JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

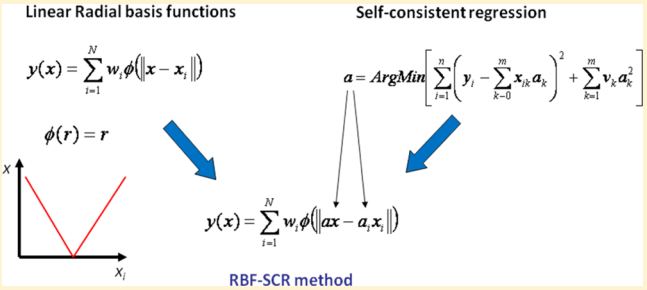
A New Approach to Radial Basis Function Approximation and Its Application to QSAR

Alexey V. Zakharov,[†] Megan L. Peach,[‡] Markus Sitzmann,[†,§] and Marc C. Nicklaus*[,†]

[†]CADD Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, DHHS, NCI-Frederick, , 376 Boyles St., Frederick, Maryland 21702, United States

[‡]Basic Science Program, Leidos Biomedical, Inc., Computer-Aided Drug Design Group, Chemical Biology Laboratory, Frederick National Laboratory for Cancer Research, 376 Boyles St., Frederick, Maryland 21702, United States

ABSTRACT: We describe a novel approach to RBF approximation, which combines two new elements: (1) linear radial basis functions and (2) weighting the model by each descriptor's contribution. Linear radial basis functions allow one to achieve more accurate predictions for diverse data sets. Taking into account the contribution of each descriptor produces more accurate similarity values used for model development. The method was validated on 14 public data sets comprising nine physicochemical properties and five toxicity endpoints. We also compared the new method with five different QSAR methods implemented in the EPA T.E.S.T. program. Our approach, implemented in the program GUSAR, showed a reasonable accuracy of prediction and high coverage for all external test sets, providing more accurate prediction results than the comparison methods and even the consensus of these methods. Using our new method, we have created models for physicochemical and toxicity endpoints, which we have made freely available in the form of an online service at http://cactus.nci.nih.gov/chemical/apps/cap.



INTRODUCTION

The aim of drug discovery is to find promising compounds that show good potency and selectivity against selected targets. Potential hits need to have reasonable properties in the areas of absorption, distribution, metabolism, elimination, and toxicity (ADMET).[1] Compounds must be absorbed by the human body, be transported to the target, and then interact with the target receptors or enzymes. To reach a tissue, the compound usually is taken up into the bloodstream, which usually occurs via mucous surfaces such as the digestive tract (intestinal absorption). Factors such as poor compound solubility, gastric emptying time, intestinal transit time, chemical instability in the stomach, and inability to penetrate the intestinal wall can all reduce drug absorption after oral administration. Solubility of compounds in turn depends on solvent properties. A solvent's main physicochemical characteristics, which can be classified as surface versus transport properties of a liquid, are surface tension, viscosity, and thermal conductivity.[2] The magnitudes of these properties are dependent upon intermolecular interactions between the solvent molecules. Therefore, the physicochemical properties of chemical structures play important roles in the design and optimization of ADMET parameters of potential drug compounds.[3]

In addition to possessing balanced ADME properties, potential drug candidates should provide the desirable effect while avoiding toxicity and side effects. Toxicity is considered one of the most important factors for success or failure in drug development.[4] Toxic and unwanted side effects observed in drug candidates or marketed drugs can be caused by many different modes of action such as interactions with enzymes, receptors, and ion channels.[5]

Although experimental testing of ADME properties and toxicity is generally thought to provide the most reliable data about the interaction of a given compound with a biological system, it is very time consuming and expensive and thus not suitable for the screening of large sets of compounds.[6] For that purpose, several different computational approaches based on quantitative structure−activity relationships (QSAR) have been used[7−10] instead.

The main idea of QSAR methods is to describe relationships between activity measures and structural descriptors of compounds and to create models that can be used for prediction of the same activity for new compounds. There are numerous different techniques that have been used for this task. Among the currently most widely used ones are Support Vector Machines,[7] Random Forests,[9] and Artificial Neural Networks (ANN).[11] These methods allow the creation of nonlinear models that can successfully describe the structure−activity relationships of multi-faceted properties such as ADME, physicochemical properties, and toxicity. ANNs typically provide better results in comparison to other methods[8] due to their capability for construction of nonlinear models of any

Received: November 29, 2013
Published: January 22, 2014

   dx.doi.org/10.1021/ci400704f | J. Chem. Inf. Model. 2014, 54, 713−719

level of complexity, especially in cases where the general form of the analytic dependence is unknown.

