Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression

Minghu Song,[†] Curt M. Breneman,*[,†] Jinbo Bi,[‡] N. Sukumar,[†] Kristin P. Bennett,[‡] Steven Cramer,[§] and Nihal Tugcu[§]

Departments of Chemistry, Mathematics, and Chemical Engineering, Rensselaer Polytechnic Institute, 110 8th Street, Troy, New York 12180

Received August 14, 2002

Quantitative Structure-Retention Relationship (QSRR) models are developed for the prediction of protein retention times in anion-exchange chromatography systems. Topological, subdivided surface area, and TAE (Transferable Atom Equivalent) electron-density-based descriptors are computed directly for a set of proteins using molecular connectivity patterns and crystal structure geometries. A novel algorithm based on Support Vector Machine (SVM) regression has been employed to obtain predictive QSRR models using a two-step computational strategy. In the first step, a sparse linear SVM was utilized as a feature selection procedure to remove irrelevant or redundant information. Subsequently, the selected features were used to produce an ensemble of nonlinear SVM regression models that were combined using bootstrap aggregation (bagging) techniques, where various combinations of training and validation data sets were selected from the pool of available data. A visualization scheme (star plots) was used to display the relative importance of each selected descriptor in the final set of "bagged" models. Once these predictive models have been validated, they can be used as an automated prediction tool for virtual high-throughput screening (VHTS).

I. INTRODUCTION

*I*on-*E*xchange *C*hromatography (IEC) is a widely accepted standard bioseparation technique that has been growing in importance during the past decade in keeping with current rapid developments in biotechnology. To date, there are two main kinds of IEC: cation-exchange and anion-exchange chromatography, determined by whether a negative charge (cation-exchange) or a positive charge (anion-exchange) is carried by the functional groups on the surface of the IEC stationary phase. The ionic biopolymers, such as proteins, are separated primarily through the electrostatics interactions between the charged surface of the ion-exchange resin and the ionic solutes bearing the opposite charge. In the case of anion-exchange chromatography, negatively charged proteins bind in a transient fashion to the positively charged stationary phase sites, as long as the salt concentration is kept low. Proteins bound with different degrees of interaction can be separated with the aid of an increasing salt gradient. The selectivity of this technique can be optimized by varying the composition of the stationary phase as well as the pH of the mobile phase. Consequently, one of the major challenges in ion exchange bioseparation is to select appropriate chromatographic materials for a given biological mixture. It has been suggested that virtual screening of separation materials in a manner that parallels current QSAR (*Q*uantitative *S*tructure-*A*ctivity *R*elationship) methods in drug design would facilitate the selection of proper chromatographic conditions and speed up development processes.

As a result, there is increasing interest within the chromatography community in the development of *Q*uantitative *S*tructure-*R*etention *R*elationship (QSRR) models[1] based on linear or nonlinear modeling techniques, including *P*rincipal *C*omponent *R*egression (PCR),[2] *P*artial *L*east *S*quares (PLS),[3] and *A*rtificial *N*eural *N*etworks (ANN).[4,5] The major aims of these studies are to construct improved QSRR models to predict the retention behavior of solutes in different stationary phases or salt conditions as well as to build a valuable chromatographic interpretation tool for the solute retention mechanisms. Due to computational bottlenecks in descriptor generation and machine learning algorithms, most current approaches are only applicable for small molecules. Recent research has focused on the adaptation of the TAE (*T*ransferable *A*tom *E*quivalent) electron density-derived descriptor technique to large molecules such as proteins.[6] In that study, partial least-squares models constructed using subsets of TAE descriptors were found to be capable of predicting protein retention with good accuracy. In the current study, we present a novel modeling approach based on *S*upport *V*ector *M*achine (SVM) Regression[7] to predict the retention time of proteins in anion exchange systems. A visualization tool, the star plot, is employed to aid in model interpretation. The predictive power of the resulting models is demonstrated by testing them on unseen data that were not used during either descriptor selection or model generation.

II. DATA SET AND DESCRIPTOR GENERATION

Protein Retention Data Set. The crystal structures of 24 structurally diverse proteins with similar isoelectric points (PI) were downloaded from the RSCB Protein Data Bank[8] for analysis. The retention times for these proteins were obtained by carrying out linear gradient chromatography using the anion exchange stationary phase Source 15Q. The

* Corresponding author phone: (518)276-2678; fax: (518)276-4887, e-mail: brenec@rpi.edu.
[†] Department of Chemistry.
[‡] Department of Mathematics.
[§] Department of Chemical Engineering.

10.1021/ci025580t CCC: $22.00
Published on Web 10/15/2002

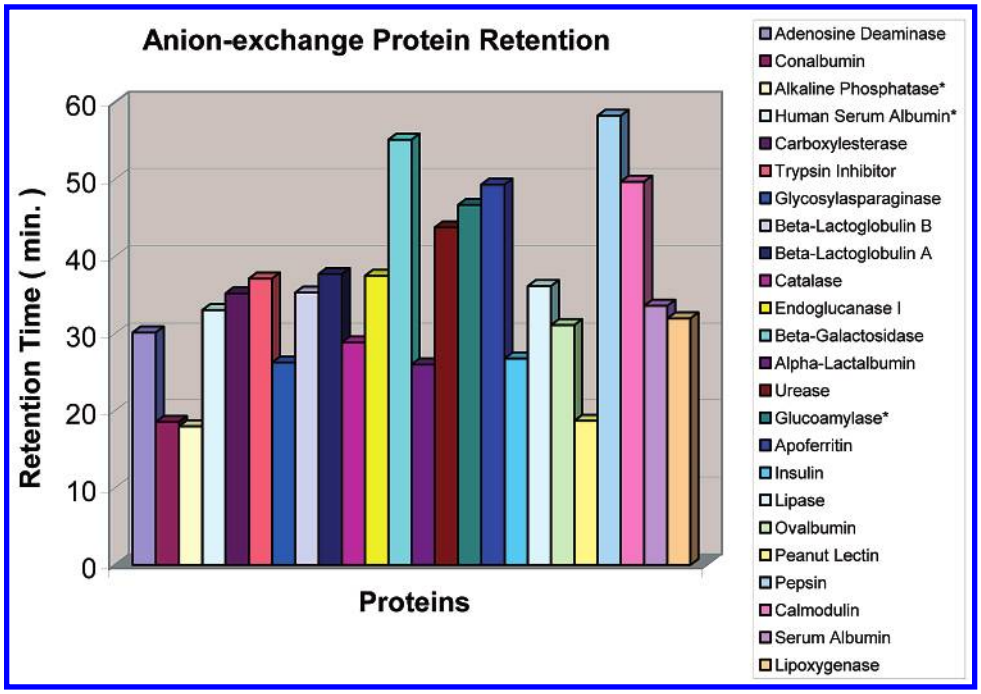

Figure 1. Proteins and their experimental retention times. Entries marked by * were used as the test set.

names and the experimental retention times of the 24 proteins are provided in Figure 1. Three proteins were randomly selected as external test cases from this original list.

SYBYL v6.5 software[9] was used to preprocess the raw macromolecular structures by eliminating the waters of crystallization and adding hydrogen atoms to satisfy neutral valences on all atoms. A total of 243 descriptors was then computed for these proteins using both RECON[10] and MOE (Chemical Computing Group, Inc.) programs to give a composite set of traditional and electron density-derived TAE descriptors.

