Anal. Chem. **2003,** *75,* 3563−3572

Prediction of the Effect of Mobile-Phase Salt Type on Protein Retention and Selectivity in Anion Exchange Systems

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

Department of Chemical Engineering, Department of Chemistry, and Department of Mathematics, Rensselaer Polytechnic Institute, Troy, New York 12180

This study examines the effect of different salt types on protein retention and selectivity in anion exchange systems. Particularly, linear retention data for various proteins were obtained on two structurally different anion exchange stationary-phase materials in the presence of three salts with different counterions. The data indicated that the effects are, for the most part, nonspecific, although various specific effects could also be observed. Quantitative structure retention relationship (QSRR) models based on support vector machine feature selection and regression models were developed using the experimental chromatographic data in conjunction with various molecular descriptors computed from protein crystal structure geometries. Star plots for each descriptor used in the final model were generated to aid in interpretation. The resulting QSRR models were predictive, with cross-validated r^2 values of 0.9445, 0.9676, and 0.8897 for Source 15Q and 0.9561, 0.9876, and 0.9760 for Q Sepharose resins in the presence of three different salts. The predictive power of these models was validated using a set of test proteins that were not used in the generation of these models. Interpretation of the models revealed that particular trends for proteins and salts could be captured using QSRR techniques.

Each separation and purification process presents unique challenges due to the variety of proteins, the different nature of contaminants and impurities, and the quantity of product to separate from the media. Developing elution methodologies with enough selectivity to remove closely related impurities may require a lot of scouting to find the proper combination of stationary-phase material and mobile-phase conditions. Finding the right operating conditions can be challenging where various mobile-phase modifiers, salt types, pH, and possibly displacer molecules need to be screened to optimize the selectivity of the chromatographic system. The combination of the operating conditions and the type of chromatographic material has a strong impact on developing efficient separation methodologies.

To date, a number of publications have appeared in the literature to address the effect of displacing salt type and concentration on protein selectivity for ion exchange systems. The results indicated that both co-ion and counterion had an effect on the protein selectivity.[1−7] Kopaciewicz and co-workers[1] had demonstrated in their work that protein retention on an ionic surface is the result of protein charge, surface charge, and charge characteristics of the surrounding medium, as well as the type of the displacing salt. Their results indicated that while cation type slightly altered the selectivity, the anion type significantly affected the retention time as well as the selectivity in anion exchange systems. Results also indicated that a classification of salts into different categories according to the power of the displacing salt (weak, moderate, and strong) could be possible.[2]

Barron and Fritz[3] demonstrated the dependence of the strength of the salt type on the functional groups located on the stationary-phase surface; the changes were attributed to the “water−structure induced ion pairing” for ion exchange systems. The mechanism of the water−structure induced ion pairing was successful in explaining the favorable association of the large and polarizable ions with the quaternary ammonium groups on the surface. It has also been demonstrated that the chaotropic and kosmotropic salts, and their combinations, would affect the number of charged groups involved in the adsorption/desorption process.[4] In addition, the changes in the displacing salt type and gradient mode significantly enhanced the selectivity of the closely related variants.[5]

Most of the previous work reported in the literature concluded that displacing salt-type effects are specific to the proteins employed[2−6] for ion exchange systems. On the other hand, Malmquist and Lundell[7] had concluded that the effect was nonspecific and can be explained by the changes in the apparent

* Corresponding author. Tel: (518) 276-6198. Fax: (518) 276-4030. E-mail: crames@rpi.edu.

[†] Department of Chemical Engineering.

[‡] Current address: Merck Research Labs, RY805S-100, P.0. Box 2000, 126 E. Lincoln Ave., Rahway, NJ 07065.

[§] Department of Chemistry.

[||] Department of Mathematics.

(1) Kopaciewicz, W.; Rounds, M. A.; Fausnaugh, J.; Regnier, F. E. *J. Chromatogr.* **1983**, *266*, 3−21.

(2) Kopaciewicz, W.; Regnier, F. E. *Anal. Chem.* **1983**, *55*, 251−259.

(3) Barron, R. E.; Fritz, J. S. *J. Chromatogr.* **1984**, *284*, 13−25.

(4) Hodder, A. N.; Aguilar, M. I.; Hearn, M. T. W. *J. Chromatogr.* **1989**, *476*, 391−411.

(5) Hodder, A. N.; Aguilar, M. I.; Hearn, M. T. W. *J. Chromatogr.* **1990**, *506*, 17−34.

(6) Rounds, M. A.; Regnier, F. E. *J. Chromatogr.* **1984**, *283*, 37−45.

(7) Malmquist, G.; Lundell, N. *J. Chromatogr.* **1992**, *627*, 107−124.

10.1021/ac0263519 CCC: $25.00 © 2003 American Chemical Society
Published on Web 05/31/2003

gradient slope. However, to date, no theory could be established that would quantitatively explain the effect of displacing salt type.[1,6,7]

In this work, the main focus of interest is evaluating the effect of salt type on chromatographic retention for several proteins in anion exchange systems. A fundamental understanding of the efficacy of different salts requires a sufficient amount of experimental data. Differing from the previous studies, a large number and variety of proteins were evaluated for their chromatographic retention on anion exchange systems, which will enable the discussion of the specific and nonspecific effects of displacing salt type. Protein retention data were obtained using linear gradient experiments performed using a constant gradient slope in the presence of three salts: NaCl, NaBr, and Na_2SO_4. Experiments were carried out on two different stationary phases, Source 15Q (poly(styrene−divinylbenzene)) and Q Sepharose (agarose). These experiments enabled us to capture the selectivity changes due to either stationary-phase backbone chemistry or salt type.

The retention data on two stationary phases were used to generate quantitative structure retention relationship (QSRR) models for different salts as previously discussed by Mazza and co-workers.[8] In the present work, 2D, 3D (molecular operating environment, MOE), and transferable atom equivalent (TAE) descriptors[9,10] were calculated using the crystal structure geometries (PDB files) of the proteins. Linear methods, such as multiple linear regression or partial least-squares regression, have been widely applied in the QSAR/QSRR area based on an assumption of linearity between the descriptors (indirectly representing free energy increments) and an investigated experimental property. This assumption of linearity does not hold true in all cases, especially when complicated biological phenomena are represented by sets of necessarily imperfect descriptors. Additionally, the retention behavior of proteins on chromatographic media is controlled by a complex balance of interactions between the solute, resin, and solvent, resulting in apparently nonlinear relationships between the descriptors and the retention times. In this work, a set of predictive nonlinear QSRR models were derived using state-of-the-art machine learning methods, support vector machine (SVM) regression with bootstrapping techniques, in an attempt to extract a maximum amount of information from the descriptors used in the study.[11] Using this methodology, the original training set of proteins is further subdivided into a *validation set*, with the remaining proteins used for a training set. This procedure is repeated 20 times, resulting in the construction of 20 models using different training and validation sets. The predictive quality of the models is initially determined by their performance on the validation sets, but the true predictive power is only revealed when predictions are made using the "true unknowns"—the proteins held back as the test set. When the predictions made by all 20 models are combined, the result is a bootstrap aggregates (BAGGED) result. It is shown that the resulting BAGGED QSRR models are able to successfully predict the retention behavior of the remaining proteins (test set) in the database. Interpretation of the resulting models enables the importance of various structural and electronic features of the proteins to be revealed. Elucidation of the changes in importance of these features under different salt conditions may allow both prediction of protein retention under different conditions and interpretation of the chemical effects associated with these changes.