The process of neural network training includes two steps. First, the network architecture, which includes the number of hidden layers and neurons, needs to be constructed. Second, the network parameters associated with the neurons are determined using different optimization algorithms, which try to minimize the errors of the network's predictions compared to the observed values in the training set examples. This is the training procedure, during which the network learns the relationships between the input and output variables. Most ANN learning algorithms require a lot of computational time due to their optimization methods.[12]

Radial basis function (RBF) neural networks form a class of ANNs that has certain advantages over other types of ANNs, including better approximation capabilities, a simpler network architecture, and faster learning algorithms.[12] The main idea of RBF neural networks is to create the proper number of hidden neurons and determine the weight of each neuron. Functions that depend only on the distance from a center vector of the neuron (calculated in descriptor space) and are radially symmetric around that vector are called radial basis functions. They provide a nonlinear approximation of the input data. Often the selection of the center vectors of the neurons is performed with different clustering methods. Some of them require setting up the initial number of neurons (centroid-based clustering), and other methods calculate the optimal number of neurons (distribution-based clustering). After selection of the neurons, it is necessary to calculate the weight of each neuron. For this purpose, the simple least-squares method can be used.

Although RBF networks are a very powerful approach, one disadvantage has to be mentioned. As with most ANN methods, RBF neural networks need to select the hidden neurons, which is both ambiguous and sometimes poorly reproducible. To avoid this problem, the radial basis function interpolation approach can be applied. The difference between the radial basis function interpolation approach and a general RBF network is that the former has a number of hidden neurons equal to the number of input variables (training set members), whereas the latter has a significantly reduced number. Thus, the learning procedure of the RBF interpolation approach uses all the elements in the training set. However, the RBF interpolation approach can be sensitive to noise created by both a huge number of descriptors and low-quality data.

Earlier, we had shown that self-consistent regression (SCR) could successfully be used to generate models from a large number of descriptors for different noise levels in the data.[13] In this work, we propose a new approach that combines the advantages of both the RBF interpolation and self-consistent regression methods.[14] We call this approach RBF-SCR. We compare the RBF-SCR method with the radial basis function interpolation method and RBF neural networks with *k*-means clustering. For these comparisons, 14 publicly available data sets were used: nine data sets with physicochemical properties and five data sets with toxicity endpoints. In addition, we compare the RBF-SCR method with different QSAR methods implemented in the U.S. Environmental Protection Agency (EPA) T.E.S.T. program on the same data sets. All QSAR models developed with the RBF-SCR method have been made freely available in our Chemical Activity Predictor Web service: http://cactus.nci.nih.gov/chemical/apps/cap.

MATERIALS AND METHODS

Data Sets. All 14 data sets were downloaded from the EPA Web site.[15] Nine are related to physicochemical properties and five to toxicity. Each data set includes training and test sets. We used the same partitioning of the training and test sets as presented on the EPA Web site. To allow unbiased comparison, the data was not curated, and we used it "as is." Each data set is briefly described in the following.

Physicochemical Data Sets. *Boiling Point.* The normal boiling point is defined as the temperature at which a chemical boils at atmospheric pressure. The total number of compounds was 5758. The training set included 4607 compounds, and the test set contained 1151 compounds. The modeled property is the boiling point in °C, which varied from −128 to 548 °C in the EPA set.

Density. The total number of compounds was 8908. The training set included 7125 compounds and the test set contained 1783 compounds. The modeled property is the density in g/cm^3, and it varied from 0.53 to 4.008 g/cm^3 in the EPA set.

Flash Point. The flash point of a chemical is defined as the lowest temperature in °C at which it can vaporize to form an ignitable mixture in air. The total number of compounds was 8362. The training set included 6690 compounds, and the test set contained 1672 compounds. The modeled property is the temperature in °C, and it ranged from −136 to 902.8 °C in the EPA set.

Thermal Conductivity. The thermal conductivity is defined as the property of a material in units of mW/(m·K) reflecting its ability to conduct heat. The total number of compounds was 442. The training set included 352 compounds, and the test set contained 90 compounds. The modeled property ranged from 35.55 to 352 mW/(m·K) in the EPA set.

Viscosity. The viscosity is defined as a measure of the resistance of a fluid to flow in cP defined as the proportionality constant between shear rate and shear stress. The total number of compounds was 557. The training set included 444 compounds, and the test set contained 113 compounds. The modeled property ranged from −0.859 to 2.975 cP in the EPA set.

Surface Tension. The surface tension is defined as a property of the surface in dyn/cm of a liquid that allows it to resist an external force. The total number of compounds was 1416. The training set included 1133 compounds, and the test set contained 283 compounds. The modeled property ranged from 9.42 to 66.178 dyn/cm in the EPA set.

Water Solubility. The water solubility is defined as the amount of a chemical that will dissolve in liquid water to form a homogeneous solution. The total number of compounds was 5020. The training set included 4016 compounds, and the test set contained 1004 compounds. The modeled property ranged from −1.494 to 13.172 mg/L in the EPA set.

Vapor Pressure. The vapor pressure is defined as the pressure of a chemical's vapor in mmHg in thermodynamic equilibrium with its condensed phases in a closed system. The total number of compounds was 2510. The training set included 2006 compounds, and the test set contained 504 compounds. The modeled property ranged from −17.699 to 5.243 mmHg in the EPA set.