***Quantum Theory of Atoms in Molecules* (QT-AIM) and TAE/RECON Descriptors.** Quantum chemical descriptors offer an attractive alternative to traditional QSAR/QSPR molecular descriptors by expressing a more accurate and detailed description of the electronic and geometric molecular properties and the interaction between them.[11] However, even with the rapid advances in computer architecture and the anticipated continued growth in computational power, a direct calculation of the properties of large molecules at a high level of theory is prohibitive. Bader's quantum theory of Atoms in Molecules (AIM)[12,13] provides the framework for reconstructing large complicated molecules from a number of small electron density fragments while still achieving an good approximation to the properties of the intact molecules. In AIM theory, the electron density of a molecule can be partitioned into distinct electron density basins (the regions of space occupied by the corresponding atoms), each containing an atomic nucleus. These electron density fragments are essentially bounded by surfaces of zero net flux in the electron density, which correspond to the steepest descent pathways from each bond critical point. An atomic property (A) can then be expressed as the integral of a corresponding property density $\rho_A(r)$ over an atomic basin:

$$A(\Omega) = \int_\Omega d\tau \rho_A(r) \text{ where } \rho_A(r) = (N/2) \int d\tau' \{\psi^* \hat{A}\psi + (\hat{A}\psi)^* \psi\} \quad (1)$$

Table 1. TAE Atomic Electronic Surface Properties

EP	electrostatic potential
Del(Rho)·N	electron density gradient normal to 0.002 e/au^3 electron density isosurface
G	electronic kinetic energy density G = $(-(\hbar/4m) \int \{\nabla\psi^* \cdot \nabla\psi\}\, d\tau)$
K	electronic kinetic energy density $K = -(\hbar/4m) \int \{\psi^* \nabla^2 \psi + \psi \nabla^2 \psi^*\}\, d\tau)$
Del(K)·N	gradient of K electronic kinetic energy density normal to surface
Del(G)·N	gradient of B electronic kinetic energy density normal to surface
Fuk	Fukui F^+ function scalar value
Lapl	Laplacian of the electron density $\nabla^2\rho$
BNP	bare nuclear potential $BNP_{(j)} = \sum_{i=1}^{n} q_i/r_{ij}$
PIP	local average ionization potential $PIP(r) = \sum_i \rho_i(r)\lvert\epsilon_i\rvert/\rho(r)$

These atomic properties possess a high degree of transferability from the electronic environment in one molecule to another molecule with a similar environment. Consequently, the properties of a functional group or whole molecule can be obtained by adding these atomic properties together:

$$A_{molecule} = \sum_{\Omega} A(\Omega) \quad (2)$$

Based on AIM theory, Breneman introduced the concept of "Transferable Atom Equivalents" (TAEs),[10,14] which are composed of atomic electron density fragments bounded by interatomic zero-flux surfaces ($\nabla\rho(r) \cdot n(r) = 0$, for all points on the surface) and an extended $\rho = 0.002$electron/au^3 isodensity surface that approximates the condensed-phase van der Waals surface. TAE fragments carry 10 atomic charge density-derived properties (listed in Table 1) that were precomputed from small molecules using ab initio wave functions at the 6-31+G* level of theory. As evident from the table, TAE electron density reconstructions provide not only molecular electron densities but also electronic kinetic energy densities and local average ionization potentials as well as other first- and second-derivative properties of

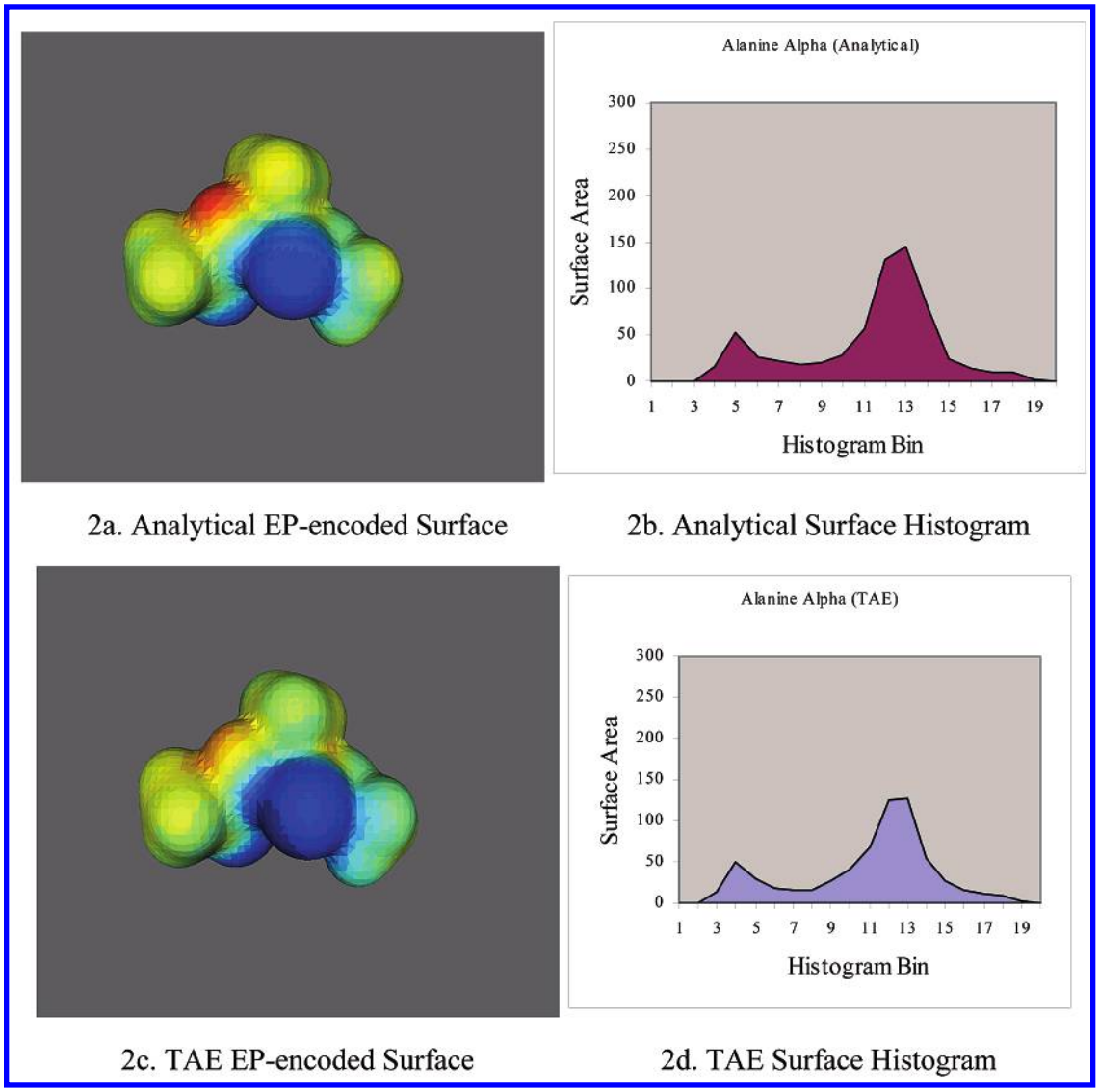

Figure 2. Electrostatic potential surface distributions and histograms generated for alanine using both TAE and analytical ab initio methods. The color scheme in parts a and c corresponds to different values of EP on the molecular surface. The distributions of the surface electrostatic potentials are characterized as histogram descriptors using binning techniques as illustrated in parts b and d. Descriptors for larger molecules or proteins can be computed following a similar scheme. TAE reconstruction of the proteins used in this study required approximately 60 s on a single 1.7 GHz processor Linux PC.

thedensity. The distributions of these electronic properties computed on 0.002-e/au^3 electronic density isosurfaces may be characterized as molecular property descriptors in several ways. TAE histogram descriptors can be produced by recording the distribution of the properties as surface histograms that quantified the molecular surface areas with specific ranges of each property value. In addition to these histogram descriptors, property extrema, average values and standard deviations of the property distributions (in some cases with separate σ values for positive and negative portions of the range) were also included in the TAE descriptor set.