EXPERIMENTAL PROTOCOL

Materials. Strong anion exchange (quaternary ammonium) materials, Source 15Q (15 μm) and Q Sepharose HP (34 μm) were donated by Amersham Biosciences (Uppsala, Sweden). These bulk stationary phases were slurry packed into 50 × 5 mm i.d. columns that were donated by Amersham Biosciences. The following proteins were purchased from Sigma (St. Louis, MO): adenosine deaminase, alkaline phosphatase, human serum albumin (HSA), bovine serum albumin (BSA), carboxylesterase, trypsin inhibitor, glycosylasparaginase, bovine β-lactoglobulin B, bovine β-lactoglobulin A, catalase, endoglucanase I, β-galactosidase, α-lactalbumin, amyloglucosidase, insulin, ovalbumin, lectin (peanut), lipase, ovalbumin, pepsin, calmodulin, and lipoxygenase. Apoferritin and urease were purchased from ICN Biomedicals, Inc. (Aurora, OH). Sodium chloride, sodium bromide, and sodium sulfate were purchased from Fischer Scientific (Pittsburgh, PA). Tris-HCl and Tris-base were purchased from Sigma.

Apparatus. Linear gradient experiments were carried out using a model 600 multisolvent delivery system, a model 712 WISP autoinjector, and a model 996 Photodiode array absorbance detector controlled by a Millenium chromatography manager (Waters, Milford, MA).

Procedures. Linear gradient experiments were carried out with a constant slope between buffer A (20 mM Tris, pH 7.5) and buffer B (20 mM Tris with 600 mM NaCl, NaBr, or Na_2SO_4). The linear gradient slope for these experiments was 6 mM salt concentration per column volume. Aliquots of 20 μL of protein solutions with a concentration of 4 mg/mL were injected, and the experiments were carried out in duplicate at a flow of 0.5 mL/min. For these experiments, the absorbance was monitored between 215 and 280 nm.

QSRR MODELING AND SVM REGRESSION MODELS

To construct a set of informative QSRR models, electron density-based TAE quantum mechanics descriptors were combined with a set of traditional 2D and 3D descriptors obtained using the MOE program from CCG. Using this hybrid set of descriptors, a SVM sparse regression algorithm is applied in a feature selection mode to determine a subset of relevant molecular property descriptors for each of the training sets involved in the bootstrapping procedure. Subsequently, nonlinear SVM models are built based on those relevant descriptors. The overall modeling scheme is shown in Figure 1.

Implementation. MOE (Chemical Computing Group, Inc, Montreal, Canada) software was used to obtain a set of traditional 2D topological and 3D geometry-dependent molecular descriptors. The locally developed RECON2000 program was employed for generating TAE descriptors for all proteins used in the study. An in-house SVM regression program was developed independently by Dr. Bennett and Jinbo Bi in Department of Mathematics at Rensselaer Polytechnic Institute.[12]

[8] Mazza, C. B.; Sukumar, N.; Breneman, C. M.; Cramer, S. M. *Anal. Chem.* **2001**, *73*, 5457−5461.

[9] Breneman, C. M.; Rhem, M. *J. Comput. Chem.* **1997**, *18*, 182−197.

[10] RECON2000, program locally developed by Breneman, C. M. and Sukumar, N., RPI, Troy, NY, 2000.

[11] Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.

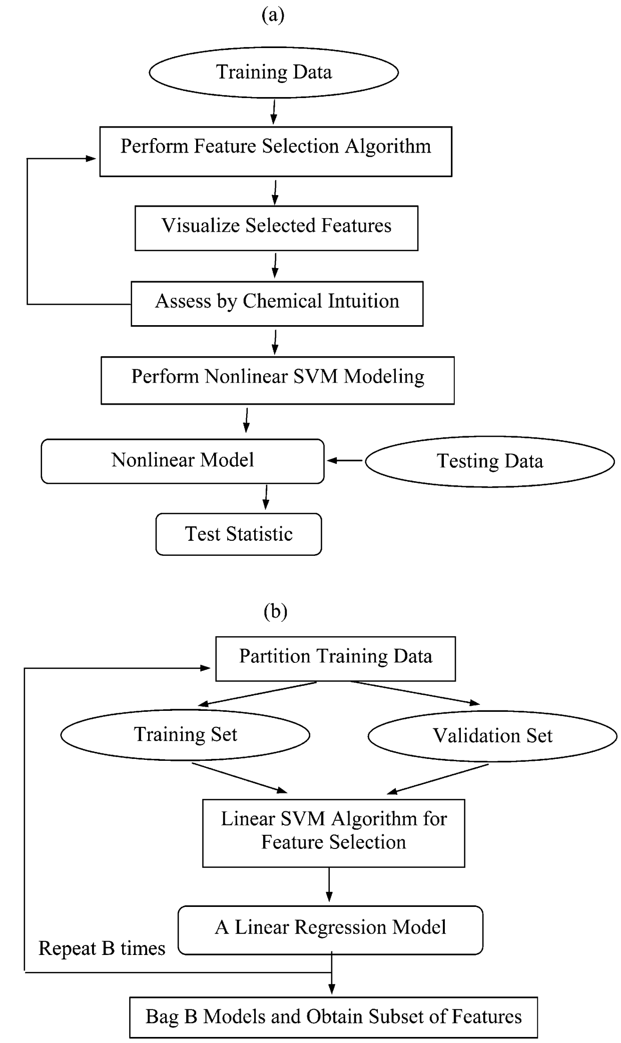
(a)
Training Data
Perform Feature Selection Algorithm
Visualize Selected Features
Assess by Chemical Intuition
Perform Nonlinear SVM Modeling
Nonlinear Model
Testing Data
Test Statistic
(b)
Partition Training Data
Training Set
Validation Set
Linear SVM Algorithm for Feature Selection
A Linear Regression Model
Repeat B times
Bag B Models and Obtain Subset of Features

Figure 1. Computational chemical property design and model validation.