Melting Point. The melting point is defined as the temperature in °C at which a chemical in the solid state changes to a liquid state. The total number of compounds was

dx.doi.org/10.1021/ci400704f | J. Chem. Inf. Model. 2014, 54, 713−719

9384. The training set included 7509 compounds, and the test set contained 1875 compounds. The modeled property ranged from −196 to 492.5 mmHg in the EPA set.

Toxicity Data Sets. *Fathead Minnow.* The fathead minnow LC_{50} endpoint represents the concentration in water which kills half of a population of fathead minnows (*Pimephales promelas*) in 4 days (96 h). The total number of compounds was 823. The training set included 659 compounds, and the test set contained 164 compounds. The experimental data are represented by $-\mathrm{Log}_{10}$ (LC_{50} [mol/L]) and ranged from 0.037 to 9.261 in the EPA set.

Daphnia magna. The *Daphnia magna* LC_{50} endpoint represents the concentration in water which kills half of a population of *Daphnia magna* (a water flea) in 48 h. The total number of compounds was 353. The training set included 283 compounds, and the test set contained 70 compounds. The modeled property is $-\mathrm{Log}_{10}$ (LC_{50} [mol/L]) and varied from 0.117 to 10.064 in the EPA set.

Tetrahymena pyriformis. The *Tetrahymena pyriformis* IGC_{50} endpoint represents the 50% growth inhibitory concentration of the *T. pyriformis* organism (a protozoan ciliate) after 40 h. The total number of compounds was 1792. The training set included 1434 compounds, and the test set contained 358 compounds. The modeled property is $-\mathrm{Log}_{10}$ (IGC_{50} [mol/L]) and varied from 0.334 to 6.36 in the EPA set.

Oral Rat Acute Toxicity. The oral rat LD_{50} endpoint represents the amount of the chemical (mass of the chemical per body weight of the rat) which when orally ingested kills half of the rats. The total number of compounds was 7413. The training set included 5931 compounds, and the test set contained 1482 compounds. The modeled property is $-\mathrm{Log}_{10}$ (LD_{50} [mol/kg]) and varied from 0.291 to 7.207 in the EPA set.

Bioconcentration Factor. The bioconcentration factor (BCF) is defined as the ratio of the chemical concentration in biota as a result of absorption via the respiratory surface to that in water at steady state. The total number of compounds was 676. The training set included 541 compounds, and the test set contained 135 compounds. The modeled property is Log_{10}(BCF) and varied from −1.7 to 5.694 in the EPA set.

METHODS

Descriptors. The QSAR models in this study were developed using the GUSAR program.[16,5,17] GUSAR uses a combination of three types of descriptors: whole-molecule descriptors, QNA (Quantitative Neighborhoods of Atoms) descriptors,[14] and descriptors based on predictions from the PASS algorithm for predicting the biological activity spectra of compounds.[18]

The whole-molecule descriptors used in GUSAR are topological length, topological volume, lipophilicity, number of positive charges, number of negative charges, number of hydrogen bond acceptors, number of hydrogen bond donors, number of aromatic atoms, molecular weight, and number of halogen atoms.

QNA descriptors are defined by two functions, P and Q. The values for P and Q for each atom i are calculated as

$$P_i = B_i \sum_k \left(\exp\left(-\frac{1}{2}C\right)\right)_{ik} B_k \tag{1}$$

$$Q_i = B_i \sum_k \left(\exp\left(-\frac{1}{2}C\right)\right)_{ik} B_k A_k \tag{2}$$

where the k are all other atoms in the molecule and

$$A_k = \frac{1}{2}(\mathrm{IP}_k + \mathrm{EA}_k),\ B_k = (\mathrm{IP}_k - \mathrm{EA}_k)^{-1/2} \tag{3}$$

IP is the ionization potential, EA is the electron affinity for each atom, and C is the connectivity matrix for the molecule as a whole.[14] Two-dimensional Chebyshev polynomials are used for approximating the functions P and Q over all atoms of the molecule.

The PASS biological descriptors are calculated using the PASS algorithm,[18] which predicts a wide range of biological outcomes including transporter protein binding, gene expression activities, and various mechanisms of action, adding up to about 6400 "biological activities" at a mean prediction accuracy threshold of at least 95%. The output from PASS is the probability, for each predicted outcome, that the compound will be active (P_a) and the probability that it will be inactive (P_i). The difference between these two values ($P_a - P_i$), for a randomly selected subset of the predicted activities, was used as a molecular descriptor for the regression analysis in GUSAR.

RBF Neural Networks. We used Radial Basis Function networks with three layers: an input layer, a hidden layer with a nonlinear RBF activation function, and a linear output layer. The input is represented as a vector of real numbers $x \in R^n$. The output of the network is then a scalar function of the input vector, $y{:}R^n \rightarrow R$, and is determined by

$$y(x) = \sum_{i=1}^{N} w_i \phi(\| x - c_i \|) \tag{4}$$

where N is the number of neurons in the hidden layer, c_i is the center vector for neuron i, and w_i is the weight of neuron i in the linear output neuron.