The TAE library consists of a set of precalculated atomic fragments structured in a form that allows the atomic fragments involved in the new molecule to be rapidly retrieved. The RECON (*RECON*struction) program reads the atomic connectivity information of the protein and assigns the closest fragment match from the TAE library to each atom based on atom type, hybridization and structural environment. By summing up the corresponding atomic properties of the constituent fragments, we can obtain a large set of electron density-based TAE descriptors for macromolecules. These descriptors provide information about basicity, hydrophobicity, hydrogen-bonding capacity and polarity as well as molecular polarizability. For example, surface property histograms such as the electrostatic potential distribution of alanine histogram shown in Figure 2 may be computed using TAE/RECON program. As shown in the figure, the TAE electrostatic potential distribution represents the analytical ab initio result quite effectively.

The TAE/RECON approach has been shown to be effective in QSPR studies.[15] It is a resource-efficient alternative to HF/SCF or DFT ab initio calculations, which can be prohibitive even for molecules of modest size. The CPU and disk resources required for TAE reconstruction are comparable to those utilized by molecular mechanics energy computations. The TAE QSPR descriptors for individual proteins or large databases can be computed within seconds on modest workstations.

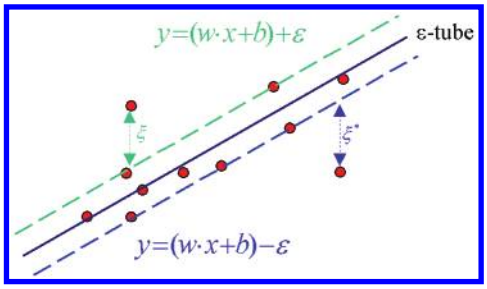

Figure 3. Graphical depiction of an ϵ-insensitive loss function and an ϵ-tube. Only the deviations of data points outside the ϵ-tube, such as ξ and ξ^* , will be considered as the errors and thus be penalized—in this case two points are used as examples.

MOE Descriptors. The MOE program provides a widely applicable set of classical molecular descriptors, including traditional physicochemical properties, connectivity-based topological 2D and shape-dependent 3D molecular features. These descriptors have been applied to the construction of QSAR/QSPR models for boiling point, vapor pressure, and the free energy of solvation in water as well as water solubility and blood-brain barrier penetration.[16]

III. MODELING METHODOLOGY

***Support Vector Regression* (SVR) Overview.** In recent years, there has been a lot of interest in studying support vector machines (SVMs) in the field of machine learning. SVMs are a class of supervised learning algorithms initially proposed by Vapnik.[17,18] To date, SVMs have been applied successfully to a wide range of pattern recognition problems, such as image recognition,[19] microarray gene expression classification,[20] protein folding recognition,[21] protein structural class prediction,[22] identification of protein cleavage sites,[23] QSAR and other pharmaceutical data analysis.[20,24] Although SVMs were originally developed for classification, Vapnik enabled them to solve regression problems by choosing a suitable cost function (ϵ-insensitive loss function) that enables a sparse set of support vectors to be obtained.[17]

Normal regression procedures are often stated as the processes deriving a function f(x) that has the least deviation between predicted and experimentally observed responses for all training examples. One of the main characteristics of SVR is that instead of minimizing the observed training error, SVR attempts to minimize the generalization error bound so as to achieve higher generalization performance. This generalization error bound is the combination of the training error and a regularization term that controls the complexity of hypothesis space. The first term is calculated by the ϵ-insensitive losses[17]

$$L_\epsilon(y - f(x)) := |y - f(x)|_\epsilon = \min(0, |y - f(x)| - \epsilon) \quad (3)$$

in which ϵ is the tolerance to error and we only consider those deviations larger than ϵ as errors. The l_2-norm $1/2||\omega||^2$ of normal vector is typically adopted as a regularization factor and ω is the weight vector to be determined in the function f. This algorithm, called ϵ-SVR, seeks to find a function $f^* \in F = \{f: R^N \rightarrow R\}$ based on a training set of M examples (x_i, y_i) with $x_i \in R^N$ by minimizing the overall regularized risk functional[25]

$$\frac{1}{2}||\omega||^2 + C\sum_{i=1}^{M}|y_i - f(x_i)|_\epsilon \quad (4)$$

where C is a fixed regularization constant determining the tradeoff between training error (empirical loss) and model complexity. Figure 3 illustrates what the ϵ-insensitive loss function looks like.

If the hypothesis space F consists of a linear function in the form $\langle w \cdot x \rangle + b$, then the SVR problem can be posed as a convex optimization problem as follows:

$$\left[\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*) \\ \text{subject to} & y_i - \langle w \cdot x_i \rangle - b \le \epsilon + \xi_i,\ \xi_i \ge 0, \\ & \langle w \cdot x_i \rangle + b - y_i \le \epsilon + \xi_i^*,\ \xi_i^* \ge 0, \\ & i = 1,2,\ldots,M \end{array}\right] \quad (5)$$

A favorable property of the above formulation is that its solution is robust with respect to small changes in the training set.

Another major characteristic of Support Vector methods is that it implicitly maps the original input space to a high dimensional feature space $x \mapsto \Phi(x)$ by means of so-called kernel functions based on Mercer's theorem, whereupon a linear regression function $f(x) = \langle w \cdot \Phi(x) \rangle + b$ is constructed upon the feature space to achieve a nonlinear model in the original input space. Thus Support Vector generalization error, unlike those of other machine learning methods, is not directly related to the original input dimensionality of the problem. By the optimality conditions of the quadratic programming formulation of SVMs, the normal vector w can be expressed as $w = \sum_{i=1}^{M} \alpha_i \Phi(x_i)$ and the function f can be written in the form of a kernel expansion as

$$f(x) = \sum_{i=1}^{M} \alpha_i k(x_i, x) + b \text{ where } k(x_i, x) = \langle \Phi(x_i) \cdot \Phi(x) \rangle \quad (6)$$

In classical support vector regression, the proper value for the parameter ϵ is difficult to determine beforehand. Fortunately, this problem is partially resolved in a new algorithm, ν support vector regression (ν-SVR),[26,27] in which ϵ itself is a variable in the optimization process and is controlled by another new parameter $\nu \in (0,1]$. ν is the upper bound on the fraction of error points or the lower bound on the fraction of points inside the ϵ-insensitive tube. Thus a good ϵ can be automatically found by choosing ν, which adjusts the accuracy level to the data at hand. This makes ν a more convenient parameter than the one used in ϵ-SVR.

