ESP: A Method To Predict Toxicity and Pharmacological Properties of Chemicals Using Multiple MCASE Databases

Gilles Klopman,*,†,‡ Suman K. Chakravarti,† Hao Zhu,† Julian M. Ivanov,‡ and Roustem D. Saiakhov‡

Department of Chemistry, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, and MULTICASE Inc., 23811 Chagrin Boulevard, Suite 305, Beachwood, Ohio 44122

Received September 9, 2003

We describe here the development of a computer program which uses a new method called Expert System Prediction (ESP), to predict toxic end points and pharmacological properties of chemicals based on multiple modules created by the MCASE artificial intelligence system. The modules are generally based on different biological models measuring related end points. The purpose is to improve the decision making process regarding the overall activity or inactivity of the chemicals and also to enable rapid in silico screening. ESP evaluates the significance of the biophores from a different viewpoint and uses this information for predicting the activity of new chemicals. We have used a unique encoding system to represent relevant features of a chemical in the form of a pattern vector and a genetic artificial neural network (GA-ANN) to gain knowledge of the relationship between these patterns and the overall pharmacological property. The effectiveness of ESP is illustrated in the prediction of general carcinogenicity of a diverse set of chemicals using MCASE male/female rats and mice carcinogenicity modules.

INTRODUCTION

Combinatorial libraries are structurally diverse collections of a large number of chemicals. They have become the starting place for many drug discovery programs and are often used for lead identification. High throughput screening and ultrahigh throughput screening techniques are rapid methods to screen such libraries for lead molecules but still are relatively costly. In contrast, in silico approaches can potentially provide quick answers for a large number of compounds in an unbiased manner.[1] In silico screening of compound libraries for potential undesirable attributes such as toxicity before they become barriers in the development process is a key. Although many toxicity end points are of potential interest for screening, mutagenicity and carcinogenicity end points have been the focus of the most intensive efforts to develop predictive systems because these end points tend to be of prominent concern from a health and regulatory standpoint.

Both mutagenicity and carcinogenicity end points are required to be evaluated against multiple biological models according to regulatory guidelines. The most common mutagenicity test, the Ames test, measures the ability of a compound to cause reverse mutations in a series of different *Salmonella typhimurium* strains.[2] The assay is conducted in the absence and in the presence of liver microsomal preparations.[3] Similarly, carcinogenicity of a chemical is generally assessed against different species and different genders of rodents, for example, male and female mice and rats.[4] Screening against multiple SAR/QSAR models may increase the probability of correct detection of significantly active and inactive compounds because they cover a wider range of biological and mechanistic domain.[5,6]

The carcinogenic effect of a chemical is of major concern if it is observed in several organs and/or animal species. If the effect of the chemical is specific to a single organ of a single species, then it is likely that it is irrelevant to humans. In contrast, if a chemical is truly genotoxic, its effect will be seen in different cells and the specific structural alert identified by the program for this chemical should be the same (or similar) for the majority of cells where a toxic effect is observed. Therefore, relevant toxicity prediction models would benefit from including as many submodels as possible covering a wide chemical/mechanistic space and different animal species and gender.[7] However, submodels representing different species/gender cannot generally be combined to make a global single model.[8] Indeed, the relationship between different models is often not straightforward and can blur the causality of the composite model, complicating the process of assessment of the overall activity of the chemicals. Methods which can identify the relationship between different SAR/QSAR models and combine the results of screening against different models may therefore yield the best predictions. Such a method should provide consistent and good results, should withstand weaknesses present in the individual models, and help identify the true structural features responsible for activity.

The existing computer based toxicity prediction programs are generally rule based approaches, although models also exist based on simple quantitative structure−activity relationships (QSAR). The latter are most of the time developed for small sets of congeneric molecules, but sometimes correlations are also sought between the activity of large data sets and miscellaneous preselected parameters (TOPKAT).[9]

* Corresponding author phone: (216)368-3618; fax: (216)368-5921; e-mail: gxk6@po.cwru.edu.
† Case Western Reserve University.
‡ MULTICASE Inc.

10.1021/ci030298n CCC: $27.50
Published on Web 01/03/2004

Rule-based approaches differ in the way they derive the rules. Some programs develop rules from prior human expert's knowledge of the relationship between the chemical structure and the likely mechanism of toxicity and then use these rules to predict the toxicity of the test compounds (DEREK[10] and ONCOLOGIC[11]). CASE[12] and MCASE,[13] on the other hand, are hybrids. Using automated algorithms, they develop their rules automatically from a correlative study of a knowledge database consisting of activity and structure of molecules for which experimental activity is known. CASE and MCASE identify the most relevant descriptors called "biophores" or "structural alerts" and automatically create expert systems capable of recognizing the existence of these structural alerts in new chemicals. MCASE further creates QSAR type correlations for each group of molecules containing a specific structural alert. This learned information is then used in the analysis of test compounds.

The simple rule based approaches are somewhat biased as they are based on prior conceptions and experience, whereas the correlative approach MCASE strives to avoid such bias in the interest of discovering new SAR/QSAR associations potentially useful for prediction. In the absence of prior knowledge of the relation between structure and a specific activity, pure rule based approaches cannot be created. Their development in such cases will depend on new knowledge often discovered independently by one of the correlative approaches. On the other hand, CASE and MCASE mine existing stores of mutagenicity and carcinogenicity data in an effort to recognize and codify new structure−activity relationships beyond the limits of human perception.

The CASE and MCASE programs have been applied to the study of noncongeneric databases associated with a range of toxicity end points and used in a variety of creative ways as a research tool, e.g., to explore properties of databases, and the dependence of statistical prediction ability on database size and diversity.[14−16] In such a study, Matthews et al.[17] made an effort to combine MCASE single database results to predict rodent carcinogenicity, and this resulted in a significant improvement over the single database predictions. In this method the authors made some changes in the conventional interpretation of MCASE results by redefining the biological significance of the MCASE biophores based on multidatabase principles. A major drawback of this approach was its specificity for rodent carcinogenicity. The approach was not automatic, needed significant human expert intervention to arrive at the final decision, and did not provide any statistical confidence measure of the activity prediction.

Encouraged by the work of Matthews et al. we have developed a computer program which uses a new method called Expert System Prediction or ESP, to predict overall biological property of a chemical using several MCASE modules. The objective of the present paper is to discuss this method and analyze its performance. This method offers significant developmental advantages over the method of Matthews et al. It redefines the biological significance of the biophores (fragments which are relevant to the activity) after performing a substructural search through all the training databases being used. It then uses a genetic artificial neural network (GA-ANN) to learn the relationship between MCASE prediction patterns and the overall biological property of the chemicals. This acquired knowledge can then be used to predict activity of new chemicals. The training step needs to be performed only once, and thereafter any number of chemicals can be tested. ESP also provides a quantitative confidence score of the prediction results. Thus, this method is very suitable for in silico screening of a large number of chemicals. The method also offers the opportunity to study the relationship between the biophores present across different models. Moreover, it is fully automated and supported by a friendly graphical user interface in Windows environment. To analyze the performance of ESP, a diverse set of totally new chemicals with known rodent carcinogenicity was tested. The results are presented and discussed in conjunction with similar results obtained without the use of ESP.

Table 1. Features of the Rodent Carcinogenicity Databases Used To Build MC4PC Master Modules

module name	rodent species/ gender	total number of chemicals	I/M/A[a]	percent of active chemicals
AF1	male rat	1002	458/65/479	47.8
AF2	female rat	991	502/52/437	44.1
AF3	male mouse	889	451/51/387	43.5
AF4	female mouse	902	451/44/407	45.1

[a] Numbers of inactive/marginally active/active chemicals in database.

DATA SETS

In the present study, several rodent carcinogenicity[2,3] databases were used. Primarily three different data sets were employed: 1. a set of four data sets of male/female rats and mice carcinogenicity was used to build MCASE master modules; 2. a small database to train ESP; and 3. a set of new chemicals to test the ability of ESP to predict overall carcinogenicity of a chemical.