Descriptor Generation. The TAE/reconstruction (TAE/RECON)[13] method utilizes a new, rapid charge density reconstruction algorithm that utilizes atomic charge density fragments that have been precomputed using ab initio wave functions. In principle, a library of atomic charge density components (TAEs) can be used to construct molecular electron densities in a form that allows for rapid retrieval of the molecular surface properties needed to generate descriptors. For each calculated molecule, the RECON program reads in molecular structure information and then reconstructs the electronic properties of the molecular surface from the atomic fragments. The distributions of several electronic properties on molecular surfaces may then be quantified to give a large variety of numerical QSRR descriptors. The CPU and disk resources required for TAE reconstruction are minimal—the electronic property distributions of ~25 proteins may be computed in ~80 s on a single-headed 1.7 GHz Linux workstation.

MOE Descriptors. The MOE program provides a combination of several types of traditional molecular property descriptors, including connectivity-based topological 2D descriptors, physicochemical property descriptors, shape-independent 3D molecular features, and some pharmacophoric descriptors. These descriptors were calculated for the proteins using the QuaSAR descriptors module in the MOE package.

Support Vector Machine Modeling. The SVM method proposed by Vapnik and co-workers[11] is based on statistical learning theory. This method has proven to be very effective for addressing general purpose classification and regression problems. SVMs have been successfully applied to a wide range of pattern recognition problems, including quality control classifications, "needle in a haystack" classification searches, and robust regression modeling. In most of these cases, the performance of SVM modeling either matches or is significantly better than that of traditional machine learning approaches, including artificial neural networks. The SVM method has a number of interesting properties, including an effective avoidance of overfitting, which improves its ability to build models using large numbers of molecular property descriptors with relatively few experimental results in the training set. Although SVM was originally developed for pattern recognition, it was later extended to solve the regression problem by Vapnik.[11] In this paper, we only focus on the use of support vector regression for creating QSRR models of protein retention. To summarize the operation of an SVM modeling procedure, it is important to consider some fundamental principles of SVMs: With a given set of training data, its objective is to find a function $f(x)$, called the ϵ-insensitive loss function, that has less than ϵ deviation from the experimental protein retention data for all cases in the training set. In other words, those predicted retention times within the ϵ distance of actual response are not penalized for being erroneous. Only those prediction points beyond ϵ of real response values are considered to contain modeling errors and are included in the "loss function". In a classical SVR, a function $f(x_i)$ is found that minimizes the overall regularized risk:

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

In the above formula, the first term $\Sigma_{i=1}^{M}|y_i - f(x_i)|_\epsilon$ represents the ϵ-insensitive losses associated with the training error and the l_2-norm $^1/_2||w||^2$ of normal vector is a regularization factor that controls the model complexity. ω is a weight vector to be determined in the function f. The parameter C is a regularization factor that determines the tradeoff between the above two terms.

This technique helps to control the complexity of the model and tends to minimize the risk of overfitting. In typical QSPR studies, many more descriptors are initially available than the number of molecules in the data set and usually include some redundant or irrelevant variables. To identify only the relevant descriptors for a particular problem, variable selection techniques are always employed to choose informative descriptors and eliminate irrelevant descriptors from consideration. The application of this type of feature selection serves to improve the computational signal-to-noise ratio in the resulting models. In this study, we applied a feature selection approach based on the linear l_1-norm SVM regression.[12] The regulation factor is applied using the l_1-norm $^1/_2||w||^2$, instead of the l_2-norm $^1/_2||w||^2$, so a linear

[12] Bennett, K.; Bi, J.; Embrechts, M.; Breneman, C. M.; Song, M. *J. Machine Learn. Res.*, in press.

[13] Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. *Comput. Chem.* **1995**, *19*, 161.

algorithm can be formulated for the SVM to reduce the computational cost compared with one using a quadratic algorithm.

Within this technique, a series of linear SVM models (usually 20 in our case) that exhibit good generalization are constructed. In each linear l_1-norm SVM or bootstrap, the optimal weight vector will have relatively few nonzero weights with the degree of sparsity depending on the SVM model parameters. The method exploits the fact that linear SVM with l_1-norm regularization inherently performs feature selection as a side effect of minimizing capacity in the SVM model. Those features with nonzero weights then become potential attributes to be used in the nonlinear SVM. To avoid a loss of useful information during the linear feature selection step, an ensemble of linear SVM models is used as part of the feature selection procedure. This approach captures important effects that might otherwise have been lost in a single-model style feature selection approach. As part of this algorithm, the important features for each individual linear SVM model are recorded and combined together to produce a final descriptor set that contains chemical information about protein retention behavior in a particular anion exchange system. In this way, the probability of inadvertently discarding useful descriptors is reduced. Finally, nonlinear SVM predictive models are constructed based on this final union descriptor set. Comparisons between linear and nonlinear predictions show that trends are preserved, but the use of nonlinear modeling methods significantly improves the results. To get more robust and general predictive results, multiple QSRR models based on the same feature set are built. So instead of using a single model, which is heavily and easily affected by chance correlations, the average of all model predictions is used as our final prediction results. This kind of debiasing technique or "bagging" is commonly employed by those in the statistical analysis field.[14]

RESULTS AND DISCUSSION

Chromatographic retention and selectivity changes of 24 different proteins were evaluated in the presence of three different salts on two different resins, Source 15Q and Q Sepharose.

Experiments. Selection criteria for proteins in this study were their p*I*, so that they would be retained on anion exchange resins, the availability of their crystal structures (PDB files), and the diversity of their structures. Most of the p*I* values could be found in the literature;[15] otherwise, the theoretical p*I* value was calculated using the EXPASY tool (http://www.expasy.ch/). The PDB files were downloaded from the Protein Data Bank web site (http://www.pdb.org/). Table 1 shows the names of the proteins and corresponding PDB files that were used for this work.

The two stationary phases that were used in this study were chosen due to their structural differences. Both backbone and spacer arm chemistries are different for these resins. Q Sepharose is an agarose-based matrix with hydrophilic properties, whereas Source 15Q is a hydrophilized poly(styrene−divinylbenzene) resin that is relatively more hydrophobic. Both stationary phases are strong anion exchange materials bearing quaternary ammonium functional groups.

