Prediction of Protein Retention in Ion−Exchange Systems Using Molecular Descriptors Obtained from Crystal Structure

ARTICLE *in* ANALYTICAL CHEMISTRY · DECEMBER 2001

Impact Factor: 5.64 · DOI: 10.1021/ac010797s · Source: PubMed

4 AUTHORS:

Cecilia Mazza
Rensselaer Polytechnic Institute
6 PUBLICATIONS **95** CITATIONS
SEE PROFILE

N. Sukumar
Shiv Nadar University
55 PUBLICATIONS **673** CITATIONS
SEE PROFILE

Curt M Breneman
Rensselaer Polytechnic Institute
103 PUBLICATIONS **4,458** CITATIONS
SEE PROFILE

Steven M Cramer
Rensselaer Polytechnic Institute
202 PUBLICATIONS **4,395** CITATIONS
SEE PROFILE

Prediction of Protein Retention in Ion-Exchange Systems Using Molecular Descriptors Obtained from Crystal Structure

C. B. Mazza,[†] N. Sukumar,[‡] C. M. Breneman,[‡] and S. M. Cramer*,[†]

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

In this paper, a novel approach is described for the a priori prediction of protein retention in ion exchange systems. Quantitative structure retention relationship (QSRR) models based on a genetic algorithm/partial least squares approach were developed using experimental chromatographic data in concert with molecular descriptors computed using protein crystal structures. The resulting QSRR models were well-correlated, with cross-validated r^2 values of 0.938 and 0.907, and the predictive power of these models was demonstrated using proteins not included in the derivation of the models. Importantly, these models were able to predict selectivity reversals observed with two different stationary phase materials. To our knowledge, this is the first published example of predictive QSRR models of protein retention based on crystal structure data.

One of the major challenges in bioprocessing is selecting the appropriate chromatographic material for a given biological mixture. The generation of predictive quantitative structure-based models for relatively small molecules in various modes of chromatography has been the focus of several publications.[1−5] However, these reports are based on the generation of models with a relatively small number of predetermined descriptors. Recently, it has been shown that quantitative structure chromatography models can be successfully derived for small molecules in ion-exchange (Mazza, C. B.; Rege, K.; Breneman, C. M.; Dordick, J. S.; Cramer, S. M. Submitted 2001.) and reverse-phase chromatographic systems[6] employing a partial least squares modeling approach (PLS) with genetic algorithm (GA) feature selection.[7] This approach is a departure from traditional QSRR modeling methods, because a large number of variables are initially calculated for the molecules in the data set, followed by a capacity-controlled GA feature selection routine. The resulting descriptor set is then employed to generate the QSRR model based on the PLS approach. Molecules not present in the training set are employed for validating the predictive power of the resulting QSRR model.

QSAR AND MOLECULAR PROPERTY MODELING

The chemical literature abounds with examples of quantitative structure−activity relationship (QSAR) approaches designed to assist in the development of bioactive compounds.[7−14] The basis for QSAR is built around the concept of linear free-energy relationships in that variations in the binding behavior of small molecules to biological systems may be quantitatively attributed to changes in structure. When structural variations among the small molecules within a dataset are not too great, useful correlations between molecular structure and biological responses may be generated and used to predict the activity of unknown compounds. The predictive power of a QSAR model is often determined by the quality and type of molecular property variables being used and by the regression method employed for model building. Although several successful methods are known, including nonlinear neural network techniques, the current work is based on the use of the robust partial least-squares regression method (PLS) because of its utility in handling large sets of nonorthogonal variables and its resistance to over-training problems.[15]

Property modeling techniques are not limited to drug design and have been generalized into a broader field known as quantitative structure−property analysis (QSPR) In such studies, numerous physical properties of molecular systems have been successfully modeled, including boiling points, aqueous solubility

[†] Department of Chemical Engineering.

[‡] Department of Chemistry.