The four rodent carcinogenicity databases (AF1−4) were assembled by FDA's Center for Drug Evaluation and Research's (CDER) Office of Testing and Research (OTR) under a cooperative research and development agreement with MULTICASE Inc. The four database modules were constructed to predict carcinogenicity in male rats (AF1), female rats (AF2), male mice (AF3), and female mice (AF4). Specific features of these databases are summarized in Table 1.

The databases which have been used to train and test ESP were obtained from the Handbook of Carcinogenic Potency and Genotoxicity Databases, but the biological activity of the chemicals contained therein was binary and only denote whether the chemical is overall carcinogenic or noncarcinogenic.[18] Proper care was taken to ensure that any chemical present in the test set is not present in the master modules and in the training sets.

Biological Activity. The representation of carcinogenic potency of the chemicals in the four main rodent carcinogenicity modules (AF1−4) is described in detail by Matthews et al.[17] In short, the majority of the studies used an oral route of exposure (feed, gavage, or drinking water) although some compounds were tested by nonoral routes of administration. The duration of acceptable carcinogenicity studies was limited to >= 18 months for inactive compounds. All studies with positive findings were acceptable regardless of duration of treatment with the exception that positive nonoral studies

were included if tumors were induced at other than the site of application. The carcinogenic potential of a compound is expressed on a scale of CASE units ranging from 10 to 79 depending on the tumor response of the chemical at single tissue/organ site or multiple tissue/organ sites. The corresponding CASE units and qualitative tumor responses include the following: inactive (10−19), marginal (20−29), weakly active (30−34), active (35−39), very active (40−49), and extremely active (50−79).

For rest of the databases which have been used to train and test ESP, the biological activity was represented on a binary scale, i.e., 0 for noncarcinogenic and 1 for carcinogenic. For the purpose of this study, a carcinogen must induce single tissue/organ site tumors in two or more modules (male/female − rat/mouse modules) or have a multiple tissue/organ site tumors in at least a single module.

METHODS

MCASE Methodology. The MCASE methodology[12,13] and its applications have been extensively published in the scientific literature.[19−22] In short, the MCASE program is based on a hierarchical statistical analysis of a database (called the training set) composed of a number of chemicals with their biological activity data. The program aims to discover substructures that appear mostly in active molecules and may therefore be responsible for the observed activity. It starts by identifying the statistically most significant substructure existing within the learning set. This fragment, labeled the top biophore, is seen as responsible for the activity of the largest possible number of active molecules. The active molecules containing this biophore are then removed from the database, and the remaining ones are submitted to a new analysis leading to the identification of the next biophore. This procedure is repeated until either the activity of all the molecules in the learning set have been accounted for or no additional statistically significant substructure can be found. For each set of molecules containing a specific biophore, MCASE identifies additional parameters, deemed modulators, which can be used in the construction of a quantitative structure−activity relationship within this reduced set of congeneric molecules. Modulators consist of the presence of certain substructures or the value of calculated parameters such as the highest occupied and lowest unoccupied orbital energies, octanol−water partition coefficient, and so on. The process is automated and proceeds with minimal human intervention and bias. The knowledge that the program gains during the training process can then be used to predict the biological activity of new chemicals that were not included in the training set. For the present work, a new generation Windows based PC version of MCASE, called MC4PC, is used, and all the calculations presented herein were performed using an Intel Pentium III 450 MHz processor based personal workstation.

ESP Methodology. The ESP methodology was implemented in a computer program called WinESP which was written in C++ in our laboratory. Currently, it has a Windows based interface and can be run on personal computers. The methodology consists of the following steps: 1. building the MC4PC master modules, 2. performing a MC4PC analysis of WinESP training set and test data sets, 3. building a biophore database, 4. training of WinESP, and 5. predicting the overall activity of the test set chemicals using the trained WinESP. The flow of ESP methodology is shown in Figure 1.

1. Building the MC4PC Master Modules. Four MC4PC master modules were prepared by submitting the SMILES code of the chemicals with their experimental carcinogenicity in male/female mice/rats in the form of four separate databases. For each module, MC4PC generates a file, called the S-File, containing a list of biophores with their relevant details. The details include distribution of chemicals in active, inactive, and marginally active categories containing the biophore as well as other statistical results.

2. MC4PC Analysis of ESP Training Set and Test Data Sets. The data set for WinESP training and the data set which contains the test set chemicals were analyzed using the MC4PC master modules, and the results of the analysis were sent to another file called the R-File. This file contains records of all the details of the analysis for the chemicals of the data set. The R-files are used by the ESP program during the training and prediction phases. The textual representation of a record in an R-file is presented in Figure 2 showing the MC4PC analysis of 2-acetamidofluorene against the AF1 module.

3. Building a Biophore Database. The data of all the S-files are combined to create a biophore database. During this step the significance of the biophores is defined by assigning a score to each of them. The score is calculated by multiplying the average CASE unit activity of the biophore by the number of chemicals containing the biophore (total CASE activity) and then adding the calculated value from all the MC4PC modules containing the biophore. [Personal communication from Dr. Edwin J. Matthews of the U.S. Food and Drug Administration.] Each biophore is then classified in one of four possible categories: significant, nonsignificant, possibly significant, and unreliable biophore. A "significant" biophore has at least a score of 200, "possibly significant" biophores have scores ranging between 150 and 199, and "nonsignificant" biophores have scores less than 150. If a biophore exists in less than 67% of active compounds or MC4PC finds a substructure to exist predominantly in inactive molecules (biophobe) in another model, it is classified as "unreliable". Thus, rather than relying on the biophores identified in individual MC4PC modules, WinESP requires confirmatory experimental findings across correlated biological end points, a high degree of representation of the biophores in the training data set, and a high biological potency for activity. Table 2 shows some of the most significant biophores compiled from the four master modules.

4. Training of WinESP. Since WinESP uses a GA-ANN for the prediction of new chemicals, it has to be trained for this purpose. Basically, the results of an MC4PC analysis (R-files) of a set of chemicals with known biological activity are submitted to WinESP. The program then learns the relationship between the patterns generated by MC4PC analysis and the activity of the chemical using the pattern recognition capability of the GA-ANN. This knowledge is later used for activity prediction of new molecules. In general, the training process consists of the following steps:

(a) Generation of Pattern Vectors from MC4PC Analysis. A pattern vector is essentially a 13 member real number array generated for each chemical of the training