Three salts were employed for this study: NaCl, NaBr, and Na_2SO_4. In anion exchange, the relative binding strength of these ions is known to be SO_4^{-2} > Br^{-1} > Cl^{-1}. The anion with the higher binding strength will be a stronger displacing salt for these systems. Among these ions, SO_4^{-2} is a kosmotrope (strongly hydrated ion), Cl^{-1} is a borderline chaotrope, and Br^{-1} is a chaotrope (weakly hydrated ion).[16]

Table 1. Proteins and Their Corresponding PDB Files Employed for This Work

PDB code	protein name	function
1A4L	adenosine deaminase	hydrolase
1AIV	conalbumin	Iron transport protein
1AJC*	alkaline phosphatase	nonspecific monoesterase
1AO6*	human serum albumin	carrier protein
1AUO	carboxylesterase	hydrolase
1AVU	trypsin inhibitor	serine protease inhibitor
1AYY	glycosylasparaginase	hydrolase
1BEB	bovine *β*-lactoglobulin B	lipocalin
1BSO	bovine *β*-lactoglobulin A	transport protein
8CAT	catalase	oxidoreductase (H_2O acceptor)
1EG1	endoglucanase I	cellulose degradation
1F4H	*β*-galactosidase	hydrolase
1F6S	α-lactalbumin	metal-binding protein
1FWE	urease	hydrolase
3GLY*	glucoamylase	hydrolase
1IES	apoferritin	iron storage
4INS	insulin	hormone
1LPN	lipase	hydrolase
1OVA	ovalbumin	serpin
2PEL	peanut lectin	lectin (agglutinin)
3PEP	pepsin	hydrolase (acid proteinase)
1QIW	calmodulin	calcium-binding protein
1UOR	serum albumin	metal-binding protein
1YGE	lipoxygenase	dioxygenase

The comparison of the average retention times of the proteins on the Source15Q resin in the presence of three different displacing salts is shown in Figure 2a. The average retention times are sorted in increasing order for the case of NaCl. As seen in the plot, the general retention behavior of these proteins in the presence of NaCl and NaBr is similar. On the other hand, the presence of Na_2SO_4 causes the retention times for the proteins to be significantly decreased. In addition, there are a few cases where the order of elution was observed to change. One such protein pair is amyloglucosidase (AMY) and apoferritin (APO). Even though these proteins displayed similar retention behavior in NaCl and NaBr (retention of APO was slightly higher than that for AMY), the elution order was found to be reversed in the presence of Na_2SO_4. Another interesting result was obtained for pepsin. Its retention time decreased when the salt type was changed from NaCl to NaBr, as was the case for several of the proteins; however, its retention time was greater in the presence of Na_2SO_4 than it was in the presence of NaBr—an observation that was unique for this protein. These results confirm previous results reported in the literature[1−7] and indicate that the retention of different proteins is altered to various degrees depending on the type of displacing salt type that is employed.

Figure 2b shows the comparison of the average retention times of proteins on Q Sepharose HP stationary-phase material in the presence of different salt types. One of the first observations is a

[14] Breiman, L. *Machine Learn.* **1996**, *24*, 123−140.

[15] Righetti, P. G.; Caravaggio, T. *J. Chromatogr.* **1976**, *127*, 1.

[16] Collins, K. D. *Biophys. J.* **1997**, *72*, 65−76.

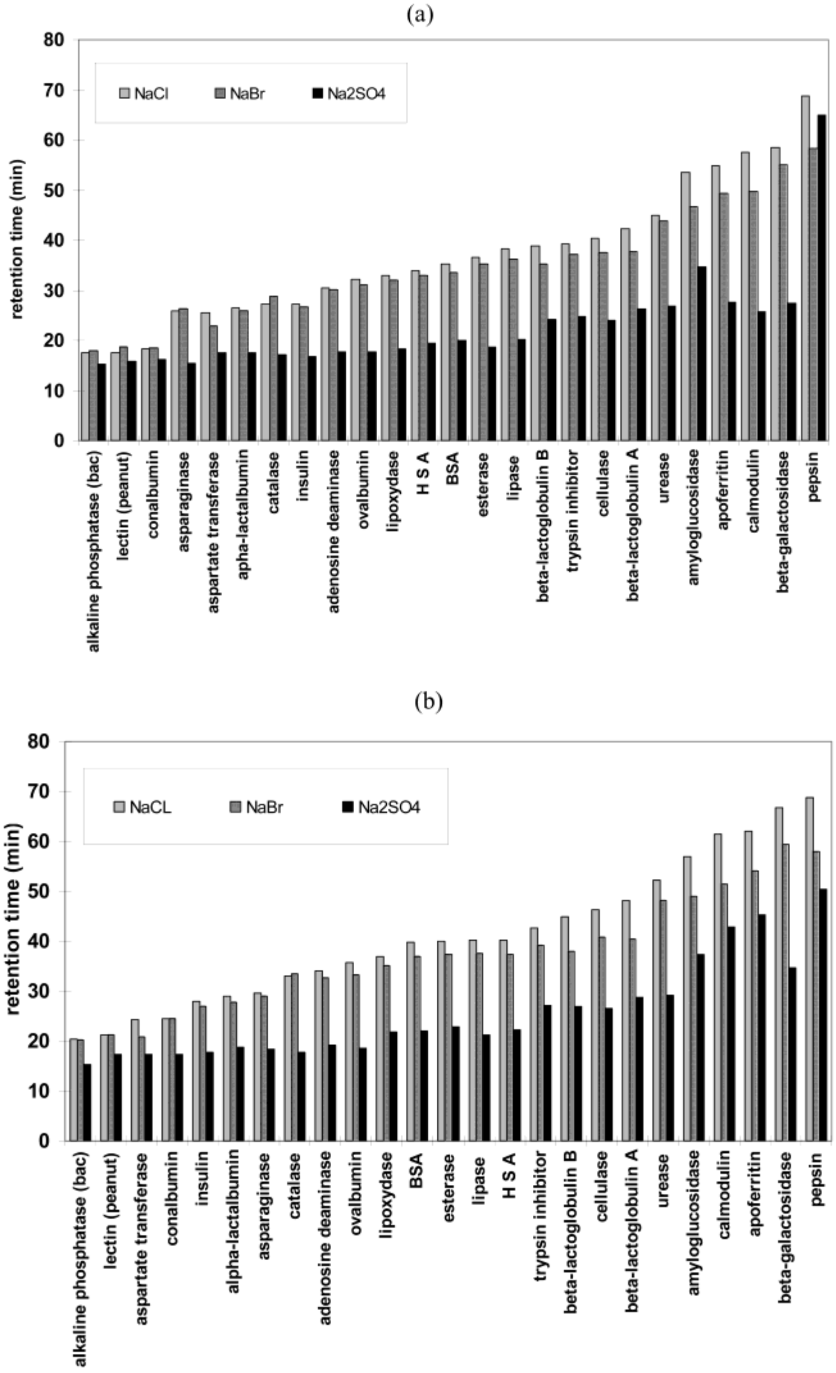

Figure 2. Protein retention times on (a) Source 15Q and (b) Q Sepharose HP in the presence of three different salts.

general increase in the average retention time of all proteins when the stationary-phase type was changed from Source 15Q to Q Sepharose HP. This agrees with the results that were obtained in our laboratory for cation exchange systems[8] using stationary phases with the same backbone chemistry. While the retention was slightly decreased in NaBr as compared to NaCl, the presence of Na_2SO_4 was associated with significantly decreased retention times. In addition, the elution order was significantly different in Na_2SO_4 as compared to the other salts. For example, β-galactosidase and apoferritin represent a pair where selectivity reversal had taken place. In the presence of NaCl, β-galactosidase was more retained than apoferritin, whereas the elution order changed in favor of apoferritin when the salt type was changed to Na_2SO_4. On the other hand, the retention time for pepsin did not show any unique trends as in the previous case using Source15 Q.