Since solving quadratic programming problems is usually more computationally expensive than solving linear programming problems, efforts have been made to derive a linear programming formulation for SVR. Instead of using the Euclidean norm i.e., l_2-norm regularization of w, the sparse ν-SVR always regularizes through applying l_1-norm, a sparse favoring norm, directly to coefficients $\alpha_j, j = 1, \cdots M$ in the kernel expansion of f. The l_1-norm of the vector α is $\sum_{j=1}^{M}|\alpha_j|$, which can be rewritten as $\sum_{j=1}^{M}(\alpha_j + \alpha_j^*)$ if we define $\alpha_j = \alpha_j - \alpha_j^*$, where $\alpha_j \ge 0$ and $\alpha_j^* \ge 0$.

Due to these features of linear ν support vector regression, we adopted it for our numerical experiments on the QSRR problem. A two-step computational strategy was adopted: First, a sparse linear SVM was utilized as a variable selection method to identify relevant molecular descriptors; and then in the next step, a set of nonlinear SVM models derived by kernel mapping were constructed using the selected features. In addition, a statistical technique called "bagging" (Bootstrap Aggregation) was employed to improve model generalization performance.

***l_1*-Norm SVR Linear Feature Selection.** In ion-exchange chromatography systems, the solutes interact with the stationary phase in the column through a combination of intermolecular interactions as the mobile phase flows down through the column. Since it is not possible to know a priori which molecular descriptors are most relevant for describing these interactions, a comprehensive set of descriptors is employed in the initial steps of QSRR model generation. This results in a situation where there are far fewer observations than the number of molecular descriptors. As is well-known in both the chemical and statistical communities, the accuracy of prediction is not monotonic with respect to the number of features employed in the model, because some descriptors may be found to be unnecessary or irrelevant, while inclusion of too many descriptors may produce fortuitous correlations and over-trained models. Therefore, in this extreme of very few observations with very many descriptors, it is essential to utilize efficient feature selection and regularization methods. Even though SVMs are claimed to be insensitive to the problem of dimensionality with kernels implemented as discussed above, reduction of the input space can still help to speed up the learning process by removing irrelevant features and emphasizing only a few relevant features to make the interpretation more convenient. That is why feature selection methods have received much attention recently in QSAR or QSPR studies. Several algorithms, such as forward selection,[28] simulated annealing,[29] genetic algorithms,[30,31] K-nearest neighbor,[32] evolutionary programming,[33,34] artificial ants[35,36] and binary particle swarms,[37] have been implemented for feature selection in the scientific literature.

The feature selection method used in this work exploits the fact that sparse SVM modeling using a linear hypotheses with l_1-norm regularization inherently performs feature selection as a side effect of minimizing function capacity during the modeling process.[38] In a linear regression model of the form $y = \langle a \cdot x \rangle + b$, each component of α provides a weight for the corresponding feature, thus providing a measure of its significance in the model. Moreover, the sign of each component α_i indicates the effect of the i^{th} feature on the hypothesis. If $\alpha_i > 0$, the feature contributes positively to the observed response y, and when negative it diminishes y. In linear Support Vector regression, the pertinent process involves maximization of the "margin", a term that is inversely proportional to the norm of the weights $||w||$. The margin is defined as the geometric size of the ϵ-tube. In the case of linear SVMs, this size of the margin provides a measure of model complexity. An effect of maximizing the margin (or minimizing the norm of the weights) is to make the optimal weight vector more sparse. Sparsity is defined here as the average number of nonzero components (descriptor weights) in the optimal weight vector. This method of feature selection is formulated as a sparse ν-SVR without kernel mapping, which can be stated in the following manner:

Figure 4. The weights of irrelevant descriptors will converge to zero much faster when using the l_1-norm compared to the l_2-norm.

$$\left[\begin{array}{ll} \text{minimize} & \frac{1}{2}\sum_{j=1}^{N}(\alpha_j + \alpha_j^*) + C\frac{1}{M}\sum_{i=1}^{M}(\xi_i + \xi_i^*) + C\nu\epsilon \\ \text{subject to} & y_i - \sum_{j=1}^{N}(\alpha_j - \alpha_j^*)x_{ij} - b \le \epsilon + \xi_i,\ i = 1,2,...,M \\ & \sum_{j=1}^{N}(\alpha_j - \alpha_j^*)x_{ij} + b - y_i \le \epsilon + \xi_i,\ i = 1,2,...,M \\ & \alpha_j, \alpha_j^*, \xi_i, \xi_i, \epsilon \ge 0,\ j = 1,2,...,N,\ i = 1,2,...,M \end{array}\right] \quad (7)$$

One-norm sparse SVR optimization can enhance the sparsity of the l_1-norm of α as shown in Figure 4, because it is easier to drive the weights of irrelevant descriptors to zero. Those descriptors with nonzero weights then become potentially relevant features to be selected and used to build a subsequent nonlinear model.

Since QSRR data are sometimes comprised of relatively few examples represented by many correlated descriptors, even small perturbations of the training set may lead to large variations in the learning process. This eventuality results in the generation of different linear models and different sets of nonzero-weighted descriptors for related training sets. Recent research reported in the literature has shown that if used with care, ensemble modeling can improve the generalization performance particularly for unstable nonlinear models, such as those involving neural networks.[39] Thus to stabilize the learning process and ensure that a robust set of features are selected in the present work, the technique of bootstrap aggregation (or "bagging") was used in the form originally proposed by Breiman.[40,41] The idea is to construct a series of individual sparse SVR predictors (models) using a bootstrap resampling technique,[42] record the selected descriptors for each individual bootstrap and then take a union of all descriptors into a single final feature set.

The overall feature selection scheme is illustrated in Figure 5. The following process was carried out in this work:

- Multiple training and validation sets were developed from a master training data set using a bootstrapping protocol;
- A series of sparse linear SVMs was created that exhibit good generalization following the common accession using n-fold cross-validation and quantified by cross-validated correlation coefficients;
- Subsets of features having nonzero weights in the linear models were selected;
- Finally, all features obtained in the last step were aggregated to produce the final candidate set of descriptors.

Nonlinear Regression Bagging Models. Once a set of features is selected, a nonlinear ν-SVR with a kernel