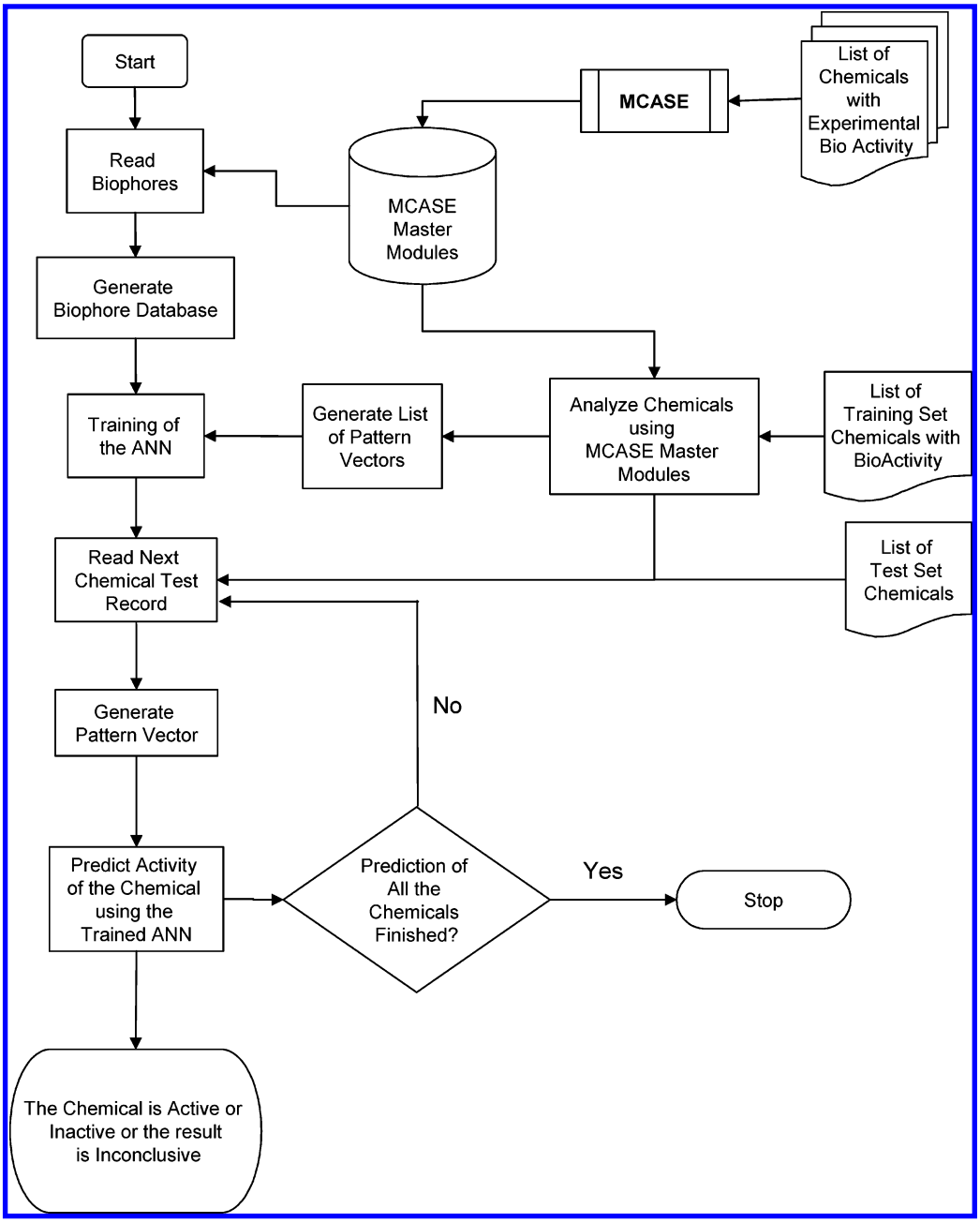

Figure 1. Flowchart of the ESP methodology.

set based on its analysis against selected MC4PC master modules, and it is a digitized summary of the chemical's test against the MC4PC master modules. These pattern vectors are used as inputs to the ANN. The process of the pattern vector generation for a chemical is shown in Figure 3. Each component of a pattern vector is a real number ranging from 0.1 to 0.9 and represents a feature of the chemical obtained from its MC4PC analysis. A list of the thirteen features is given in Table 3. We have selected these features so as to provide the best possible representation of a chemical's multiple MC4PC analysis in a compact form.

(b) Training of the GA-ANN. ANNs are computer-based simulations which contain some elements that appear to exist in living nervous systems. There have recently been numerous applications of ANNs to chemical problems.[23−25] For details of the ANN implementation, readers are referred to relevant references or standard texts.[26,27] In this work, a standard feed forward fully connected ANN was implemented utilizing a genetic algorithm for training (Figure 4). The network uses a sigmoidal transfer function. The transfer function has minimum and maximum output values of 0 and 1, respectively. The network accepts the pattern vectors of the chemicals in the training set as inputs. The ANN has 13 neurons in the input layer, each accepting a component of a pattern vector, 4 neurons in the hidden layer and the output layer has one neuron. Depending on the value at the output neuron, the chemical presented to the network is either predicted as active or inactive. The hidden and the output layers also have one bias neuron each.

A genetic algorithm was used to adjust the weights between the neurons in the ANN and train it to make the best predictions. Genetic algorithms are stochastic optimization methods that have been inspired by evolutionary principles.[28] The distinctive aspect of a genetic algorithm is that it investigates many possible solutions simultaneously, each of which explores different regions in the parameter

```
 ------------------------------------------------------------------------------
 AF1- CARCINO      - MALE RAT-REDUCED                    - XXXX    # 983 1.54
 ------------------------------------------------------------------------------
Processing... 2-Acetamidofluorene

MULTICASE-3 Prediction
 The molecule contains the Biophore   (nr.occ.= 1):
           CH2-c. =cH -c <=

***  8 out of the known   9 molecules ( 89%) containing such Biophore
         are  CARCINOGENIC  with an average activity of  63. (conf.level= 98%)
*** QSAR Contribution :                                 Constant is    72.22
                                .                                      ------
      ** Total projected QSAR activity x, (x  = response           )    72.22

  CONCLUSIONS:
  ------------
 ** The projected  CARCINOGENIC   activity is   72.0     CASE units **
 ** The compound is predicted to be EXTREMELY active **
*** The probability that this molecule is  CARCINOGENIC  is    90%  **-
```

Figure 2. Textual representation of one record in an Rxx_Sxy.ESP file.

Table 2. List of Some Important Biophores for Rodent Carcinogenicity Found in the four MC4PC Master Modules

biophore[a]	database(s) containing this biophores and their total CASE activity in parentheses
NO−N−	AF1(3500), AF2(3479), AF3(825)
NO−N−CH3	AF4(244)
NH2−NH−	AF3(902), AF4(1000)
NH2−NH2	AF1(79), AF2(79), AF3(79), AF4(79)
O^−CH2−	AF1(416), AF2(1848), AF3(294), AF4(408)
Cl−CH2−CH2−N−	AF3(1098), AF4(1140)
Cl−CH−Cl	AF3(539), AF4(495)
NH2−c=c−cH=cH−c=cH−	AF3(630)
NH2−c=cH−c.=	AF3(264)
Cl−CH=	AF1(336), AF2(318), AF4(343)
COH−N−	AF1(70), AF2(70), AF3(305)
c.=c.−cH=cH−c=	AF2(300)

[a] ^Stands for an atom of a three member ring; small letters denote aromatic atoms; a period followed by an atom denotes that the atom is part of another ring.

space.[29] Therefore, there is an increased possibility to achieve global optimization. Use of genetic algorithms in training neural networks is an attractive choice for complex problems with an error surface with many local minimums, and there is a significant risk that the traditional back-propagation method would converge to one of the local minimums.[30] These considerations and the complex nature of the present problem lead us to use such a genetic algorithm in training the ANN.

The basic design of the genetic algorithm employed in the present work is shown in Figure 5. After setting up the initial parameter list, the first step in the algorithm is to create a population of N chromosomes/individuals. Each chromosome is a vector of real numbers having a length equal to the total number of weights in the ANN. Thus, for the ANN of 13−4−1 topology, the length of each chromosome will be $(13 \times 4)+(4 \times 1)+4+1 = 61$. Each component of the chromosome can be mapped to a particular weight in the ANN. Initially, the chromosomes are assigned random numbers ranging from −10 to 10.

In the next step, the fitness of each chromosome is determined using a fitness function. For this, the input pattern

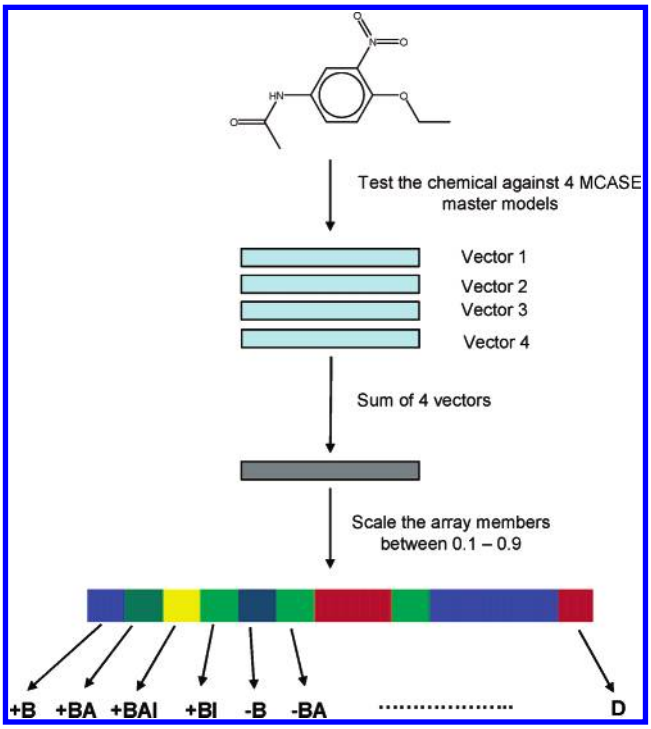

Figure 3. Generation of a pattern vector for a chemical.

vectors of all the training set chemicals are presented to the input layer of the ANN, the signal is propagated through the ANN, and the output at the output layer is compared to the desired output. The fitness of the chromosome is calculated by the RMS error (eq 1)

$$\sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{K}(Obs\_out_{i,j} - Calc\_out_{i,j})^2}{M}} \qquad (1)$$

where M is the number of training set chemicals, K is the number of output neurons (which is always 1), *Obs_out* is the experimental output (1 for active and 0 for inactive), and *Calc_out* is the output calculated by the ANN.