(1) Kaliszan, R.; van Straten, M. A.; Markuszewski, M.; Cramers, C. A.; Claessens, H. A. *J. Chromatogr. A* **1999**, *855*, 455−496.
(2) Kaliszan, R. *J. Chrom. A.* **1993**, *656*, 417−435.
(3) Carr, P. W. *Microchem. J.* **1993**, *48*, 4−28.
(4) Timerbaev, A. R.; Semenova, O. P.; Tsoi, I. G.; Petrukhin, O. M. *J. Chromatogr.* **1993**, *648*, 307−314.
(5) Escuder-Gilabert, L.; Sagrado, S.; Villanueva-Camanas, R. M.; Medina-Hernandez, M. J. *Anal. Chem.* **1998**, *70*, 28−34.
(6) Breneman, C. M.; Rhem, M. J. *Comput. Chem.* **1997**, *18*, 182−197.
(7) GA/PLS, MIT software library, MA.
(8) Hansch, C.; Leo, A. *Exploring QSAR*. American Chemical Society: Washington, DC, 1995.
(9) Hopfinger, A. J. *J. Am. Chem. Soc.* **1980**, *102*, 7196.
(10) Horwell, D. C.; Howson, W.; Higginbottom, M.; Naylor, D.; Ratcliffe, G. S.; Williams, S. *J. Med. Chem.* **1995**, *38*, 4454−4462.
(11) Mazerska, Z.; Augustin, E.; Dziegielewski, J.; Cholody, M. W.; Konopa, J. *Anti-Cancer Drug Des.* **1996**, *11*, 73−88.
(12) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palyulin, V. A.; Zefirov, N. S. *J.ournal of Chem. Inf. Comput. Sci.* **1993**, *33*, 630−634.
(13) Seydel, J. K.; Trettin, D.; Cordes, H. P. *J. Med. Chem.* **1980**, *23*, 607−613.
(14) Karcher, W.; Karabunarliev, S. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 672−677.
(15) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III *SIAM J. Sci. Stat. Comput.* **1984**, *5* (3), 735.

10.1021/ac010797s CCC: $20.00
Published on Web 00/00/0000

and polymer properties[1,5,16−19] In this work, we introduce the use of new electron density-derived descriptors that are used in conjunction with traditional QSAR and QSPR descriptors to build chromatographic retention models for protein systems. This descriptor set has facilitated the rapid, direct modeling of the noncovalent interactions between chromatographic stationary phase media and each member of a protein dataset for a given set of conditions.

Traditionally, connectivity-based 2D as well as 3D descriptors have been employed for both drug design and chromatography modeling.[20−23] Improvements to these models may be obtained by also including electron-density-derived descriptors. A variety of molecular descriptors can be derived from the electron density distributions obtained from ab initio calculations;[24] however, these calculations often require extensive computational time. Although much faster, semiempirical methods are not capable of producing the type of electron density information required for appropriate descriptor generation. In order to accurately obtain molecular electron-density-derived descriptors with a substantial reduction in computational time, Breneman and co-workers have developed the transferable atom equivalent (TAE) method[25,26] that is based on the theory of atoms in molecules reported by Bader.[27] In the TAE/RECON method, atomic contributions are used to rapidly generate whole molecule electron-density-derived descriptors that approximate those available through ab initio calculations. Although the descriptors generated by the TAE method have been employed for modeling the properties of small molecules[25] (Mazza, C. B.; Rege, K.; Breneman, C. M.; Dordick, J. S.; Cramer, S. M. Manuscript submitted 2001. Tugcu, N.; Mazza, C. B.; Moore, J. A.; Breneman, C. M.; Sanghvi, Y. S.; Cramer, S. M. Manuscript submitted 2000), they have not been previously used to generate descriptors based on the electron density distributions of proteins. In this paper, it is shown that when the traditional descriptors available through the Molecular Operating Environment (MOE)[28] software are used together with a set of TAE descriptors computed using protein crystal structures as input to the RECON2000 program, they produce predictive PLS QSRR models of ion-exchange chromatographic behavior.

EXPERIMENTAL SECTION

Materials. Fast Flow S P Sepharose and Source 15S columns (1 mL) were donated by Amersham Pharmacia (Uppsala, Sweden). The following proteins were purchased from Sigma (St. Louis, MO): bovine heart cytochrome *c*, horse heart cytochrome *c*, chicken egg lysozyme, turkey lysozyme, pyruvate kinase, bee venom phospholipase A2, bovine phospholipase, pig phospholipase, protease carlsberg, trypsinogen, protease nagarse, lentil lectin, elastase, human hemoglobin, bovine hemoglobin, ribonuclease A, ribonuclease B, papain, α-chymotrypsin, γ-chymotrypsin, α-chymotrypsinogen A, and hen egg avidin. Sodium acetate was purchased from Sigma (St. Louis, MO); glacial acetic acid and NaCl were purchased from Aldrich (Milwaukee, WI).

Equipment. Analytical linear gradient experiments were carried out using a Waters 600 multisolvent delivery system, a Waters 712 WISP autoinjector and a Waters 484 UV-vis absorbance detector controlled by a Millenium chromatography software manager (Waters, Milford, MA).