These results using two different stationary phases have shown that the effect of displacing salt type is, for the most part, nonspecific; that is, as the displacing salt-type strength was increased, the retention times of most of the proteins were decreased. However, the order of elution for several proteins changed, indicating specific effects of displacing salts on selectivity. These experimental results indicate that significant changes can occur depending on the stationary-phase chemistry and displacing salt type. This is the reason that various salts and resins (as well as pH) are often examined during methods development to determine the most selective ion exchange gradient systems.

Table 2. Definition of the Descriptors Selected in the Feature Selection for Modeling Protein Retention

descriptor name	chemical information encoded in these descriptors
	Descriptors with Negative Contribution
PMIX (MOE)	*x* component of the principal moment of inertia (external coordinates)
PMIZ (MOE)	*z* component of the principal moment of inertia (external coordinates)
DEL. K. IA (TAE)	gradient of the K electronic kinetic energy normal to the molecular surface that describes the differences in the polarizability and hydrophobicity of molecular regions; electrophilicity
PEOE.VSA.FPPOS (MOE)	fraction of positive polar van der Waals surface; the partial equalization of orbital electronegativities (PEOE) is a method of calculating the atomic charges[17]
STD.DIM 1 and 2 (MOE)	square root of the first and second largest eigenvalues of the covariance matrix of the atomic coordinates; a standard dimension is equivalent to the standard deviation along a principal component axis. “size and shape”-related descriptor
VSA.POL (MOE)	sum of VDW surface of “polar” atoms
FCHARGE (MOE)	formal charge, negative contribution means negative charge is favored
	Descriptors with Positive Contribution
STD.DIM 3 (MOE)	the square root of the third largest eigenvalue of the covariance matrix of the atomic coordinates
SIEP (TAE)	surface integral of electrostatic potential as determined on the VDW surface (electron density surface)
SIEPIA (TAE)	integral average of the SIEP
DIPOLE, DIPOLE Z (MOE)	first derivative of energy with respect to an applied electric field; it is a measure of the asymmetry in the molecular charge distribution; it has three components: DIPOLEX, DIPOLEY, DIPOLEZ along the *x*, *y*, and *z* axes, as well as a magnitude DIPOLE; moment of charge in the molecule
PIP1 (TAE)	Politzer ionization potential; the first histogram bin of PIP property; local average ionization potential in the low 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 to SIKIA (opposite sign); complementary over the integral of the whole molecule
PEOE.VSA.−3 (MOE)	sum of VDW surface area where charge is in the range of [−0.20, −0.15)
GLOB (MOE)	globularity, or inverse condition number (smallest eigenvalue divided by the argest eigenvalue), of the covariance matrix of atomic coordinates
DEL. G.NMIN (TAE)	minimum value of the surface integral of G kinetic energy

QSRR Modeling. This set of experimental retention data was also employed to generate predictive QSRR models using electron density-based descriptors and SVM regression modeling. The aim of this work was to generate models that could predict protein retention directly from crystal structure data and that could explain or capture the main interactions that take place in anion exchange systems, as well as the effects associated with the presence of different salts. As part of the modeling process, star plots were generated to give a better understanding of the relative importance of each descriptor for the models being generated. A star plot is a multiplot radial visualization of a multivariate data matrix, and it is explained in the following section. The relevant QSRR descriptors include shape, size, surface property, and electron density-derived descriptors. The definitions of all these descriptor types are given in Table 2. Although some of the descriptors are difficult to interpret in the physical/chemical sense, some of them are linked with chemical features that are relevant to our experimental system. These implied relationships are discussed in detail in the following section and can facilitate our understanding of QSRRs for protein retention.

Figures 3a, 4a, and 5a show the correlation between the experimental and predicted results for Source 15Q material in the presence of three salts, NaCl, NaBr, and Na_2SO_4, respectively. The cross-validated r^2 for these models were 0.9445, 0.9676, and 0.8897, which indicates that the predicted values for protein retention are in good agreement with the experimental data. In fact, the retention values that were predicted for three proteins not included in the training set (alkaline phosphatase, HSA, amyloglucosidase) successfully verified the predictive power of these models (Figures 3b, 4b, and 5b). Similarly, QSRR models generated for Q Sepharose (figures not shown here for brevity) had very good correlations between the experimental and predicted protein retention values with cross-validated r^2 values of 0.9561, 0.9876, and 0.9760 in the presence of NaCl, NaBr, and Na_2SO_4, respectively. Furthermore, the retention times of three proteins in the test set were successfully predicted by the QSRR models for the Sepharose material.

These results indicate that the QSRR models generated for these experimental systems were very well correlated with the experimental data and are capable of making good a priori predictions for a wide range of proteins. The main challenge still lies in the interpretation of the descriptors used for these models. To facilitate interpretation, it was necessary to determine which descriptors were consistently important when different combinations of training and validation molecules were used. Each different set of training molecules is called a “fold”, and the model created using this set is used to make predictions on the validation molecules left out of the training set for that particular “fold”. Star plots were then created to evaluate the relative importance of each of these descriptors selected throughout each of the 20 bootstrap “folds” used for creating the composite model set. In these plots, each star corresponds to a specific descriptor (column) in the investigated weight matrix from linear models and the radius of each stroke in the star represents the weight or importance of this descriptor in each one of the 20 bootstrap iterations or

[17] Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.