Next, a sexual reproduction takes place so that the new offspring contains characteristics from both of the parents.

Table 3. List of the Components of a Pattern Vector

no.	symbol	vector component
1	+B	number of significant biophores
2	+BA	number of significant biophores with activating modulators
3	+BAI	number of significant biophores with activating and inactivating modulators
4	+BI	number of significant biophores with inactivating modulators
5	−B	number of nonsignificant biophores
6	−BA	number of nonsignificant biophores with activating modulators
7	−BAI	number of nonsignificant biophores with activating and inactivating modulators
8	−BI	number of nonsignificant biophores with inactivating modulators
9	(+)B	number of possibly significant biophores
10	(+)BA	number of possibly significant biophores with activating modulators
11	(+)BAI	number of possibly significant biophores with activating and inactivating modulators
12	(+)BI	number of possibly significant biophores with inactivating modulators
13	D	number of deactivating fragments

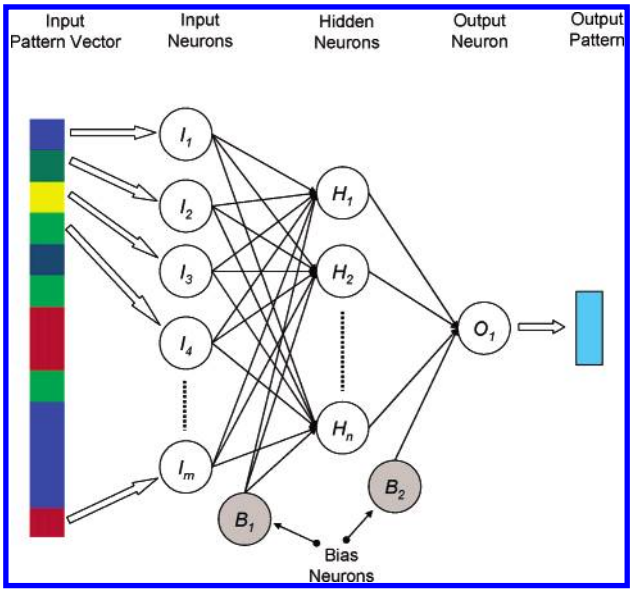

Figure 4. The artificial neural network used in WinESP.

Two individuals are selected probabilistically on the basis of their fitness scores and serve as parents. Next, in a two-point crossover operation each parent contributes a randomly selected part of its genetic material, and two children are constructed by switching these two parts of "genetic code". Finally, this child is subjected to a random mutation in one of its genes; i.e., one chromosome component is replaced by a random number. This selection-crossover-mutation process was repeated until all of the N parents in the gene pool were replaced by their children. The fitness score of each member of this new generation is again evaluated, and the reproductive cycle was continued until a desired number of generations or target fitness score was reached. The cycle can optionally be operated in the "elitism" mode, in which the chromosome with the best fitness score is always passed on to the next generation. After the reproductive cycle is complete, the final population contains a number of good solutions, and some of the top scoring chromosomes can be used to assign weights to the ANN and to predict the activity of new chemicals.

5. Prediction of Overall Activity of the Test Set Chemicals Using the Trained WinESP. The last step in any WinESP run is to predict the biological or toxicity activity of untested chemicals using the structure−activity relationship knowledge gained by the ANN. To accomplish this, the R-files of the test chemicals (containing the results of the MC4PC tests against different master databases) are submitted to WinESP, which in turn generates a pattern vector for each tested chemical and send this vector as an input to the ANN to obtain a decision regarding the activity or inactivity of the test chemical. One can also calculate a quantitative confidence score for the prediction results using eq 2.

$$\text{Abs(value at output neuron} - 0.5) \times 2 \times 100 \quad (2)$$

Thus, if the value at the output neuron is 0.996, then the chemical will be predicted as active with a confidence score of 99.2%. If this confidence score is very low, the prediction result can be considered as not sufficient to indicate that the chemical is active and the chemical should be subjected to further evaluation.

RESULTS AND DISCUSSION

The Biophores. As pointed earlier, ESP builds a biophore database and in the present study, a database containing a total of 821 biophores was generated from the four rodent carcinogenicity master modules. ESP assigned a score to each biophore and thus was able to identify several biophores from the database which are significant according to the new WinESP scoring system. Five hundred thirty-five biophores were found to be nonsignificant. Some of the important biophores with high score are shown in Table 2. It can be seen that several biophores are present in more than one module which results in a high score assigned to the biophore and during the prediction phase if any chemical contains one of these biophores, it stands a good chance to be predicted as active. In contrast, the biophores which were found in a single module are most likely to be classified as not significant unless they have a very high total CASE Activity Index in a single MC4PC module.

Biophores identified in the present WinESP analysis represent functionalities of some common carcinogenic chemicals, e.g. *N*-nitroso compounds, hydrazine derivatives, alkylating agents, polynuclear aromatic compounds, aromatic amines, organic halogenated compounds, epoxy compounds, and others. Detailed discussion of each type of biophore is beyond the scope of this paper, and we are limiting the discussion to some major ones.

N-Nitroso compounds namely nitrosamines and nitrosamides are well-known potent carcinogens. Nitrosamines are chemically stable compounds requiring metabolic activation for their carcinogenic effect. *N*-Nitrosamides are chemically reactive compounds and in most cases do not require metabolic activation.[31] In Table 2 the biophores NO−N−

