	High	High	0.96518	0.03482
987	High	High	0.97448	0.02552
Riddelline	High	High	0.99913	0.00087
Chlorendic acid	High	High	0.99997	2.76E-05
65% Minimum Probability				
Coverage	70%			
Concordance	89%			
Sensitivity	93%			
Specificity	86%			

models and 2.discriminant analysis models. These are based on the following simple assumption: each constructed model has a certain applicability region in the space of independent variables. Specifically, if our molecule is found to exist in an observation space "far" from the set used to build the model, the prediction for the object should be treated with caution, with a less degree of confidence than in the case when the model built used objects found "nearer" to our observation space.

Let X be a n*×p matrix of data with columns being n-dimensional vectors of variables X_i and rows being p-dimensional vectors of observations x_j, m_X be the row of means,

$$X_c = \begin{pmatrix} x_1 - m_X \\ x_2 - m_X \\ \dots \\ x_n - m_X \end{pmatrix}$$

be a centered matrix of observations, and $A = (1/(n - 1))X_{\_}^{T} X_{\_}$ be a covariance matrix. A reasonable measure of proximity of a molecule to the training set in the observation space is the Mahalonobis distance, evaluated for new observation row x_{new} using formula

$$M = (x_{new} - m_X)\, A^{-1} (x_{new} - m_X)^T$$

and its special case (at A = E), common Euclidean space D:

$$D^2 = (x_{new} - m_X)\, (x_{new} - m_X)^T$$

Having chosen the distance in the observation space, one should also decide which distances are to be considered "large" or "small". In the case of regression models, it is natural to use an obvious analogy between outliers in the training set and "far-flung" observations. First note that the sum of Mahalonobis distances for all observations from the training set is p(n – 1). Consider quantity

$$d = M / (n - 1)$$

which is referred to as centered leverage value and for observations from the training set lies between 0 and 1. Based on the rules for separation of outliers in the observation space, which are recommended in regression analysis, the degree of applicability of a regression model to an object that is not a member of the training set is evaluated in MDL®QSAR as follows:

If d exceeds p/n, its average across the training set, more than twofold, one should treat the prediction for such a case with caution.

If $d > 0.5$, the degree of applicability of a model to an object is taken to be very low; if $0.2 < d \leq 0.5$ low, and if $d \leq 0.2$ we consider it optimal to use the model.

For discriminant analysis models, such as the model contained in the MDL® Carcinogenicity Module, similar methods for separation of outliers are not used. In order to partition distances into "large" and "small", an approach to data standardization is applied that is traditional for statistics. Suppose that distance D (Euclidian, see definition above) is normally distributed across the general population of data. Then, the standardized distance d_1 (the difference between the distance and its sample mean over the training set)

divided by the sample estimate of mean-square deviation, will approximately follow a normal distribution with parameters 0, 1. Thus probabilities for d_1 to fall into one or another interval of the real axis are found from the tabulated values of the Laplace function. For example, consider points $x_0 = 1.65$ and $x_1 = 2.33$ marked on the axis, such that the probability to fall to the left of them is, respectively, 0.95 and 0.99. If a new observation produces value $d_1 \geq x_1$, the degree of applicability of a model to such an object may be taken to be very low; if $x_0 \leq d_1 < x_1$ – low.

VI. CONCLUSIONS

The FDA/CDER rodent carcinogenic database provides a sound basis for development of a model for the prediction of carcinogenicity risk. The combination of the experimental data and the experience in the FDA provided the basis for qualifying 1072 compounds for the database. The MDL®QSAR software provided a useful basis for calculating the molecular descriptors and performing the discriminant analysis to establish a classification algorithm. The topological structure descriptors included electrotopological state (E-State) descriptors and molecular connectivity chi indices that have been shown to provide a sound basis for classifying molecular structures. A non-parametric method was used to obtain the final model based on the normal kernel method. Descriptors were selected by examining models with varying numbers of descriptors and deciding upon the model with the best classification statistics on the training set. The discriminant model presented here demonstrates good prediction statistics on the external validation test set of fifty compounds with sensitivity of 76% and specificity of 84% in addition to concordance of 80%. This test set includes 38 pharmaceuticals and 12 industrial chemicals. Nine of the pharmaceuticals are newer compounds that are still under regulatory review and confidential. Twenty-five of the test set compounds are considered high risk. Based on these results the model appears useful as an indicator of potential carcinogenicity risk for candidate molecules in a design process and for regulatory decision support.

Data transformation is an essential component in the QSAR modeling of carcinogenicity from rodent bioassay. This process converts tumor incidence findings into weighted numerical form with the highest score given to multi-site and trans-species tumor findings. This simulates aspects of the weight of evidence process used in regulatory risk analysis. Tumor site is not considered in this modeling process because there is poor tumor site concordance between rats and mice making it a poor factor for QSAR modeling [21].

Converting molecular structure into electrotopological state (E-State) descriptors and molecular connectivity chi indices also provides a means for modeling proprietary molecular structures that does not disclose the exact structure and identity of proprietary molecules. In this report proprietary compounds were included in the training data set and in the 50 test compounds. The name and structure of proprietary compounds was encoded and kept confidential by the FDA. The proprietary structure information descriptors contained sufficient information for successful modeling but are insufficient to unambiguously recreate a molecular structure, making this a valuable tool for data sharing while preserving confidentiality.