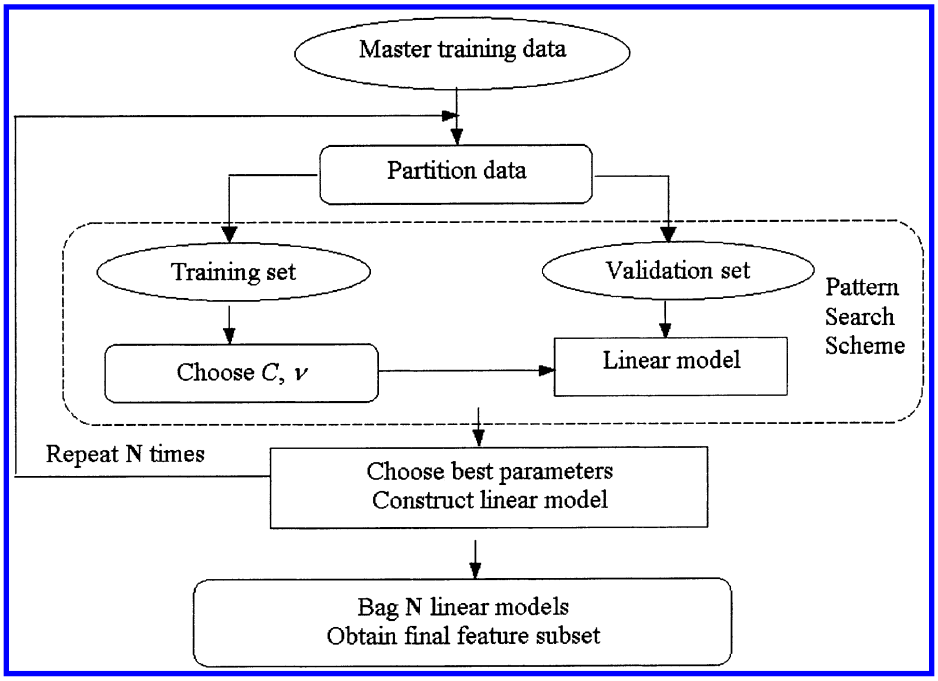

Figure 5. General framework of feature selection scheme.

formulation such as shown in eq 9 is used to construct the QSRR models. The Radial Basis Function (RBF)

$$k(x,x') = \exp(-||x - x'||^2/2\sigma^2) \tag{8}$$

was chosen as the kernel in our computational studies.

This allows us to obtain the regression function f as a linear combination of only a few kernel functions. The sparse ν-SVR is formulated as follows:

$$\left[\begin{array}{ll} \text{minimize} & \frac{1}{2}\sum_{j=1}^{M}(\alpha_j + \alpha_j^*) + C\frac{1}{M}\sum_{i=1}^{M}(\xi_i + \xi_i^*) + C\nu\epsilon \\ \text{subject to} & y_i - \sum_{j=1}^{M}(\alpha_j - \alpha_j^*)k(x_i,x_j) - b \le \epsilon + \xi_i,\ i = 1,2,...,M \\ & \sum_{j=1}^{M}(\alpha_j - \alpha_j^*)k(x_i,x_j) + b - y_i \le \epsilon + \xi_i,\ i = 1,2,...,M \\ & \alpha_j, \alpha_j^*, \xi_i, \xi_i, \epsilon \ge 0,\ i,j = 1,2,...,M \end{array}\right] \tag{9}$$

A simple grid search[43] was employed to choose appropriate values for the kernel parameter σ as well as the capacity factor C and the parameter ν. More details of how the parameters C,ν are selected using a pattern search technique can be found in Bennett's recent publication.[38] To again reduce the variance of the predicted values, the same "bagging" technique was utilized in training the final regression model over the selected features based on the nonlinear SVR predictors$\phi_n(x)$.

$$\phi_{bag}(x) = \frac{1}{N}\sum_{n=1}^{N}\phi_n(x), \quad \text{where } N \text{ is the cardinality of the ensemble} \tag{10}$$

The same cross-validation procedure as described earlier was used to quantify the predictive capabilities of individual predictors and that of the whole predictor ensemble.

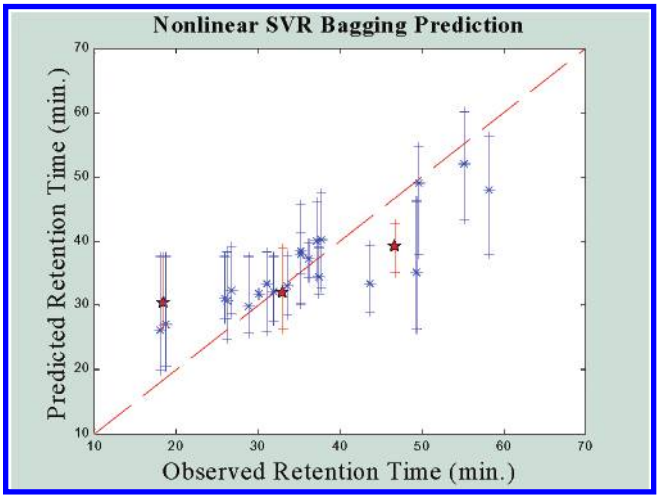

Figure 6. The prediction scatter plot using all descriptors before feature selection.

Implementation. The SVR feature selection and modeling program was implemented using the CPLEX optimization toolbox[44] and the C programming language as available in the Department of Mathematics at RPI and installed on an IBM-AIX Unix platform. Star plot visualization graphics were generated using the S-PLUS 2000 software package.[45]

IV. RESULTS AND DISCUSSION

SVR Feature Selection and Bagging Prediction Results. The aim of this work was to generate predictive models for protein ion-exchange chromatographic retention times with high accuracy as well as to characterize the main interaction mechanisms that account for the retention behavior in anion exchange systems.

Figure 6 shows retention time modeling results obtained before any feature selection using all topological and quantum mechanical descriptors mentioned in Section II. In this figure, the observed retention times (horizontal axis) are

Table 2. Definition of the Relevant Descriptors Obtained from Sparse SVR Feature Selection

descriptor name	chemical information encoded in these descriptors
PEOE.VSA.FPPOS (MOE)	Fraction of positive polar van der Waals surface area. The Partial Equalization of Orbital Electronegativities (PEOE) method of calculating the atomic charges was developed by Gasteiger[46]
FCHARGE (MOE)	Total charge of molecule (sum of formal charges)
PIP2 (TAE)	The second histogram bin of PIP property. Local average ionization potential in the low range
PIP20 (TAE)	The last histogram bin of PIP property. Local average ionization potential in the high range
SIKIA(TAE)	K electronic kinetic energy density, which correlates with the presence and strength of Bronsted basic sites. (integral average)
SIGIA (TAE)	Derived from the G electronic kinetic energy density on the molecular surface. Similar in interpretation to SIKIA, but provide supplemental information.
VSA.POL	Sum of van der Waals surface of "polar" atoms

plotted against the corresponding predicted values for each protein obtained using nonlinear SVR models. The blind test data, as indicated in red, were held out and were not involved in model generation or validation. The asterisk on each vertical bar shows the bagged result of 12 bootstraps for each protein, and the length of the bar represents the full prediction range of retention time for each protein generated by the 12 bagged models. The cross-validation step produced an $R^2_{CV} = 0.851$ and the blind test set an $R^2_{bag} = 0.926$.

As discussed above, sparse ν-SVR approaches were adapted to select only those features relevant to anion-exchange protein retention under the experimental conditions used to develop the data set. In this feature selection procedure, 20 sparse linear SVM models were constructed based on 20 different random partitions of the training data. In the final aggregate SVR model, there were only seven descriptors remaining with nonzero weights. These seven descriptors and their primary definitions are shown in Table 2. Although some of the descriptors are not directly associated with specific physicochemical effects, they have been found to contain chemical information relevant to the interaction mechanisms involved in the anion exchange system. As explained below, this can facilitate the understanding of QSRR modeling for protein retention.