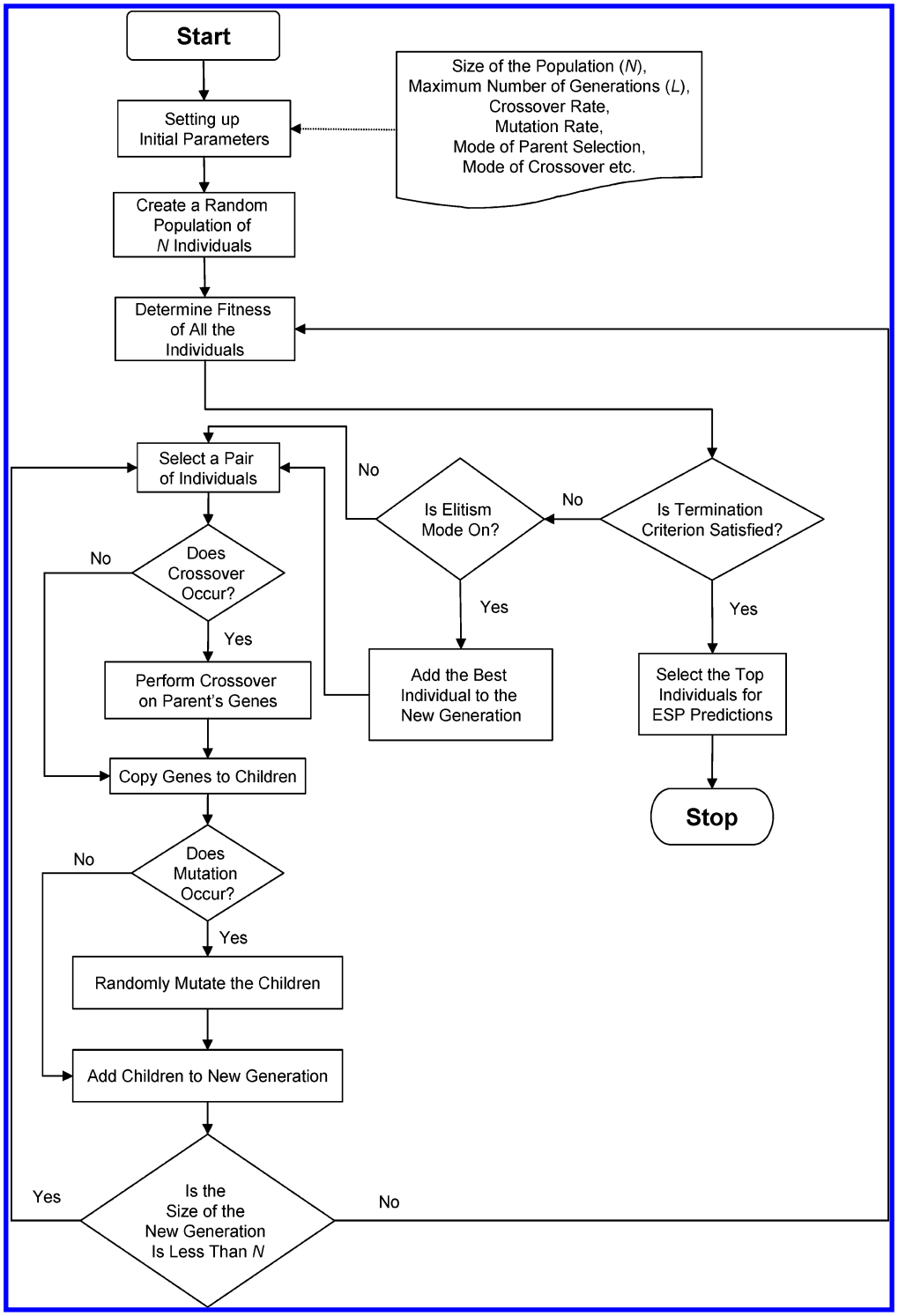

Figure 5. Design of the genetic algorithm used to train the ANN.

and NO−N−CH3 are representative of *N*-nitroso compounds, and it is evident that NO−N− has a very high total CASE Activity Index and is present in three modules.

Hydrazine and its derivatives are potent rodent carcinogens. Hydrazine itself can produce adenocarcinomas and hepatocarcinomas if given orally to rats and mice. *N*-Methyl-*N*-formylhydrazine induces tumors of the lungs, liver, gall bladder, bile ducts, and blood vessels in mice and hamsters.[32] The antituberculor drug isonicotinic acid hydrazide (INH) can induce lung tumors in mice.[32] In Table 2 the biophore NH2−NH2 represents hydrazine and NH2−NH− represents hydrazine derivatives. Significance and potency of the biophores can be assessed by their high total CASE activity and multimodule presence.

Alkylating agents are considered to be archetypical carcinogens. In fact the majority of other chemical carcinogens are only active after being metabolized to alkylating or arylating agents; prominent examples are aflatoxins, dialkyl-nitrosamines, polycyclic aromatic hydrocarbons, and vinyl chloride. In Table 2 Cl−CH2−CH2−N− represents nitrogen mustard type alkylating agents.

Genetic ANN Training. Several parameters in the genetic ANN can be adjusted to achieve a satisfactory learning of the association between input pattern vectors and the

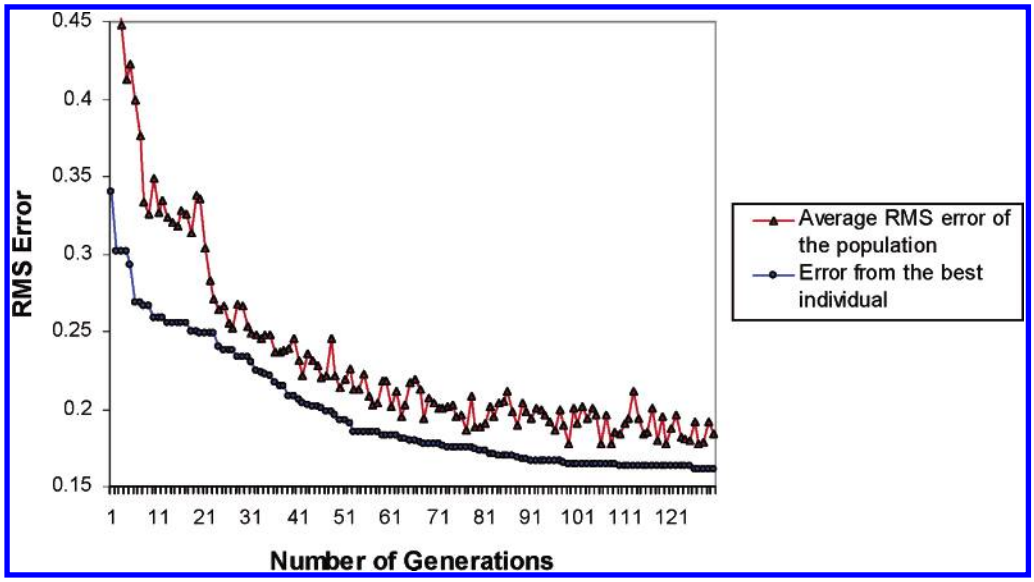

Figure 6. Plot of RMS error of the ANN during its training using genetic algorithm.

experimental overall carcinogenicity of the training set chemicals. The only adjustable parameter in the topology of the ANN is the number of hidden layer neurons. After several trials, we have found that four hidden neurons give satisfactory learning. Too many hidden neurons results into overtraining and too few of them are slow or unable to learn.

For the genetic algorithm, the adjustable parameters are the size of the population, the number of evolutions to be performed, the crossover probability, the mutation probability, the mode of parent selection, and the mode of cross over. Again, after several trials we experienced that a population size of 100, 130 evolutions or epochs, cross over probability and mutation probability of 1.0 and 0.6 respectively, a tournament mode (available options are roulette wheel, tournament and rank selection modes) for parent selection, and two point cross over mode is satisfactory in the present case of rodent carcinogenicity prediction. Elitism mode was kept "on" during the learning phase. It should be noted that these settings may not be satisfactory with other pharmacological/toxic end points and may need readjustment. In Figure 6, the average RMS error of the whole population and of the best individual through 130 generations of genetic evolution is shown. In this particular case, the above-mentioned parameter values were used, and 150 chemicals were present in the training set. It can be seen that the RMS error becomes stable approximately after 90 generations. According to our experience as well as others, there is always a chance of overtraining with long training sessions, and this should be avoided.

We have also tried training sets of different size. It was found that a training set of 130−150 chemicals is optimum for the present purpose, and increasing the size of the training set does not help in improving the prediction of unknown chemicals. Rather, the ratio of active and inactive chemicals in the training set is important and an unusually high proportion of either active or inactive chemicals should be avoided. For the present purpose we have used a training set of 150 chemicals in which the number of inactive chemicals is roughly twice that of actives.

A satisfactorily trained ANN model can be saved and can be used repeatedly for the activity prediction of different test sets.

Carcinogenicity Prediction of New Chemicals. An actual test of the capability of the ESP method is to predict carcinogenicity of entirely new chemicals which have neither been present in the MC4PC master modules nor in the WinESP training sets. It is also very important to observe the advantage it offers over the predictions made by the MC4PC program alone. As stated earlier, we have compiled a test set of 113 diverse chemicals for this purpose. This set contains 80 active and 33 inactive chemicals. The result of the overall carcinogenicity prediction of these 113 chemicals using WinESP and without WinESP (MC4PC alone) is shown in Table 4. The statistics for the predictions from both methods are presented in Table 5. We have used the GA solution with the highest fitness score for the prediction results in the above tables. Since WinESP has an option to let the user utilize other GA solutions also, we have tried some of them to see if their prediction results differ substantially from the top one. We found that although the fitness of the top GA solutions differ slightly, they produce almost the same result in predicting the activity and inactivity of the chemicals. GA solutions with poor fitness should not be used because they do not produce optimum results.