To determine center vectors (centroids), k-mean clustering was used. There are many radial basis functions that can be applied to construct neural networks: linear, Gaussian, multiquadric, polyharmonic spline, etc. In this work, we decided to compare two RBF functions: Gaussian and linear radial basis functions. For this purpose, we have used Gaussian radial basis functions for the RBF NN and linear functions for the RBF interpolation and RBF-SCR methods.

RBF Interpolation. As mentioned above, the difference between an RBF interpolation and an RBF network is that the first method uses each input variable as a centroid. Therefore, the learning process is performed across all input variables in the training set

$$y(x) = \sum_{i=1}^{N} w_i \varphi(\| x - x_i \|) = \Phi w \tag{5}$$

where the approximating function $y(x)$ is represented as a sum of N radial basis functions, each associated with a different center x_i, and weighted by an appropriate coefficient w_i.

If the points x_i are distinct, then the interpolation matrix Φ in the above equation is nonsingular. Thus, the weights w can be solved simply by

$$w = \Phi^{-1} y \tag{6}$$

dx.doi.org/10.1021/ci400704f | J. Chem. Inf. Model. 2014, 54, 713−719

To find the weights, a simple least-squares method was used. A linear radial basis function was applied for the RBF interpolation.

RBF-SCR. We showed earlier that self-consistent regression (SCR) can be successfully applied to different QSAR tasks.[14,18,19,16] The basic purpose of the SCR method is to remove those variables that poorly describe the target value. The final number of variables in the QSAR equation selected after the SCR procedure is significantly smaller compared to the initial number of variables. Also, it has been shown that SCR is robust against noise in the data.[13] Self-consistent regression is a regularized least-squares method

$$a = \mathrm{ArgMin}[\sum_{i=1}^{n}(y_i - \sum_{k=0}^{m} x_{ik}a_k)^2 + \sum_{k=1}^{m} v_k a_k^2] \quad (7)$$

where a is the regression coefficient, n is the number of objects, y_i is the response value of the i^{th} object, m is the number of independent variables, x_{ik} is the value of the k^{th} independent variable of the i^{th} object, a_k is the k^{th} value of the regression coefficients, and v_k is the k^{th} value of the regularization parameters.

Equation 7 has the following solution

$$a = \mathbf{T}\mathbf{X}^T y,\ \mathbf{T} = (\mathbf{X}^T\mathbf{X} + \mathbf{V})^{-1}$$

where $\mathbf{X}^T$ is the transposed regression matrix of $\mathbf{X}$, and $\mathbf{V}$ is the diagonal matrix of the regularization parameters.

The regression coefficients, obtained from SCR, reflect the contribution of each particular descriptor (variable) to the final equation. The higher the absolute value of the coefficient, the greater its contribution. Thus, regression coefficients obtained after SCR can be used for weighting of descriptors (variables) according to their importance. We used this advantage to create a new machine learning approach that combines self-consistent regression with the radial basis function interpolation method, which we therefore call RBF-SCR.

Typically, the radial basis function is calculated using the Euclidean distance (similarity) between descriptor vectors and centroids. In the case of RBF interpolation, the same type of distance (similarity) is calculated between input vectors of descriptors. If one takes into account the contribution of each descriptor, a more accurate distance (similarity) value can be obtained and thus a more accurate prediction be achieved. For this purpose, the descriptors are weighted during the calculation of the radial basis functions by the coefficients obtained from SCR. Thus, RBF-SCR can be expressed as the equation

$$y(x) = \sum_{i=1}^{N} w_i \varphi(\| ax - a_i x_i \|) \quad (8)$$

where a is taken from eq 7 (SCR). The weights a_i are the novel element compared with eq 5.

Therefore, RBF-SCR can be described as a 3-step algorithm: (1) Self-consistent regression determines coefficients and selects descriptors. (2) Radial basis functions are calculated using similarity that is weighted by SCR coefficients. (3) RBF weights are determined by the least-squares method.

A general scheme of the RBF-SCR approach is shown in Figure 1.

A linear radial basis function was used for the RBF interpolation and RBF-SCR methods. In contrast to the commonly used Gaussian function, a linear function has the effect that the more dissimilar the input compounds (represented by descriptor vectors) are, the more contribution they provide (Figure 2).

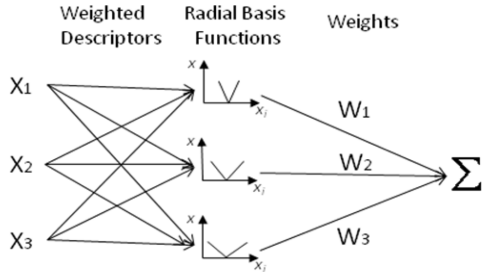


Figure 1. Schematic representation of the RBF-SCR radial basis function approach.