During the model construction, one of main tasks is to determine the significance of the selected QSRR descriptors for later model interpretation. In earlier work, traditional QSRR equations made up of linear combinations of physically interpretable structural descriptors were employed to elucidate the relative importance of several molecular mechanisms involved in chromatographic processes.[47]

In contrast to earlier techniques that often used descriptor weights within single models for chemical interpretation, a graphic visualization tool known as "star plots"[48] was used in the current work to characterize the relative importance of the seven selected descriptors across the multiple models present in the bootstrap aggregate. In most multivariate visualization applications, star plots are generated in a multiplot format where each plot represents one case, and each radial line represents the magnitude of a particular variable (or column) in the data matrix. When the endpoints of the rays are connected together with a line, the resulting figure resembles a "star". In the current work, each star corresponds to a single selected relevant descriptor, where the radius of each spoke is the weight of that descriptor in one of the sparse SVR models used in the bootstrap (normalized by the maximum magnitude of the weights of all descriptors in the same bootstrap). This technique visually represents the relative importance of each descriptor in each of the predictor models used in the bootstrap aggregate and provides a measure of the consistent importance of the descriptor over all of the bootstrap models. For each descriptor, the sum or average of all 20 radii (or the surface area of the star) can be used to represent the overall relative importance of the descriptor over all 20 bootstraps. The descriptor weights from all 20 of the linear SVR models used in the bootstrap aggregation procedure are mapped onto the star plots in the manner shown in Figure 7. In the example shown in that figure, descriptor 1 is consistently important to all models, while descriptor 2 has less uniform significance.

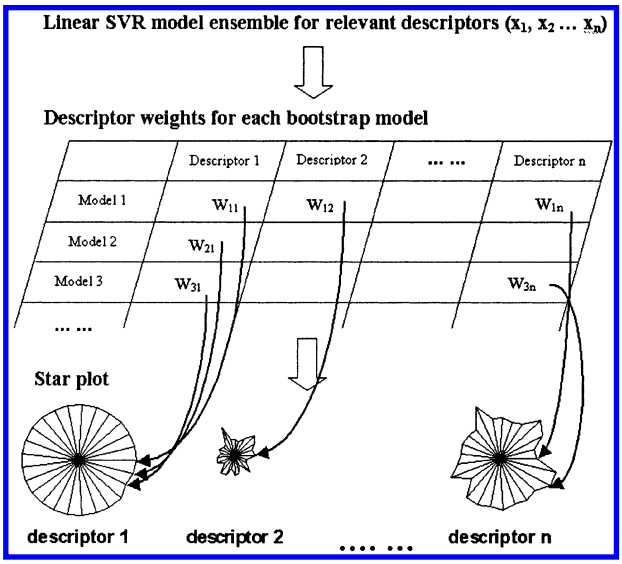

Figure 7. Star plot generation process.

Since descriptor contributions may be either positive or negative, background color is used indicate the consistent sign of the weight across all bootstraps. Finally, the descriptors may be ranked over all 20 bootstrap iterations, such that the most significant negatively weighted descriptor appears in the upper left of the graphic, while the most positively weighed descriptor appears in the lower right-hand side of the figure. The ordering is performed in a column-wise fashion. For instance, in Figure 8, the star plot graph shows seven stars representing the weights of the seven selected descriptors over 20 bootstraps. As shown in the figure, PEOE.VSA.FPPOS has the largest negative effect on retention time and PIP2 has the largest positive effect on retention time. This kind of graphical approach offers a direct way to

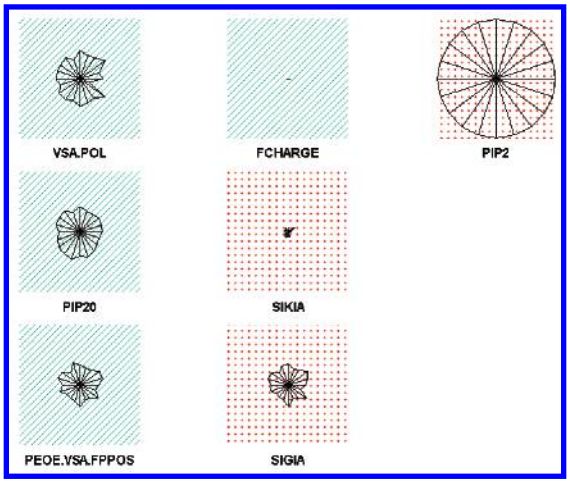

Figure 8. Star plots for the seven descriptors selected by the feature selection algorithm. Descriptor starplots with a cyan background have negative contributions to the retention time, while those with a red dot background have a positive effect on retention.

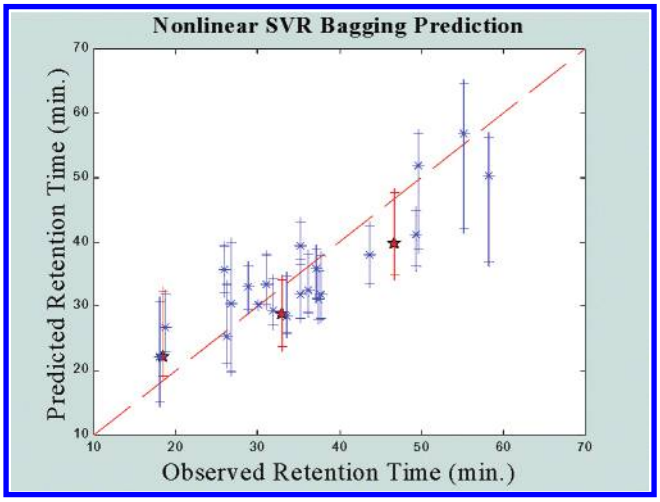

Figure 9. The prediction scatter plot using the nonlinear SVR model with seven selected descriptors.

examine the relative significance of molecular property descriptors in a semiquantitative manner.

The scatter plot for nonlinear SVR prediction based on these seven descriptors with 12 bootstraps is shown in Figure 9. In this case, the cross-validated $R^2_{CV} = 0.882$ and the test set $R^2_{bag} = 0.988$. It may be observed that the final nonlinear model performs better with only seven features than with the original 243 descriptors. The reduction in features also simplifies the model and allows for better interpretation. While the predictive accuracy of the model is subject to improvement, the technique is clearly capable of providing useful estimates of retention time that should prove useful in chromatography planning.

Model Interpretation. Besides the development of direct prediction models, one of main challenges in QSRR lies in extracting chemical meaning from the descriptor patterns found in the models. It is hypothesized that the application of a fundamental physicochemical modeling approach such as that used in this investigation will aid in the understanding of the interaction mechanisms of ion-exchange systems. The development of predictive models and a greater level of understanding of the underlying processes of protein chromatography will be valuable for future experimental design.

Despite its widespread application, the exact mechanism of protein retention in ion-exchange chromatography is still controversial. Protein retention on an anion-exchange resin is controlled by the balance of interactions between the protein and a set of charged functional groups as well as by the characteristics of the surrounding medium and the stationary phase solid matrix. It is known that this kind of mixed-mode separation mechanism within ion-exchange columns can offer unique selectivity for the separation of proteins. One such scenario is depicted in Figure 10.

Due to the nature of the ion-exchange processes, it is expected that electrostatic effects will play a dominant role in protein retention. This dominant electrostatic effect arises because the acidic amino acid side chains, i.e., aspartate and glutamate, are partially deprotonated under the experimental conditions in which the mobile phase is buffered at pH = 7.4 and produce negative charges on the periphery of the protein. Since anion exchange sites (quaternary ammonium functional groups $N(CH_3)_3^+$) on the resin surface are completely ionized under these conditions, proteins with high negative charge densities on their surfaces will show greater affinity for these sites and will elute later. Proteins with low negative charge densities will interact more weakly with the resin and will elute first. As described later, additional effects are also present that can influence elution selectivity, including overall charge asymmetry and other factors.