The first advantage offered by WinESP is prediction in the form of a single value which is either overall activity or inactivity of the test chemical based on the quality of biophores present in the different modules for the test chemical. This prediction is based on the knowledge gained during the GA-ANN training, whereas MC4PC gives 4 different predictions for test against each individual master module. Since MC4PC predictions do not give an indication about the single/multiple tissue/organ specificity of tumor response, therefore, it is very difficult to obtain an overall carcinogenicity prediction from these 4 different prediction results. For comparison purposes we made an attempt to determine an overall prediction from the 4 different predictions of MC4PC, and they are shown in Table 4. A chemical is declared active if it is predicted active in two or more modules and inactive if it is predicted to be inactive in all four modules or active in not more than one module. The result is declared inconclusive if the chemical is predicted to be possibly active, marginal, or inconclusive in only one module. In case of WinESP overall predictions, the result is

Table 4. WinESP Prediction Results of Chemicals of the Test Set

no.	CAS RN	name	ESP pred.[a]	value at the output neuron of the ANN: output	value at the output neuron of the ANN: C.S.[b]	exp.[c]	MC4PC predictions[d]: AF1	AF2	AF3	AF4	ovp.[e]
1	94360	benzoyl peroxide	-	0.087907	82.42	-	-	-	-	-	-
2	578767	7-methylguanine	-	0.087907	82.42	-	-	-	-	-	-
3	517282	hematoxylin	+	0.990099	98.02	+	+	+	-	-	+
4	106923	allyl glycidyl ether	+	0.999817	99.96	-	+	+	+	+	+
5	14026030	2-pipecoline, 1-nitroso-, (R)-	+	0.813184	62.64	+	+	+	(−)	-	+
6	115026	azaserine	+	0.998668	99.73	+	+	+	m	(−)	+
7	6452739	oxprenolol hydrochloride	-	0.180560	63.89	-	-	-	-	m	?
8	64755	tetracycline hydrochloride	-	0.179094	64.18	-	(+)	m	-	-	?
9	2353459	fast green FCF	-	0.000524	99.90	-	-	-	-	-	-
10	16423680	erythrosine B	-	0.180560	63.89	-	-	-	m	-	?
11	65765073	SaH 47−480	-	0.087907	82.42	-	-	-	-	-	-
12	71556	1,1,1-trichloroethane	-	0.159271	68.15	-	-	-	+	-	-
13	74317	1,4-benzenediamine, *N*,*N*′-diphenyl	-	0.139163	72.17	-	+	-	-	-	-
14	592314	*n*-butylurea	-	0.012750	97.45	-	-	+	+	+	+
15	59865133	cyclosporin A	?	0.359988	28.00	-	+	-	-	-	-
16	628944	adipamide	-	0.196710	60.66	-	+	m	-	-	+
17	1330785	tricresyl phosphate	-	0.204620	59.08	-	-	-	-	+	-
18	100643967	indolidan	-	0.315124	36.98	+	+	-	+	-	+
19	86315528	isomazole	-	0.088381	82.32	+	m	-	-	-	?
20	55984515	*N*-(nitrosomethyl)−2-oxopropylamine	+	0.823979	64.80	+	+	+	+	+	+
21	16301261	azoxyethane	-	0.000375	99.92	+	+	-	-	-	-
22	75411835	2-hydroxypropylmethylnitro-samine	+	0.813184	62.64	+	(+)	(+)	+	-	+
23	63642171	L-ornithine, N5-[(methylnitrosoamino) carbonyl]-	+	0.742084	48.42	+	(−)	(−)	+	(+)	+
24	816579	*N*-propyl-*N*-nitrosourea	+	0.957248	91.45	+	+	+	+	-	+
25	91930	3,3′-dimethoxy-4,4′-biphenylylene isocyanate	-	0.087907	82.42	+	-	-	-	-	-
26	57497344	Z-methyl-*O*,*N*,*N*-azoxyethane	-	0.000375	99.92	+	+	-	-	-	-
27	57497297	Z-ethyl-*O*,*N*,*N*-azoxymethane	-	0.087907	82.42	+	-	-	-	-	-
28	129431	1-hydroxyanthraquinone	?	0.579149	15.83	+	(+)	+	-	-	+
29	18774851	*N*-hexylnitrosourea	+	0.957248	91.45	+	+	+	+	-	+
30	760565	1-allyl-1-nitrosourea	+	0.957248	91.45	+	+	+	+	-	+
31	57307	phenobarbital sodium	-	0.188731	62.25	+	-	-	+	(+)	+
32	135206	cupferron	+	0.813184	62.64	+	+	+	+	-	+
33	88133113	bemitradine	-	0.087907	82.42	+	-	-	-	-	-
34	22571955	symphytine	+	0.998410	99.68	+	+	+	+	+	+
35	684935	*N*-nitroso-*N*-methylurea	+	0.936479	87.30	+	+	+	+	+	+
36	53609646	2-propanol, 1,1′-(nitrosoimino)bis-	+	0.813184	62.64	+	+	+	+	-	+
37	1912249	atrazine	-	0.088381	82.32	+	-	-	(+)	-	?
38	33372393	4-bis(2-hydroxyethyl)amino-2-(5-nitro-2-thienyl)quinazoline	-	0.087907	82.42	+	-	-	-	-	-
39	598572	methylnitramine	-	0.087907	82.42	+	-	-	-	-	-
40	120627	piperonyl sulfoxide	-	0.000192	99.96	+	m	m	+	-	+
41	8001352	toxaphene	-	0.087907	82.42	+	-	-	-	-	-
42	108601	bis-chloroisopropyl ether	+	0.882961	76.59	+	-	+	+	+	+
43	1582098	trifluralin	?	0.366396	26.72	+	-	-	-	(+)	?
44	759739	*N*-nitroso-*N*-ethylurea	+	0.957248	91.45	+	+	+	+	-	+
45	54749905	chlorozotocin	+	0.970191	94.04	+	+	+	+	m	+
46	443481	flagyl	-	0.087907	82.42	+	-	-	-	-	-
47	120809	catechol	-	0.088381	82.32	+	+	-	-	-	-
48	533313	sesamol	+	0.872160	74.43	+	+	-	+	-	+
49	71752700	*N*-(3-hydroxypropyl)-*N*-nitrosourea	+	0.957248	91.45	+	+	+	+	-	+
50	61825	1H-1,2,4-triazol-3-amine	-	0.307846	38.43	+	-	-	-	+	-
51	60560	methimazole	-	0.087907	82.42	+	-	-	-	-	-
52	91308713	2-propanone, 1-(nitroso-2-propenylamino)-	+	0.813184	62.64	+	+	+	+	-	+
53	86451378	1,2-propanediol, 3-(methylnitrosoamino)-	+	0.934161	86.83	+	+	+	+	+	+
54	91308702	2-propanol, 1-(nitroso-2-propenylamino)-	+	0.813184	62.64	+	+	+	+	-	+
55	621647	*N*-nitrosodipropylamine	+	0.813184	62.64	+	+	+	+	-	+
56	55557001	dinitrosohomopiperazine	+	0.905423	81.08	+	+	+	+	+	+
57	81795075	4H-1,3,5-dithiazine, dihydro-2,4,6-trimethyl-5-nitroso-	+	0.813184	62.64	+	+	+	+	-	+
58	55556928	1-nitroso-1,2,3,6-tetrahydropyridine	+	0.905423	81.08	+	+	+	+	+	+
59	75896332	2-propanol, 1-[(2-hydroxyethyl)nitrosoamino]-	+	0.813184	62.64	+	+	+	+	-	+
60	88208166	*N*-nitroso-2,3-dihydroxypropylallylamine	+	0.946365	89.27	+	+	+	+	+	+
61	64005625	nitrosoamylurethane	+	0.905423	81.08	+	+	+	+	-	+
62	614959	nitrosoethylurethane	+	0.905423	81.08	+	+	+	+	-	+
63	92177509	2-propanone, 1-[(2,3-dihydroxypropyl)nitrosoamino]-	+	0.843403	68.68	+	+	+	+	+	+
64	89911795	1,2-propanediol, 3-[(2-hydroxypropyl)nitrosoamino]-	+	0.946365	89.27	+	+	+	+	+	+
65	25843452	azoxymethane	-	0.087907	82.42	+	-	-	-	-	-
66	363177	acetamide, *N*-fluoren-2-yl-2,2,2-trifluoro-	+	0.997345	99.47	+	+	+	+	+	+
67	3096502	acetamide, *N*-(9-oxo-9H-fluoren-2-yl)-	+	0.986888	97.38	+	-	+	+	+	+
68	63019658	*N*-1-diacetamidofluorene	+	0.850647	70.13	+	-	+	+	-	+