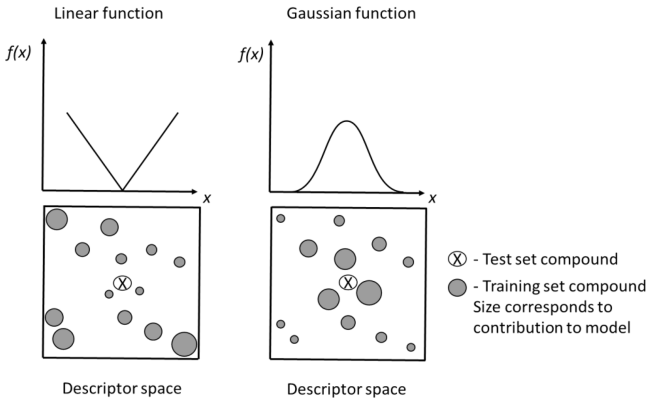


Figure 2. Representation of linear and Gaussian functions in the descriptor space.

Thus, the linear radial basis function can be used for modeling of diverse training sets with a high level of dissimilarity between the objects. We have implemented the RBF-SCR method in the program GUSAR.

Applicability Domain Estimation. GUSAR uses three different approaches for estimation of model applicability domains: similarity, leverage, and accuracy assessment.

Similarity. For each compound, the pairwise distance to each of its three nearest neighbors in the training set is calculated using Pearson's correlation coefficient in the space of the independent variables obtained after SCR. The compound is considered to be in the applicability domain of the model if the average of these three distances is less than or equal to 0.7.

Leverage. Leverage calculations are a method for identifying outliers based on the contribution of each molecule to its own predicted value

$$\text{Leverage} = x^T(\mathbf{X}^T\mathbf{X})^{-1}x$$

where x is the vector of the descriptors for a test compound, and $\mathbf{X}$ is the matrix formed from the rows corresponding to the descriptors of all the molecules in the training set. A compound is considered outside the applicability domain of a model if its leverage is higher than the 99th percentile in the distribution of the leverage values calculated for the training set.

Accuracy Assessment. Here, the applicability domain prediction for each compound is calculated based on the error of prediction for the three most similar compounds in the

dx.doi.org/10.1021/ci400704f | J. Chem. Inf. Model. 2014, 54, 713−719

Table 1. Comparison of QSAR Models Generated by GUSAR with Different RBF Methods[a]

activity name	RBF NN (Gaussian functions) R^2	RBF interpolation (linear functions) R^2	RBF-SCR (linear functions) R^2	RBF NN (Gaussian functions) RMSE	RBF interpolation (linear functions) RMSE	RBF-SCR (linear functions) RMSE
Physicochemical						
boiling point (°C)	0.84	0.95	0.95	34.63	20.11	**19.50**
density (g/cm^3)	0.93	0.97	0.97	0.08	0.06	**0.05**
flash point (°C)	0.78	**0.89**	0.88	39.13	**27.20**	28.51
thermal conductivity (mW/(m·K))	0.85	0.90	**0.93**	15.39	12.20	**10.08**
viscosity ($\log_{10}$(cP))	0.65	0.87	**0.88**	0.34	0.22	**0.20**
surface tension (dyn/cm)	0.83	0.88	**0.93**	2.86	2.43	**1.86**
water solubility ($\log_{10}$(mol/L))	0.83	0.87	0.87	0.90	**0.79**	0.80
vapor pressure ($\log_{10}$(mmHg))	0.86	0.95	0.95	1.34	0.82	**0.80**
melting point (°C)	0.77	0.86	0.86	49.22	**37.73**	37.97
Toxicity						
Fathead minnow, ($-\log_{10}(LC_{50})$)	0.67	0.73	**0.74**	0.84	0.76	0.76
Daphnia magna ($-\log_{10}(LC_{50})$)	0.57	0.60	**0.61**	1.16	1.10	**1.07**
Tetrahymena pyriformis ($-\log_{10}(IGC_{50})$)	0.70	0.81	**0.82**	0.55	0.43	**0.41**
oral rat acute toxicity ($-\log_{10}(LD_{50})$)	0.56	0.66	0.66	0.64	0.56	0.56
bioconcentration factor ($\log_{10}$(BCF))	0.73	0.77	**0.78**	0.71	0.66	0.65

[a]RBF-SCR is the novel method proposed in this article. Best model parameter for each endpoint is shown in bold.

training set (see the similarity metric above) relative to the training set as a whole

$$AD_{value} = RMSE_{3NN}/RMSE_{train}$$

In this study a threshold of 1 was used for the AD_{value}.

Consensus Modeling. The final predicted values for each physicochemical and toxicity endpoint are calculated using a weighted average of the predictions from several different QSAR models. Each model is based on a different set of QNA and "biological" descriptors, and its predictions for each compound are weighted according to the similarity value as calculated during the applicability domain assessment.