Computational toxicology combines databases and chemoinformatic data mining techniques with statistical methods to identify relationships between chemical structures and toxicological activities. Computational or predictive toxicology software programs are a means of evaluating knowledge accumulated from decades of toxicology studies to provide effective regulatory and product development decision support information. This approach is especially useful for prioritizing potential hazard and identifying data gaps in situations where toxicological data is limited, e.g., indirect food additives or contaminants and degredants/contaminants in the pharmaceutical manufacturing process. In drug development, the application of combinatorial chemistry and high throughput screening has resulted in an unprecedented increase in the number of compounds identified with potentially desirable pharmacological properties. The selection of lead compounds for development is currently hampered by limitations in the available toxicity screening methods. Making better use of accumulated scientific knowledge incorporated in predictive software is one way to minimize toxicity related drug failures and improve pharmaceutical risk management. Identifying serious potential toxicity early in the drug development process before significant investments in time and resources are expended is a major goal for FDA/CDER and the pharmaceutical industry. A current cause for concern is that too many drugs are failing late in the development process in phase III either for lack of efficacy or toxicity. It is estimated that 20% of total R&D costs per drug are spent on compounds that ultimately fail due to unfavorable ADME/Toxicity. The selection of drug candidates with better safety profiles will also reduce the regulatory review burden and speed the approval process by reducing the number of drugs submitted with serious safety issues that necessitate multiple review cycles or result in termination. Review resources expended for drugs that never make it to market are lost and could better be used for compounds that will succeed.

Computational or predictive toxicology has potential regulatory and drug development applications that can ultimately benefit the public health as well as refine and reduce the use of animals in the assessment of safety.

ACKNOWLEDGEMENT

We wish to express our appreciation to Vladimir Shwartz, University of St. Petersburg, St. Petersburg, Russia, for his assistance with the discriminant analysis and related statistical matters.

REFERENCES

[1] Willett P.: *Three-Dimensional Chemical Structure Handling*; John Wiley & Sons: New York, (1991).

[2] Lajiness M.S.: *Molecular Similarity-Based Methods for Selecting Compounds for Screening, In Computational Chemical Graph Theory,* Rouvray,

D.H. Ed.; Nova Science: New York, pp. 300-312, (1990).

[3] Johnson M., Maggiora G.M.: *Concepts and Applications of Molecular Similarity*: John Wiley & Sons: New York, (1990).

[4] Willett P.: *Similarity and Clustering in Chemical Information Systems*, John Wiley & Sons: New York, (1987).

[5] Warr W.: *Chemical Structures. The International Language of Chemistry*, Springer: Berlin, (1988).

[6] Hall L.H., Kier L.B.: *Electrotopological state indices for atom types: A Novel combination of electronic, topological and valence state information*, J. Chem. Inf. Comput. Sci. 35, 1039-1045, (1995).

[7] Kier L.B.; Hall L.H.: *Molecular Structure Description: The Electrotopological State:* Academic Press: San Diego, (1999).

[8] Hall L.H., Kier L.B.: *Molecular Connectivity Indices for Database Analysis and Structure-Property Modeling, in Topological Indices and Related Descriptors in QSAR and QSPR*, Devillers, J. and Balaban, A. T. Eds.; pp. 307-360, (1999).

[9] Hall L.H.; Kier L.B.: *Issues in the representation of molecular structure: The development of molecular connectivity*. J. Molecular. Model. Graphics 20, 4-18, (2001).

[10] Kier L.B., Hall L.H.: *Database organization and similarity searching with E-State indices*. SAR QSAR Environ. Res. 12, 55-74, (2001).

[11] Hall L.H.: *A structure-information approach to prediction of biological activities and properties*. Chem. Biodiversity 1, 183-201, (2004).

[12] Contrera J.F., Matthews E.J., Benz R.D.: *Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices*. Regul. Toxicol. Pharm. 38, 243-259, (2003).

[13] Gold L.S., Manley N., Slone T., Garfinkel G., Rohrback L., Ames B.N.: *The fifth plot of the carcinogenic potency database: Results of animal bioassays published in the general literature through 1988, by the National Toxicology Program through 1989*. Environrn. Health Perspect. 100, 65-135, (1993).

[14] Haseman J.K.: *A re-examination of false-positive rates for carcinogenicity studies.* Fundam. Appl. Toxicol. 3, 334-339, (1983).

[15] McConnell E.E., Solleveld H.A., Swenberg J.A., Boorman G. A.: *Guidelines for combining neoplasms for evaluation of rodent carcinogenesis studies.* JNCI, 76, 283-394, (1986).

[16] Matthews E.J., Contrera J.F.: *A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software*. Regul. Toxicol. Pharmacol. 28, 242-264, (1998).

[17] a) MDL Information Systems, 200 Wheeler Road, Burlington, MA, 01803.

b) Kellogg, G. E.; Hall, L. H.; Molconn-Z. See http://www.eslc.vabiotech.com/molconn/

c) For a list of publications illustrating applications of the descriptors in Molconn-Z, see http://www.eslc.vabiotech.com/molconn/mconpubs.html

d) See MDL®QSAR Users Guide for specific illustration of topological descriptors.

[18] Anderson T.W.: *An Introduction to Multivariate Statistical Analysis*, Second Edition, John Wiley & Sons: New York, (1984).

[19] Kendall M.G., Stuart A., Ord J.K.: *The Advanced Theory of Statistics*, Macmillan Publishing: New York, Vol. 3, Fourth Edition, (1983).

[20] Grabowski H., Vernon J., DiMasi J.: *Returns on research and development for 1990s new drug introductions.* Pharmacoeconomics 20, (Suppl. 3), 11-29, (2000).

[21] Contrera J.F., Jacobs A.C., DeGeorge J.J.: *Carcinogenicity testing and the evaluation of regulatory requirements for pharmaceuticals.* Regul. Toxicol. Pharmacol. 25, 130 -145, (1997).