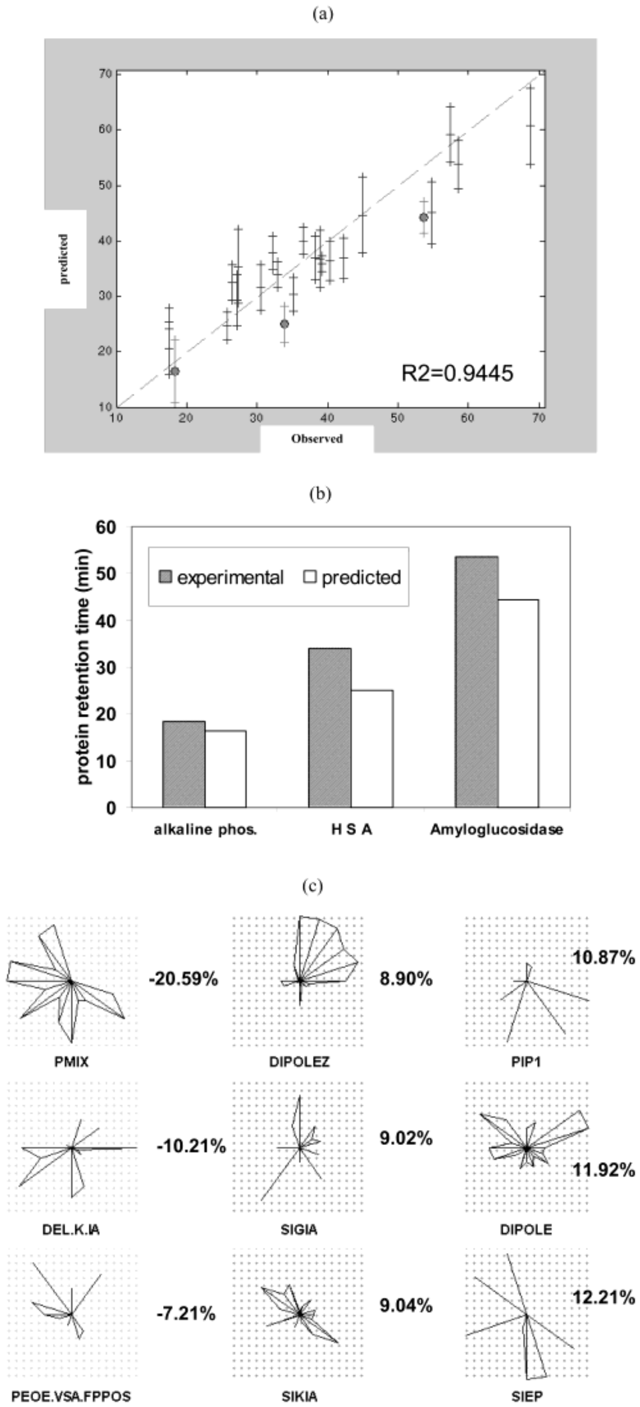

Figure 3. Source 15Q in the presence of NaCl: (a) predictions vs observed values for the entire data set, (b) comparison of predicted and experimental retention time for the test set, and (c) star plots with weight percentages for the descriptors in the model.

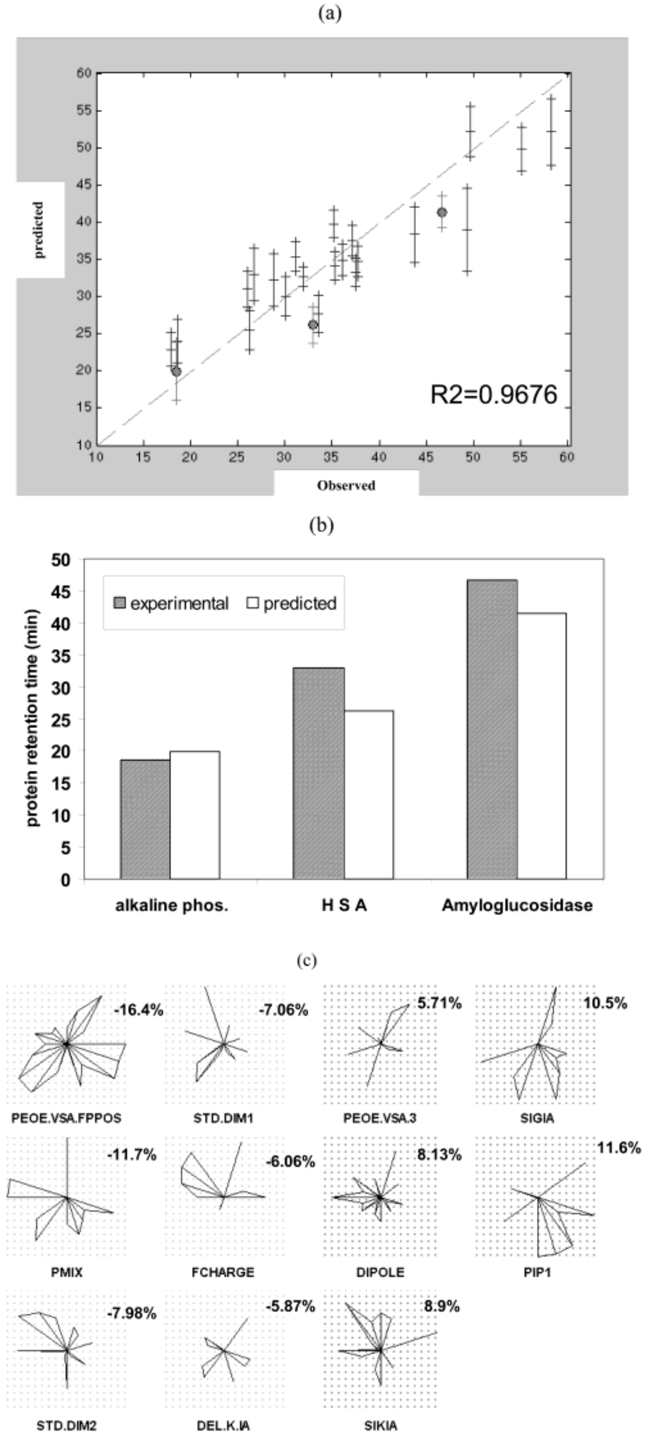

Figure 4. Source 15Q in the presence of NaBr: (a) predictions vs observed values for the entire data set, (b) comparison of predicted and experimental retention time for the test set, and (c) star plots with weight percentages for the descriptors in the model.

constructed linear SVR models. For each star plot, the selected relevant descriptors are ranked in a columnwise fashion according to their sum of the radii for all bootstrap iterations, so that the most significant descriptor with negative weight appears in the upper left-hand corner while the most significant positive contributor appears on the lower right. For instance, in Figure 4, PEOE.VSA.FPPOS has the largest negative effect on retention time while PIP1 has the largest positive effect on retention time in the case of the presence of NaBr and Source 15Q. These star plots show that surface charge-related descriptors such as PEOE.FPPOS, shape descriptors such as PMIX, and TAE descriptors such as DEL. K.IA were present in most of the models bearing highly negative coefficients (see Figures 3c, 4c, and 5c and Table 3). In addition, other TAE descriptors such as SIGIA, SIEP, and PIP1 were seen to have positive contributions on protein retention.