According to the SVM modeling results from this work, a dominant set of electrostatic interactions may be proposed to explain the protein retention behavior using three main factors: net charge, polarity/polarizability (charge asymmetry) and a desolvation penalty.

The first of the three electrostatic factors is represented by a fractional surface area descriptor and atomic formal charges. The MOE descriptor PEOE.VSA.FPPOS represents the fraction of the molecular surface area bearing a positive partial charge, as calculated by the PEOE (Partial Equalization of Orbital Electronegativities) approach. As shown in Figure 8, this descriptor bears a negative weight, meaning that greater fractional positive surface area decreases the protein retention time. This result is consistent with the net charge hypothesis, which suggests that fixed positively charged sites in the resin will exhibit a favorable affinity for negatively charged amino acids and repel positively charged regions of the protein surface. The descriptor FCHARGE represents the total formal charge of the protein, which is negative in cases where the solution pH is higher than their isoelectric point (PI). The small negative value of its weight in the sparse SVR models is consistent with the explanation that proteins with more negative charge have a tendency to interact more strongly with the positively charged function groups present on the surface of the resin. The apparent insignificance of this seemingly important descriptor is due to the fact that the electrostatic effect is better represented more by other selected electrostatic-related descriptors. The importance of this descriptor may prove to be more significant in data sets involving proteins with diverse PI.

Although the above net charge model has been frequently used to explain the phenomenon, retention mapping studies on the strong ion-exchange columns showed it to be inadequate. The influence of intramolecular charge asymmetry in the proteins has been successfully employed as an

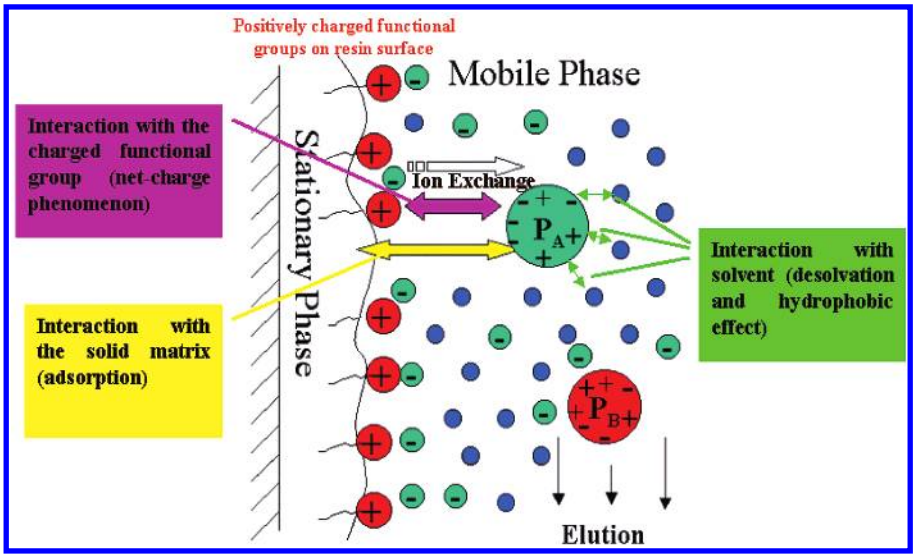

Figure 10. A simple cartoon illustration of multi-mode interaction involved in protein retention. The symbols P_A and P_B represent two proteins with different binding affinities to the stationary phase.

alternative explanation for deviations from the net charge model together with the fact that protein tertiary structure is know to affect retention.[49] Recent studies suggested that protein local dipolarity should also be taken into consideration, since it appears that only a fraction of locally charged protein surfaces interact with the stationary phase.[50] These regions of localized charge are postulated to orient the protein with respect to the oppositely charged ion-exchange support. It is clear that in large, complex macromolecules, the distribution of charged groups may not be uniform throughout the structure. As a result of this inhomogeneity, even proteins with zero net charge may exhibit significant electrostatic fields. Consequently, the retention behavior will depend not only on the net charge itself but also on the spatial distribution of charge throughout the protein structure. Other effects include the potential for reorganizing these dipoles (and higher multipoles) in response to an applied electronic field originating from the neighboring medium. Descriptors associated with dipolarity and polarizability effects are expected to account for differences among the solutes as to their propensity to participate in dipole−dipole, dipole−induced dipole and charge-transfer interactions.

Several TAE electron density-based descriptors listed in Table 2 were found to be significant to retention, e.g. SIKIA, SIGIA and PIP. These descriptors have also been found to correlate with molecular properties such as acid/base strength and polarity as well as polarizability.[10] The PIP descriptor family can be associated with regions of donor and acceptor capabilities that relate to the tendency of analytes to take part in charge-transfer interactions. Prior to feature selection, there were twenty PIP descriptors present in the descriptor set, where PIP1 and PIP2 represent regions of the molecular surface where electron density is easily ionizable, while PIP20 is associated with regions of tightly held electron density, such as on exchangeable protons. SIGIA and SIKIA describe the integrals of G and K electronic kinetic energy densities found on the molecular van der Waals surface. These descriptors are related to the Laplacian of the density and are associated with the presence and the strength of Lewis basic sites. It has been shown in this work that these dipolarity/polarizability-related descriptors (PIP2, SIKIA and SIGIA) correlate with increased retention time. This may be due to their representation of increased dipole/induced-dipole or charge/induced-dipole forces between the protein and the strong ion-exchanger groups as well as induced-dipole/induced-dipole interactions between the polarizable aromatic groups of the stationary phase and polarizable regions of the protein. The PIP20 descriptor was found to be anticorrelated with retention time, indicating that the presence of nonacidic hydrogen bond donors (serine, etc.) increases solute/mobile-phase interactions at the expense of solute/stationary phase interactions.

In addition to the charge characteristics of the protein and resin surface as well as the underlying matrix, the nature of the solvent, e.g. polarity, is also known to be an important contributor to protein retention. In the current model, this effect is described by the MOE descriptor VSA.POL, which approximates the VDW surface area of polar atoms (both hydrogen bond donors and acceptors). The importance of this descriptor implies that the hydrogen bonding capacity of proteins may also be involved in the intermolecular interactions responsible for column retention behavior. To a first approximation, most charged and polar groups on the solvated protein can interact favorably with the surrounding water before attaching to the support surface. Thus, even a protein with a moderate polarity has to pay an energetic penalty which increases in proportion to the overall polar surface of the protein. In this way a protein with more polar atoms on the exposed van der Waals surface will have a stronger hydrogen-bonding capacity with the mobile phase and will elute out of the column first, accounting for the negative effect of VSA.POL for retention shown in the star plot. Fortunately, this kind of desolvation penalty can be offset, although never completely overcome, by more favorable electrostatic interactions between the resin and the protein that lead to increased protein retention.

V. CONCLUSIONS

In this study, Support Vector Machine (SVM) regression was introduced as a method for generating predictive QSRR models of protein retention times in anion exchange chro-

matographic systems. It was demonstrated that models developed using this technique encompass a wide range of proteins including a variety of sizes, shapes, functionalities and selectivity for these resins. In the future, this method should prove useful for performing comparative QSRR studies under different chromatographic conditions and be important for determining appropriate protein purification conditions.