EPA T.E.S.T. Program. We compared our approach with well-known methods implemented in the T.E.S.T. (Toxicity Estimation Software Tool) program version 4.1 provided by the EPA.[15] This program includes models obtained using several QSAR methods.

Hierarchical Method. This method uses a weighted average of the predictions from several different models. The different models are obtained by using Ward's method to divide the training set into a series of structurally similar clusters. A genetic algorithm-based technique is used to generate models for each cluster.

FDA Method. This is an on-the-fly model that is fit to the chemicals that are most similar to the test compound.

Single Model Method. This is a multi-linear regression model that is fit to the training set (using molecular descriptors as independent variables) using a genetic algorithm-based approach. The regression model is generated prior to runtime.

Group Contribution Method. This method is a multi-linear regression model that is fit to the training set (using molecular fragment counts as independent variables). The regression model is generated prior to runtime.

Nearest Neighbor Method. This method uses an average value for the three chemicals in the training set that are most similar to the test chemical.

Consensus Method. This method uses the average of the predicted toxicities from all of the above QSAR methods (provided the predictions are within their respective applicability domains).

The T.E.S.T. program contains 797 two-dimensional descriptors spanning the following descriptor classes: E-state values and E-state counts, constitutional descriptors, topological descriptors, walk and path counts, connectivity, information content, 2D autocorrelation, Burden eigenvalue, molecular properties (such as the octanol−water partition coefficient), Kappa, hydrogen bond acceptor/donor counts, molecular distance edge, and molecular fragment counts.

We compared our results with those provided by the T.E.S.T. program as described in the User's Guide for T.E.S.T. (version 4.1).[20]

RESULTS

Comparison of RBF Methods (GUSAR) against Each Other. For each training set, 20 models based on "biological" descriptors and 20 models based on QNA descriptors were created using each of the three RBF methods implemented in GUSAR: RBF-SCR, RBF interpolation, and RBF neural network with *k*-means clustering. Thus, 40 QSAR models were created for each RBF method. These models were used for the consensus prediction of the test sets. Thus, three consensus prediction results were obtained from the three RBF methods for each particular test set. To compare the prediction results obtained for each test set, R^2 and RMSE were calculated (Table 1).

As shown in Table 1, the radial basis function (Gaussian functions) neural networks (RBF NN) showed poor results in comparison with RBF interpolation and RBF-SCR (linear functions). The reason for this is that RBF NNs are based on centroids, whose number is significantly lower than the number of objects in the training sets. In addition, RBF NNs are based on Gaussian functions, which are not as powerful for diverse data sets as the linear functions used in the RBF interpolation and RBF-SCR methods. Also, selection of the centroids is dependent on the clustering method, which is not always adequate.

The R^2 values for the test sets varied from 0.60 to 0.97 for both RBF interpolation and RBF-SCR. Thus, both methods showed good results for all types of activities. RBF-SCR showed the best results in terms of RMSE in eight cases and in terms of

dx.doi.org/10.1021/ci400704f | J. Chem. Inf. Model. 2014, 54, 713−719

Table 2. Comparison of the Results of GUSAR Using RBF-SCR with Those of the T.E.S.T. Program

endpoint	hierarchical, T.E.S.T.	single model, T.E.S.T.	FDA, T.E.S.T.	group contribution, T.E.S.T.	nearest neighbor, T.E.S.T.	T.E.S.T. consensus	RBF-SCR, GUSAR
	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE
	coverage	coverage	coverage	coverage	coverage	coverage	coverage
Physicochemical							
boiling point (°C)	18.70	N/A	21.43	27.55	29.97	19.40	**18.66**
	0.935	N/A	0.988	0.977	0.988	**0.988**	0.981
density (g/cm^3)	0.05	N/A	0.06	0.12	0.12	0.07	0.05
	0.942	N/A	0.992	0.992	0.997	0.996	**1.000**
flash point (°C)	28.90	N/A	31.48	33.63	36.83	28.50	**26.00**
	0.924	N/A	0.989	0.987	**0.993**	0.992	0.953
thermal conductivity (mW/(m·K))	11.02	11.86	16.41	15.90	12.83	12.41	**10.08**
	0.956	0.956	0.967	0.911	0.978	0.967	**1.000**
viscosity ($\log_{10}$(cP))	0.21	0.35	0.21	0.20	0.29	0.22	0.20
	0.929	0.929	0.929	0.814	0.920	0.929	**1.000**
surface tension (dyn/cm)	**1.79**	N/A	2.22	2.93	3.32	2.11	1.85
	0.919	N/A	0.979	0.926	0.936	0.968	**0.993**
water solubility ($\log_{10}$(mol/L))	0.90	N/A	0.95	1.07	1.02	0.84	**0.73**
	0.935	N/A	0.984	0.982	0.985	**0.987**	0.950
vapor pressure ($\log_{10}$(mmHg))	0.75	N/A	0.83	1.00	1.25	0.77	**0.73**
	0.940	N/A	**0.982**	0.968	0.980	0.980	0.935
melting point (°C)	44.36	N/A	45.10	54.95	52.10	41.46	**37.52**
	0.932	N/A	**0.993**	0.997	0.998	0.998	0.979
Toxicity							
Fathead minnow ($-\log_{10}(LC_{50})$)	0.80	0.80	0.92	0.81	0.88	0.77	**0.76**
	0.951	0.945	0.945	0.872	0.939	0.951	**1.000**
Daphnia magna ($-\log_{10}(LC_{50})$)	0.98	0.99	1.19	**0.80**	0.98	0.91	1.07
	0.886	0.871	0.900	0.657	0.871	0.900	**1.000**
Tetrahymena pyriformis ($-\log_{10}(IGC_{50})$)	0.54	N/A	0.49	0.58	0.64	0.48	**0.41**
	0.933	N/A	0.978	0.955	0.986	0.983	**0.989**
oral rat acute toxicity ($-\log_{10}(LD_{50})$)	0.65	N/A	0.66	N/A	0.66	0.59	**0.55**
	0.876	N/A	0.984	N/A	**0.993**	0.984	0.960
bioconcentration factor ($\log_{10}$(BCF))	0.71	0.68	0.75	0.76	0.88	0.66	**0.64**
	0.926	0.926	0.911	0.874	0.948	0.926	**1.000**