Table 4 (Continued)

no.	CAS RN	name	ESP pred.[a]	value at the output neuron of the ANN: output	value at the output neuron of the ANN: C.S.[b]	exp.[c]	MC4PC predictions[d]: AF1	AF2	AF3	AF4	ovp.[e]
69	3031514	l-5-morpholinomethyl-3-[(5-nitrofurfurylidene)amino]-2-oxazolidinone hydrochloride	?	0.485542	2.89	+	(+)	+	-	+	+
70	26049694	2-(2,2-dimethylhydrazino)-4-(5-nitro-2-furyl)thiazole	+	0.999879	99.98	+	+	+	+	+	+
71	56654525	1,3-dibutyl-1-nitrosourea	+	0.946365	89.27	+	+	+	+	-	+
72	4342034	dacarbazine	-	0.023453	95.31	+	-	+	+	+	+
73	36133887	*N*-[[3-(5-nitro-2-furyl)-1,2,4-oxadiazol-5-yl]methyl]acetamide	+	0.985323	97.06	+	+	+	+	+	+
74	3775551	1,3,4-oxadiazole, 2-amino-5-(5-nitro-2-furyl)-	+	0.962568	92.51	+	+	+	-	+	+
75	26049683	2-hydrazino-4-(5-nitro-2-furyl)thiazole	?	0.468374	6.33	+	+	+	-	+	+
76	531180	hexa (hydroxymethyl)melamine	-	0.180560	63.89	+	-	+	-	-	-
77	59358	4,6-dimethyl-2-(5-nitro-2-furyl)pyrimidine	+	0.977042	95.41	+	+	+	-	+	+
78	91308699	*N*-nitrosoallylethanolamine	+	0.813184	62.64	+	+	+	+	-	+
79	59892	*N*-nitrosomorpholine	+	0.905423	81.08	+	+	+	+	+	+
80	55738540	*trans*-2-[(dimethylamino)methylimino]-5-[2-(5-nitro-2-furyl)vinyl]−1,3,4-oxadiazole	+	0.990099	98.02	+	+	+	-	+	+
81	53757281	thiazole, 4-(5-nitro-2-furanyl)-	+	0.977042	95.41	+	+	+	-	+	+
82	17697551	1,1′-azoxypropane	-	0.087907	82.42	+	-	-	-	-	-
83	446866	azathioprine	+	0.717504	43.50	+	-	+	+	-	+
84	106514	quinone	-	0.087907	82.42	+	-	-	-	-	-
85	404864	capsaicin	-	0.088381	82.32	+	-	-	+	-	-
86	50594666	acifluorofen	+	0.766919	53.38	+	-	+	+	+	+
87	15805739	vinyl carbamate	+	0.781704	56.34	+	+	+	-	+	+
88	21928825	N6-methyl-N6-nitroso-adenosine	+	0.889640	77.93	+	+	+	(+)	+	+
89	77337543	*N*,*N*′-dipropyl-*N*-formylhydrazin	+	0.813184	62.64	+	(+)	(+)	+	-	+
90	133073	phaltan	?	0.585767	17.15	+	-	-	+	+	+
91	77094112	methyl-IQ	-	0.087907	82.42	+	-	-	-	-	-
92	57681	sulfamethazine	+	0.850460	70.09	+	+	+	+	+	+
93	21626891	diftalone	-	0.180560	63.89	+	-	-	-	+	-
94	999815	chlormequat chloride	+	0.786503	57.30	-	+	+	-	+	+
95	58333	promethazine hydrochloride	?	0.549015	9.80	-	+	+	-	-	+
96	834286	phenformin hydrochloride	-	0.088381	82.32	-	-	-	-	+	-
97	59333674	fluoxetine hydrochloride	-	0.304770	39.05	-	+	-	+	-	+
98	2198596	BASF ursol grey B	+	0.999161	99.83	-	+	+	-	m	+
99	624180	*p*-phenylenediamine dihydrochloride	-	0.087907	82.42	-	-	-	-	-	-
100	1465254	Marshall's reagent	-	0.087907	82.42	-	-	-	-	-	-
101	1934210	FD&C yellow no. 5	-	0.180560	63.89	-	(+)	-	-	-	?
102	6381777	erythorbic acid monosodium salt	-	0.087907	82.42	-	-	-	-	-	-
103	150389	EDTA, trisodium salt	-	0.127133	74.57	-	-	-	-	-	-
104	140567	sodium [4-(dimethylamino)phenyl]diazenesulfonate	-	0.139163	72.17	-	+	-	-	-	-
105	2438882	NSC 407047	?	0.642004	28.40	-	-	+	+	-	+
106	23135220	vydate	-	0.087907	82.42	-	-	-	-	-	-
107	116063	aldicarb	-	0.087907	82.42	-	-	-	-	-	-
108	2698411	*o*-chlorobenzylidenemalonic nitrile	?	0.366396	26.72	-	(+)	-	-	-	?
109	77474	hexachlorocyclopentadiene	-	0.087907	82.42	-	-	-	-	-	-
110	72435	methoxychlor	?	0.370894	25.82	-	-	(+)	+	+	+
111	77656	carbromal	-	0.000943	99.81	-	-	+	+	+	+
112	137268	thiram	-	0.293947	41.21	-	-	-	(+)	+	+
113	88960	phthalamide	-	0.000524	99.90	-	-	(+)	-	-	?

[a] Prediction made by ESP for overall carcinogenicity. [b] Confidence score calculated from the value at the output neuron of ANN. [c] Experimental overall carcinogenicity. [d] "−" inactive, "+" active, "m" marginal, "(+)" possibly active, "(−)" possibly inactive, "?" inconclusive. [e] Overall carcinogenicity predicted by MC4PC alone.

Table 5. Prediction Statistics Obtained Using WinESP and MC4PC Program Alone[a]

	WinESP	MC4PC alone
coverage (%)	91.15	92.11
concordance (%)	72.82	70.47
sensitivity (%)	66.22	74.03
specificity (%)	89.66	60.71

[a] Chemicals with inconclusive prediction were not considered in calculating the above statistics.

declared inconclusive if the confidence score obtained from the two output neurons is less than 30.0%.

The second improvement WinESP provides is a marked reduction in the false positives (specificity). It can be seen from Table 5 that specificity obtained by WinESP is 89.66%, whereas MC4PC gives only 60.71%. This is evident in the prediction of several test chemicals as shown in Table 4, e.g. chemical no. 14 which has been predicted as active by MC4PC in three modules and yet predicted as inactive by WinESP in agreement with the experimental result. Other such examples are chemical no. 16, 97, 111, 112, etc. This improvement can be attributed to the thorough analysis of the biophores by WinESP which eliminates consideration of insignificant biophores in making the prediction.

MC4PC performs better (74.03%) than WinESP (66.22%) in the prediction of actives (sensitivity). This is a result of increase sensitivity toward the quality of the biophores which

is very helpful in detecting false positives but results in somewhat increase in false negatives. WinESP fails to make a correct prediction for an active chemical whenever MC4PC could not find any biophore in all the four master modules. Notable examples are chemical no. 25, 27, 38, 39, etc. In such cases WinESP has nothing to enhance upon and consequently predicts the chemical as inactive. The root of this problem is in the master modules where the test chemical is insufficiently represented and can only be solved by increasing the diversity of the master modules.

WinESP also performed better (72.82%) than MC4PC alone (70.47%) in the overall correct predictions (concordance). The coverage which is the percentage of the test chemicals for which the result is not inconclusive is almost equal in both MC4PC and WinESP.

As pointed earlier, a confidence score can be calculated for the WinESP predictions based on the value at the output neuron. Table 4 lists the confidence score associated with each prediction. Although in the majority of the cases the score is very high, some of the test chemicals in Table 2 have very low scores. Notable examples are chemical no. 28, 69, 75, etc. Their pattern vectors contain signals that are not sufficient to distinguish between activity and inactivity and therefore resulted in values at the output neuron which are close to 0.5. We have labeled these results as inconclusive, and these chemicals should be considered for further evaluation.