In summary, the behavior of protein solutes in anion-exchange chromatography conditions may be quantified through the use of traditional and electron density-based molecular property descriptors. Models may be constructed using SVR methods for both feature selection and property prediction. Extensive cross-validation of the modeling results was accomplished using multiple sets of training and validation cases in a bootstrap scheme, the results of which are visualized and interpreted using star plots.

ACKNOWLEDGMENT

This work was supported in part by NSF grants IIS-9979860 and BES-0079436.

REFERENCES AND NOTES

(1) Kaliszan, R. Correlation between Retention Indexes and Connectivity Indexes of Alcohols and Methyl-Esters with Complex Cyclic Structure. *Chromatographia* **1977**, *10*, 529−531.
(2) Katritzky, A. R.; Petrukhin, R.; Tatham, D.; Basak, S.; Benfenati, E. et al. Interpretation of quantitative structure−property and -activity relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 679−685.
(3) Montana, M. P.; Pappano, N. B.; Debattista, N. B.; Raba, J.; Luco, J. M. High-performance liquid chromatography of chalcones: Quantitative structure-retention relationships using partial least-squares (PLS) modeling. *Chromatographia* **2000**, *51*, 727−735.
(4) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of gas chromatographic retention indices of alkylbenzenes. *Anal. Chim. Acta* **1997**, *342*, 113−122.
(5) Loukas, Y. L. Artificial neural networks in liquid chromatography: efficient and improved quantitative structure-retention relationship models. *J. Chromatogr. A* **2000**, *904*, 119−129.
(6) Mazza, C. B.; Sukumar, N.; Breneman, C. M.; Cramer, S. M. Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Anal. Chem.* **2001**, *73*, 5457−5461.
(7) Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Networks* **1999**, *10*, 988−999.
(8) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(9) *SYBYL 6.5*, Tripos Assoiciate Inc. 1699 S. Hanley Rd., Suite 303, St. Louis, MO 63144-2913.
(10) Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. Electron-Density Modeling of Large Systems Using the Transferable Atom Equivalent Method. *Comput. Chem.* **1995**, *19*, 161−179.
(11) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027−1043.
(12) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford University Press: Oxford, UK, 1994.
(13) Matta, C. F. Theoretical reconstruction of the electron density of large molecules from fragments determined as proper open quantum systems: The properties of the oripavine PEO, enkephalins, and morphine. *J. Phys. Chem. A* **2001**, *105*, 11088−11101.
(14) Breneman, C. M. Transferable Atom Equivalents. Molecular Electrostatic Potentials from the Electric Multipoles of PROAIMS Atomic Basins. *The Application of Charge Density Research to Chemistry and Drug Design*; Plenum Press: 1991; pp 357−358.
(15) Breneman, C. M.; Rhem, M. QSPR analysis of HPLC column capacity factors for a set of high-energy materials using electronic van der Waals surface property descriptors computed by transferable atom equivalent method. *J. Comput. Chem.* **1997**, *18*, 182−197.
(16) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.
(17) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer, Berlin, 1995.
(18) Cortes, C.; Vapnik, V. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273−297.
(19) Zhang, L.; Zhou, W. D.; Jiao, L. C. Support vector machine for 1-D image recognition. *J. Infrared Millimeter Waves* **2002**, *21*, 119−123.
(20) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(21) Ding, C. H. Q.; Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **2001**, *17*, 349−358.
(22) Karchin, R.; Karplus, K.; Haussler, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **2002**, *18*, 147−159.
(23) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support vector machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.* **2002**, *23*, 267−274.
(24) Czerminski, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **2001**, *20*, 227−240.
(25) Vapnik, V. N. *Estimation of Dependences Based on Emperical Data*; Springer-Verlag: Berlin, 1982.
(26) Smolar, A. J.; Scholkopf, B. Linear Programs for Automatic Accuracy Control in Regression. *Proceedings ICANN'99, Int. Conf. on Artificial Neural Networks*; Springer: Berlin, 1999.
(27) Scholkopf, B.; Smola, A. J.; Williamson, R. C.; Bartlett, P. L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207−1245.
(28) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised forward selection: A method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160−1168.
(29) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity-Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(30) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity-Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(31) So, S. S.; Karplus, M. Genetic neural networks for quantitative structure−activity relationships: Improvements and application of benzodiazepine affinity for benzodiazepine/GABA(A) receptors. *J. Med. Chem.* **1996**, *39*, 5246−5256.
(32) Zheng, W. F.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(33) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity-Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.
(34) Kubinyi, H. Variable Selection in Qsar Studies .1. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.
(35) Izrailev, S.; Agrafiotis, D. A novel method for building regression tree models for QSAR based on artificial ant colony systems. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 176−180.
(36) Izrailev, S.; Agrafiotis, D. K. Variable selection for QSAR by artificial ant colony systems. *SAR QSAR Environ. Res.* **2002**, *13*, 417−423.
(37) Agrafiotis, D. K.; Cedeno, W. Feature selection for structure−activity correlation using binary particle swarms. *J. Med. Chem.* **2002**, *45*, 1098−1107.
(38) Bennett, K. P.; Bi, J.; Embrechts M.; Breneman, C.; Song, M. Dimensionality Reduction via Sparse Support Vector Machines. *J. Machine Learning Research 2002 (Special Issue on Feature Selection)* (In press).
(39) Dimitris K. Agrafiotis, W. C.; Victor S. On the use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.
(40) Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24*, 123−140.
(41) Breiman, L. Using iterated bagging to debias regressions. *Machine Learning* **2001**, *45*, 261−277.
(42) Efron, B.; Tibshirani, R. J. *An introduction to the bootstrap*; Chapman and Hall: New York, 1993.
(43) Demiriz A. B. K.; Breneman C.; Embrechts, M. Support Vector Machine Regression in Chemometrics. *33rd Symposium on Computing Science and Statistics: Proceedings of Interface*, June, 2001.
(44) *Using the CPLEX(TM) Linear Optimizer and CPLEX(TM) Mixed Integer Optimizer (version 2.0)*. CPLEX optimization Inc., Incline Village, Nevada.
(45) *S-PLUS 2000*; Data Analysis Products Division, Mathsoft Inc., Seattle, Washington, 98109.

(46) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(47) Kaliszan, R. Quantitative Structure-Retention Relationships Applied to Reversed-Phase High-Performance Liquid-Chromatography. *J. Chromatogr. A* **1993**, *656*, 417−435.
(48) Chambers, J.; Cleveland, W.; Kleiner, B.; Tukey, P. *Graphical Methods for Data-Analysis*; Wadsworth, 1983.
(49) Kopaciewicz, W.; Rounds, M. A.; Fausnaugh, J.; Regnier, F. E. Retention Model for High-Performance Ion-Exchange Chromatography. *J. Chromatogr.* **1983**, *266*, 3-21.
(50) Cohen, B. E.; McAnaney, T. B.; Park, E. S.; Jan, Y. N.; Boxer, S. G. et al. Probing protein electrostatics with a synthetic fluorescent amino acid. *Science* **2002**, *296*, 1700−1703.

CI025580T