To aid in the interpretation, the descriptors that contribute to these models are categorized as shown in Table 4. The descriptors

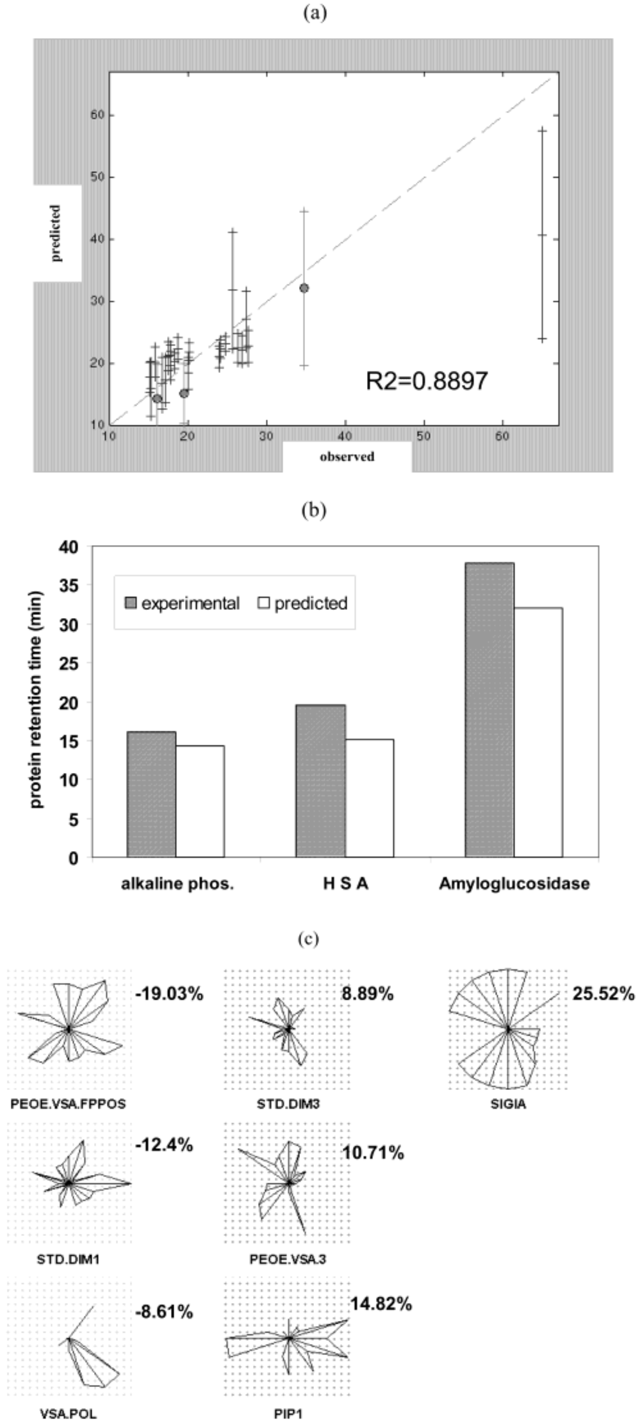

Figure 5. Source 15Q in the presence of Na_2SO_4: (a) predictions vs observed values for the entire data set, (b) comparison of predicted and experimental retention time for the test set, and (c) star plots with weight percentages for the descriptors in the model

Table 3. Correlations and Descriptor Contributions for the QSRR Models Generated for Protein Retention on Q Sepharose HP in the Presence of Three Different Salts

	descriptor contributions for Q Sepharose HP (%)		
descriptors	NaCl	NaBr	Na_2SO_4
PEOE.VSA.FPPOS	−17.65	−16.69	−16.48
PMIX	−12.87	−12.02	−9.90
STD. DIM2	−8.77	−5.96	n/a[a]
DEL. K.IA	−8.92	−5.79	−10.99
VSA. POL	n/a	−5.08	−12.15
DEL. G.NMIN	n/a	3.13	n/a
GLOB	n/a	3.62	n/a
SIKIA	7.03	4.27	6.58
DIPOLEZ	6.13	5.09	n/a
PEOE.VSA.3	6.74	5.23	n/a
DIPOLE	10.72	6.63	n/a
PIP1	n/a	6.93	n/a
SIGIA	n/a	7.58	16.09
SIEP	15.06	11.92	12.80
PMIZ	n/a	n/a	6.56
SIPEIA	n/a	n/a	8.42
FCHARGE	−6.07	n/a	n/a

[a] n/a, not available.

were found to fit into three classes: net charge-related, shape/charge-distribution, and hydrogen bond capacity.

As seen in the results, the PEOE.VSA.FPPOS descriptor is common to all six models generated for each stationary-phase material. This descriptor defines the fractional positive polar region of the molecular van der Waals (VDW) surface area. Associated with high negative contributions to retention, this descriptor suggests that as the fraction of positive polar VDW surface area increases, the protein retention will decrease. This is consistent with known aspects of the net charge phenomenon in anion exchange chromatography and shown in our experiments in that positively charged solutes will be less well retained. The same phenomenon can explain the positive contribution of the PEO-E.VSA.-3 descriptor, which is the sum of the VDW surface area associated with a specific range of negative partial charge. The FCHARGE descriptor is also related to the same effect and is therefore favorable for protein retention.

As shown in the star plots and tables, descriptors with information concerning the nature of the electron density distribution (such as PIP1 and SIGIA) as well as polar descriptors (such as DIPOLE and SIEP) have significant positive contributions in the protein retention time models. These descriptors are strongly related to the dipolarity and polarizability of the groups that make up the protein surface. This demonstrates that not only the formal charge of the protein itself may affect retention time but protein retention may also be influenced by the distribution and electronic kinetic energy of the charge density present on the protein surface, causing specific regions of that surface to be crucial to protein retention times.