RMSE: root-mean-square error. The highest coverage and accuracy values are highlighted in bold.

R^2 for seven cases out of 14. RBF interpolation was better than the other methods for three cases in terms of RMSE and for one case in terms of R^2. Thus, RBF-SCR provided more accurate predictions in comparison with RBF interpolation.

Comparison of RBF-SCR vs T.E.S.T. In addition, we compared results obtained with RBF-SCR to the results provided by FDA for the T.E.S.T. program.

To select the most predictive models obtained by RBF-SCR, a leave-10%-out cross-validation procedure was performed 20 times for each model. From the full set of 40 models, we selected only those models that satisfied the following conditions: a value of Q^2 exceeding 0.6 and a R^2 value from the leave-10%-out cross-validation procedure exceeding 0.5. The selected models were used for consensus prediction of the external test set of each activity/endpoint, taking into account the applicability domain of these models.

The methods implemented in T.E.S.T. provide only predictions that fall in the applicability domain of each model. Thus, for the same test set the various methods have different coverage. Direct comparison of the RBF-SCR method and the methods realized in T.E.S.T. is therefore not possible. However, we performed an indirect comparison taking into account both accuracy of prediction and coverage. For this indirect comparison, we analyzed which method showed higher coverage and/or better accuracy of prediction across the 14 test sets by determining both RMSE and coverage values. One can see that the T.E.S.T. consensus results are in most cases better than the results obtained by each individual T.E.S.T. method but still worse than the results achieved by RBF-SCR, which showed better results in terms of RMSE values for 10 data sets out of 14. The coverage provided by RBF-SCR for the test sets was better in eight out of 14 cases. Thus, on average the RBF-SCR method provides more accurate prediction results than the T.E.S.T. program methods and even the consensus of these methods.

Chemical Activity Predictor Web Service. Utilizing the QSAR models created with the RBF-SCR method, we have developed a freely available online service for the simultaneous prediction of the nine physicochemical properties and five toxicity endpoints described in this paper, available at http://cactus.nci.nih.gov/chemical/apps/cap. We have named this service Chemical Activity Predictor. It provides two different ways to input chemical structures. The first one is a classical online chemical editor, which allows drawing of the desired structure. The second one is based on our NCI/CADD Chemical Identifier Resolver technology and allows the input of different types of structure identifiers: InChIKey, drug names, SMILES, IUPAC names, etc. The service permits the user to input several structures simultaneously. As output, predictions of the nine physicochemical properties and five toxicity endpoints are provided for each compound. In addition, our service estimates and outputs the applicability domain of each

dx.doi.org/10.1021/ci400704f | J. Chem. Inf. Model. 2014, 54, 713−719

QSAR model. This calculation is performed for each compound with the result that each prediction is annotated with either "In AD" or "Out of AD", indicating whether one can be confident in the prediction or not. Our service performs at a reasonable computational speed (about one compound per second for the simultaneous prediction of 14 endpoints). The interpretation of the prediction results for the implemented endpoints may be done in the same way as an assessment for in vitro/in vivo experimental assays.

CONCLUSIONS

We have developed a new RBF-SCR method and have compared it with other machine learning approaches. The two crucial novel elements of this method are (a) introduction of weights for each descriptor vector used for calculation of RBF based on that descriptor's importance for the given activity as determined by SCR and (b) linear basis function used for better description of diverse data sets. A method comparison was performed on 14 data sets comprising nine physicochemical properties and five toxicity endpoints. We showed that the RBF-SCR method provides more accurate prediction results than other methods including even consensus predictions of these methods. We believe that QSAR models developed with the RBF-SCR method could successfully be used for the design and optimization of ADMET properties of potential drug compounds. We hope that our freely available online service for quantitative prediction of physicochemical properties and toxicity endpoints based on these models may be useful for researchers in their quest of finding drug-like compounds with desirable ADMET properties. It can also be used for optimizing compounds with regards to several different endpoints simultaneously.