CONCLUSIONS

A new method called ESP was developed to predict the toxicity/biological activity of chemicals based on multiple MC4PC modules. The design of the computer program and details of the methodology are discussed in the present paper. The method was successfully applied in predicting rodent carcinogenicity for 113 new chemicals. ESP contributes to a significant reduction in the false positive predictions in the test set as compared to predictions obtained without the use of ESP. ESP also performed better in the overall correct prediction of the test chemicals.

ESP uses an artificial neural network employing a genetic algorithm to learn the relationship between chemical structure and activity. After the training is complete, it can be used to predict the biological activity of a large number of chemicals and is therefore suitable for quick screening of large combinatorial libraries. The process is automatic and can easily be applied to different end points without many changes. ESP also employs a unique scoring system to rank biophores obtained from various MC4PC modules and provides an opportunity to identify important biophores and satisfactory prediction for new chemicals.

Since MC4PC has already been successfully employed to study numerous problems of the structure−activity relationship in toxicity and medicinal chemistry, we see WinESP as a very useful enhancement to the MC4PC methodology, providing an interpretation of the MC4PC results from a different point of view. Applications of WinESP to other toxicity end points such as mutagenicity and endocrine disruption and to useful therapeutic end points such as anti-HIV and anti-Parkinsonism are being studied and will be published in future papers.

ACKNOWLEDGMENT

The authors wish to thank Dr. Edwin J. Matthews, Center for Drug Evaluation and Research, U.S. FDA, for useful technical discussions. We are also grateful to the National Institute of Health for supporting their efforts through Grant no. 2R44CA090178-02.

REFERENCES AND NOTES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening: An Overview. *Drug Discovery Today* **1998**, *3*, 160−178.

(2) Ames, B. N.; McCann, H.; Yamasaki, E. Methods for Detecting Carcinogens and Mutagens with the Salmonella/Mammalian-Microsome Mutagenicity Test. *Mutat. Res.* **1975**, *31*, 347−364.

(3) Maron, D. M.; Ames B. N. Revised Methods for the Salmonella Mutagenicity Test. *Mutat. Res*. **1983**, *113*, 173−215.

(4) Gold, L. S.; Slone, T. H.; Ames, B. N. In *Handbook of Carcinogenic Potency and Genotoxicity Databases*; Gold, L. S., Zeiger, E., Eds.; CRC Press: Boca Raton, FL, 1997; Chapter 3, pp 621−623.

(5) Levin, D. E.; Yamasaki, E.; Ames, B. N. A New Salmonella Tester Strain, TA97, for the Detection of Frameshift Mutations. *Mutat. Res.* **1982**, *94*, 315−330.

(6) Prival, M. J.; Zeiger, E. Chemicals Mutagenic in *Salmonella typhimurium* Strain TA1535 but not in TA100. *Mutat. Res.* **1998**, *412*, 251−260.

(7) Benigni, R.; Richard, A. M. QSARs of Mutagens and Carcinogens: Two Case Studies Illustrating Problems in the Construction of Models for Non-Congeneric Chemicals. *Mutat. Res.* **1996**, *371*, 29−46.

(8) Mekapati, S. B.; Hansch, C. On the Parametrization of the Toxicity of Organic Chemicals to Tetrahymena pyriformis. The Problem of Establishing a Uniform Activity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 956−961.

(9) Enslein, K.; Gombar, V. K.; Blake, B. W. Use of SAR in Computer-Assisted Prediction of Carcinogenicity and Mutagenicity of Chemicals by the TOPKAT Program. *Mutat. Res.* **1994**, *305*, 47−61.

(10) Sanderson, D. M.; Earnshaw, C. G. Compuer Prediction of Possible Toxic Action from Chemical Structure; the DEREK System. *Human Exp. Toxicol*. **1991**, *10*, 261−273.

(11) Woo, Y. T.; Lai, D.; Argus, M.; Arcos, J. Development of Structure Activity Relationship Rules for Predicting Carcinogenic Potential of Chemicals. *Toxicol. Lett*. **1995**, *79*, 219−228.

(12) Klopman, G. Computer Automated Structure Evaluation of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7324.

(13) Klopman, G. MULTI−CASE 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct. Act. Relat.* **1992**, *11*, 176−184.

(14) Klopman, G.; Rosenkranz, H. S. Approaches to SAR in Carcinogenesi and Mutagenesis Prediction of Carcinogenicity/Mutagenicity using MULTI−CASE. *Mutat. Res.* **1994**, *305*, 33−46.

(15) Liu, M.; Sussman, G.; Klopman, G.; Rosenkranz, H. S. Structure−Activity and Mechanistic Relationships: the Effect of Chemical Overlap on Structural Overlap in Data Bases of Varying Size and Composition. *Mutat. Res.* **1996**, *372*, 79−85.

(16) Zhang, Y. P.; Sussman, N.; Macina, O. T.; Rosenkranz, H. S.; Klopman, G. Prediction of the Carcinogenicity of a Second Group of Organic Chemicals Undergoing Carcinogenicity Testing. *Environ. Health. Perspect.* **1996**, *104 (Suppl. 5)*, 1045−1050.

(17) 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.* **1998**, *28*, 242−264.

(18) Gold, L. S.; Slone, T. H.; Ames, B. N. In *Handbook of Carcinogenic Potency and Genotoxicity Databases*; Gold, L. S., Zeiger, E., Eds.; CRC Press: Boca Raton, FL, 1997; Chapter 3, pp 624−657.

(19) Klopman, G.; Chakravarti, S. K.; Harris, N.; Ivanov, J.; Saiakhov, R. D. In−Silico Screening of High Production Volume Chemicals for Mutagenicity Using the MCASE QSAR Expert System. *SAR QSAR Environ. Res.* **2003**, *14*, 165−180.

(20) Klopman, G.; Chakravarti, S. K. Structure−Activity Relationship Study of a Diverse Set of Estrogen Receptor Ligands (I) Using MultiCASE Expert System. *Chemosphere* **2003**, *51*, 445−459.

(21) Klopman, G.; Chakravarti, S. K. Screening of High Production Volume Chemicals for Estrogen Receptor Binding Activity (II) By the MultiCASE Expert System. *Chemosphere* **2003**, *51*, 461−468.

(22) Klopman, G.; Tu, M. Diversity Analysis of 14,156 Molecules Tested by the National Cancer Institute for Anti-HIV Activity Using the Quantitative Structure−Activity Relational Expert System MCASE. *J. Med. Chem.* **1999**, *42*, 992−998.

(23) So, S. S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: an Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.
(24) Burns, J. A.; Whitesides, G. M. Feed-Forward Neural Networks in Chemistry: Mathematical Systems for Classification and Pattern Recognition. *Chem. Rev.* **1993**, *93*, 2583.
(25) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: An Introduction*; VCH: Weinheim, 1993.
(26) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Learning Internal Representations by Error Propagation*; MIT Press: Cambridge, MA, 1986; Parallel distributed processing, Vol. 1.
(27) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley Publishing Co.: Redwood City, CA, 1991.
(28) Holland, J. H. *Adaption in Natural and Artificial Systems*; The University of Michigan Press: Ann Arbor, MI, 1975.
(29) Cartwright, H. M. *Applications of Artificial Intelligence in Chemistry*; Oxford University Press: Oxford, 1993.
(30) van Rooij, A. J. F.; Jain, L. C.; Johnson, R. P. *Neural Network Training Using Genetic Algorithms*; World Scientific: Singapore, River Edge, NJ, 1996; Vol. 26, Chapter 4, pp 56−75.
(31) Preussmann, R.; Stewart, B. W. In *Chemical Carcinogens*; Searle, C. E., Ed.; American Chemical Society: Washington, DC, 1984; Vol. 2, Chapter 12, pp 653−677.
(32) Zedeck, M. S. In *Chemical Carcinogens*; Searle, C. E., Ed.; American Chemical Society: Washington, DC, 1984; Vol. 2, Chapter 15, pp 919−923.

CI030298N