The MOE DIPOLE descriptor defines the distribution of the partial charges on the VDW surface area, and DIPOLEZ shows the dipole moment in the *Z* direction in the protein data bank orientation. These descriptors were present in most of the models generated in this work as having positive contributions to retention. These descriptors refer to the heterogeneity of the charge distribution in the proteins. SIGIA and SIKIA define the G or K electronic kinetic energy density on VDW surface. These descriptors are associated with the presence and strength of Lewis basic sites. The DEL. K.IA descriptor is associated with polarizability and the hydrophobicity of portions of molecular surfaces. Positive weighting and small values of this descriptor quantify the higher falloff rate of electronic kinetic energy density for an H-bond donor as compared to an H-bond acceptor. This descriptor was found to be negatively correlated with the PEOE.VSA.FPNEG

Table 4. Categorization of Descriptors for All the Models (Italic Descriptors Are Positively Contributing to the Model)

	Source 15Q			Q Sepharose		
salt	net charge-related	shape/charge distribution	hydrogen-bonding capacity	net charge-related	shape/charge distribution	hydrogen-bonding capacity
NaCl	PEOE.VSA.FPPOS	PMIX		PEOE.VSA.FPPOS	PMIX	
		SIEP		F.CHARGE	STD.DIM2	
		DIPOLE		*PEOE.VSA.−3*	*DIPOLEZ*	
		PIP1			*SIEP*	
		SIKIA			*DIPOLE*	
		SIGIA				
		DEL. K.IA				
		DIPOLEZ				
NaBr	PEOE.VSA.FPPOS	PMIX		PEOE.VSA.FPPOS	PMIX	VSA.POL
	PEOE.VSA.−3	STD. DIM2		*PEOE.VSA.−3*	STD. DIM2	
		STD. DIM1			*GLOB*	
		DEL. K.IA			*SIEP*	
		PIP1			*SIGIA*	
		SIGIA			*PIP1*	
		SIKIA			*DIPOLE*	
		DIPOLE			*SIKIA*	
					DEL. G.NMIN	
					DEL. K.IA	
					DIPOLEZ	
Na_2SO_4	PEOE.VSA.FPPOS	STD. DIM1	VSA.POL	PEOE.VSA.FPPOS	PMIX	VSA.POL
	PEOE.VSA.−3	*STD.DIM3*			*PMIZ*	
		SIGIA			DEL. K.IA	
		PIP1			*SIGIA*	
					SIEP	
					SIEPIA	
					SIKIA	

(molecular fraction of polar negative VDW surface area), which indicates a positive correlation with protein retention. The SIKIA has negative values and is generally anticorrelated with DEL. K.IA because the larger negative values of SIKIA imply that there will be a greater rate of falloff normal to the surface. SIGIA (sometimes associated with hydrophobicity), on the other hand, is found to be positively correlated with PEOE.VSA.FPNEG, in agreement with expectations. SIEP defines the surface integral of electrostatic potential, which in this case is found to be well correlated with the accessible surface area of the proteins. SIEPIA may indicate the average characteristics of amino acids with respect to their electrostatic potential for each protein. PIP1 relates to the lowest bin of the local ionization potential, which correlates to easy ionization or loosely held electron density. In our models, it shows a positive correlation with the size of the proteins, but this is to be expected because of a general increase in overall surface area with size. Thus, it may be said that as the size of the proteins and their accessible surface areas are increased, their retention should also increase.

Shape and size descriptors were also found to be important for all six models. For example, among these, PMIX and PMIZ descriptors define the *x* and *z* components of the principle moment of inertia, respectively. STD.DIM defines the square root of the largest eigenvalue, and GLOB defines the globularity, which is the smallest eigenvalue divided by the largest eigenvalue. For those models where PMIX and STD.DIM1 and 2 had negative contributions, descriptors such as PMIZ, STD.DIM3, and GLOB had positive contributions. STD.DIM1 and DIM3 show complementary effects. On the other hand, STD.DIM2 and PMIX are also modulated by GLOB. When these results are analyzed together, they suggest that globular conformation of proteins is favored to increase retention on either stationary-phase material.

VSA.POL descriptors define the van der Waals surface area of polar atoms, as computed on the neutral molecule. These descriptors represent the capacity of proteins for hydrogen bonding and are found to be anticorrelated with retention times. This is an expected result, since an increase in hydrogen bond capacity as indicated by a large polar surface area will facilitate the interaction of the solvent with the protein, aiding in rapid elution from the column.

In the models generated for the experimental data obtained in the presence of Na_2SO_4, it has been observed that SIGIA had a larger positive contribution than that for other models. The SIGIA descriptor contains information about molecular polarizability and the ability of proteins to become involved in charge-induced dipole and dipole−induced dipole interactions. Proteins with high values of SIGIA are more capable of participating in these interactions, making it easier to displace sulfate ions from the surface of the resin. In addition, the DIPOLE descriptor is also absent in these models, which indicates that nonpolar solutes are favored over polar ones. Therefore, hydrophobic, nonpolar solutes will be favored for retention in the presence of this salt. In addition, the VSA.POL descriptor is selected as being important for the models developed for both stationary phases in the presence of Na_2SO_4. This descriptor, which quantifies polar VDW surface area, is observed to favor elution over retention in these systems. This observation indicates that protein molecular surface polarity is not favored for retention, contrary to intuitive expectation. This result suggests that other factors such as hydrophobicity may be important in these systems or that the VSA.POL descriptor is acting as a surrogate for molecular size.

CONCLUSIONS

In this paper, the effects of different counterions were investigated for various proteins on two different stationary-phase materials. The results indicate that the displacing salt strength agrees with the common knowledge such that Na_2SO_4 is a stronger salt than NaBr and NaCl. In addition, the elution time of proteins decreased as the salt displacing strength increased, which implies nonspecific effects of different salts generic to most of the proteins. However, the elution order of proteins for various cases was changed, which indicates a specific effect of salt type on each protein. Therefore, the results with large number of proteins from various classes indicate that salt-type effects are, mostly, nonspecific; however, a significant number of specific effects can be observed, which may be indicative of selectivity changes in anion exchange systems.

The experimental chromatographic data were used in conjunction with various molecular descriptors for generating QSRR models based on SVM feature selection and regression models. The models resulted in good correlation of experimental and model predicted data and successful predictions for the test set proteins. The star plot approach represented here has been shown to be a powerful tool to aid in interpretation of the QSRR models for their physical and chemical implications. This interpretation has shown that the models can capture the general aspects of anion exchange chromatography based on protein crystal structures. In addition, some particular trends regarding proteins such as enhancement of retention for proteins with globular conformation could be addressed with these models. Furthermore, the enhancement of hydrophobic interactions in the presence Na_2SO_4, which was in agreement with the earlier work in the literature, could be captured with the models. While these results are very encouraging, as with any QSPR model, the scope and limitations of prediction and interpretation need to be further explored. Further, although it is still not completely clear what the “details” of the descriptors mean in terms of the physics of the chromatographic process, the feature selection process provides clues as to what types of interactions are more important for these different ion exchange systems. In conclusion, this modeling approach has been shown to be very powerful not only for determining a priori protein retention but also for interpreting some specific affects resulting from different mobile-phase conditions.

ACKNOWLEDGMENT

Authors thank Dr. C. B. Mazza for her contributions toward this work. The authors acknowledge NSF Grant BES-0079436 and NIH Grant GM 47372-04A2 for support of this research.

Received for review November 25, 2002. Accepted April 8, 2003.

AC0263519