AUTHOR INFORMATION

Corresponding Author

*E-mail: mn1@helix.nih.gov. Telephone: +1-301-846-5903.

Present Address

[§]Marcus Sitzmann: FIZ Karlsruhe − Leibniz-Institut für Informationsinfrastruktur, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are thankful to Dmitry Filimonov for many helpful suggestions made in the course of this project. This project has been funded in part with federal funds from the Frederick National Laboratory for Cancer Research, National Institutes of Health, under Contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States government.

ABBREVIATIONS

GUSAR, general unrestricted structure−activity relationships; PASS, prediction of activity spectra for substances; SCR, self-consistent regression; QSAR, quantitative structure−activity relationships; MNA, multi-level neighborhoods of atoms descriptors; QNA, quantitative neighborhoods of atoms descriptors

REFERENCES

(1) Wenlock, M. C.; Barton, P. In silico physicochemical parameter predictions. *Mol. Pharmacol.* **2013**, *10*, 1224−1235.

(2) Kauffman, G. W.; Jurs, P. C. Prediction of surface tension, viscosity, and thermal conductivity for common organic solvents using quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 408−418.

(3) Akamatsu, M. Importance of physicochemical properties for the design of new pesticides. *J. Agric. Food Chem.* **2011**, *59*, 2909−2917.

(4) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative structure−activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* **2009**, *22*, 1913−1921.

(5) Zakharov, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Quantitative prediction of antitarget interaction profiles for chemical compounds. *Chem. Res. Toxicol.* **2012**, *25*, 2378−2385.

(6) Mwense, M.; Wang, X. Z.; Buontempo, F. V.; Horan, N.; Young, A.; Osborn, D. Prediction of noninteractive mixture toxicity of organic compounds based on a fuzzy set method. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1763−1773.

(7) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.

(8) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Neural Networks in Building QSAR Models. In *Artificial Neural Networks: Methods and Applications*; Livingstone, D. J., Ed.; Methods in Molecular Biology Series, Vol. 458; Springer: Clifton NJ, 2008; pp 137−158.

(9) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

(10) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(11) Livingstone, D. J.; Manallack, D. T.; Tetko, I. V. Data modelling with neural networks: Advantages and limitations. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 135−142.

(12) Sarimveis, H.; Alexandridis, A.; Tsekouras, G.; Bafas, G. A fast and efficient algorithm for training radial basis function neural networks based on a fuzzy partition of the input space. *Ind. Eng. Chem. Res.* **2002**, *41*, 751−759.

(13) Filimonov, D. A.; Akimov, D. V.; Poroikov, V. V. The method of self-consistent regression for the quantitative analysis of relationships between structure and properties of chemicals. *Pharm. Chem. J.* **2004**, *38*, 21−24.

(14) Filimonov, D. A.; Zakharov, A. V.; Lagunin, A. A.; Poroikov, V. V. QNA-based "Star Track" QSAR approach. *SAR QSAR Environ. Res.* **2009**, *20*, 679−709.

(15) U.S. EPA T.E.S.T. Program. http://www.epa.gov/nrmrl/std/qsar/qsar.html (accessed June 28, 2013).

(16) Zakharov, A. V.; Peach, M. L.; Sitzmann, M.; Filippov, I. V.; McCartney, H. J.; Smith, L. H.; Pugliese, A.; Nicklaus, M. C. Computational tools and resources for metabolism-related property predictions. 2. Application to prediction of half-life time in human liver microsomes. *Future Med. Chem.* **2012**, *4*, 1933−1944.

(17) Lagunin, A. A.; Zakharov, A. V.; Filimonov, D. A.; Poroikov, V. V. A new approach to QSAR modelling of acute toxicity. *SAR QSAR Environ. Res.* **2007**, *18*, 285−298.

(18) Lagunin, A.; Zakharov, A.; Filimonov, D.; Poroikov, V. QSAR modelling of rat acute toxicity on the basis of PASS prediction. *Mol. Inf.* **2011**, *30*, 241−250.

(19) Kokurkina, G. V.; Dutov, M. D.; Shevelev, S. A.; Popkov, S. V.; Zakharov, A. V.; Poroikov, V. V. Synthesis, antifungal activity and QSAR study of 2-arylhydroxynitroindoles. *Eur. J. Med. Chem.* **2011**, *46*, 4374−4382.

(20) TEST User's Guide 4.1., U.S. EPA. http://www.epa.gov/nrmrl/std/qsar/TEST-user-guide-v41.pdf (accessed November 14, 2013).