Procedure. *Linear Gradient Chromatography.* Linear gradient experiments (buffer A, 50 mM sodium acetate, pH 5; buffer B, 50 mM sodium acetate and 600 mM sodium chloride, pH 5) were carried out using a linear gradient slope of 20 mM Na per column volume. Injections of 30 mL were employed for these experiments. The effluent was monitored at 280 nm, and the experiments were carried out at 0.5 mL/min. All experiments were carried out in duplicate.

MODEL GENERATION

Software. Protein crystal structures were obtained from the Protein Data Bank.[29] InsightII software (MSI, San Diego, CA) was employed for eliminating complexes or water molecules that were present in the protein crystal structures. Molecular Operating Environment (MOE, Chemical Computing Group, Inc, Montreal, Canada) software was used to obtain 2D (topological, connectivity-based) descriptors[30] and 3D (shape, surface area, charge-based) molecular descriptors.[31,32] RECON2000[26] was employed for generating the transferable atom equivalent (TAE) descriptors for the proteins. The GA/PLS approach (MIT module, MIT software library, MA), described below, was used to select the most important descriptors and to generate the QSRR models for each stationary-phase material.

Generation of QSRR Model. The molecular descriptors were computed for each protein using the MOE and RECON2000 software packages, and the resulting descriptor sets were used to build trial models for the experimental chromatographic retention data. A genetic algorithm was used in an iterative manner to select the most relevant descriptors for building a predictive PLS model. This "GA feature selection" approach reduces the number of descriptors by using the genetic algorithm to remove many unhelpful (noisy) descriptors from the model, using a "leave one out" technique[33] to test each trial model for predictive capability. This results in a QSRR model based on the partial least squares approach that contains only a subset of the initial descriptor field but that yields the best cross-validated model.

(16) Leo, A.; Hansch, C.; Church, C. *J. Med. Chem.* **1969**, *12*, 766−771.
(17) Hansch, C. *Drug Metab. Rev.* **1984−1985**, *15*, 1279−1294.
(18) Forgács, E.; Csartháti, T. *Molecular Basis of Chromatographic Separations*, 1st ed.; CRC Press: Boca Raton, 1997; pp 146−158.
(19) Topliss, J. G., Ed., *Quantitative Structure Activity Relationships of Drugs*; Academic Press: New York, 1983; pp 10−12.
(20) Breneman, C. M.; Rhem, M. *J. Comput. Chem.* **1997**, *18* (2), 182−197.
(21) Hopfinger, A. J.; Burke, B. J.; William J. Dunn, I. *J. Med. Chem.* **1994**, *37*, 3768−3774.
(22) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24* (4), 279−87.
(23) Breneman, C. M.; Martinov, M. *The Use of Electrostatic Potential Fields in QSAR and QSPR, in Molecular Electrostatic Potential: Concept and Applications*; Murray, J. S., Sen, K., Eds.; Elsevier: Amsterdam, 1996; pp 143−179.
(24) Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. *Comput. Chem.* **1995**, *19*, 161−179.
(25) Breneman, C. M.; Rhem, M. *J. Comput. Chem.* **1997**, *18*, 182−197.
(26) RECON2000, program locally developed by Breneman, C. M. and Sukumar, N.; RPI: Troy, NY, 2000.
(27) Bader, R. F. W.; Carroll, M. T.; Cheeseman, J. R.; Chang, C. *J. Am. Chem. Soc.* **1987**, *109*, 7968−7979.
(28) Molecular Operating Environment software, version 2000.02; Chemical Computing Group, Inc.: Montreal, Canada, 2000.
(29) http://www.rcsb.org/pdb/
(30) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd.: Hertfordshire, England, and John Wiley and Sons: New York, 1986.
(31) Stanton, D.; Jurs, P. *Anal. Chem.* **1990**, *62*, 2323−29.
(32) Gasteiger, J.; Marsali, M. *Tetrahedron* **1980**, *36*, 3219−23.
(33) Haykin, S. Neural Networks−A Comprehensive Foundation, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 1999; 217−218.

Briefly, the partial least squares method[34,35] strives to uncover a small number of "latent" variables from a much larger set of correlated descriptors. Generally, the number of latent variables is limited to 0.1 times the number of observations to prevent overdetermination. The PLS method can be expressed as

$$y = a_1\mathrm{LV}_1 + a_2\mathrm{LV}_2 + \ldots\ldots + a_m\mathrm{LV}_m \quad (1)$$

where y is the dependent variable (i.e., chromatographic retention time); LV_i, the ith latent variable and a_i is the ith regression coefficient corresponding to LV_i. Each latent variable, LV_i, can be expressed as a linear combination of the independent variables x_i

$$\mathrm{LV}_i = b_1x_1 + b_2x_2 + b_3x_3 + \ldots. + b_nx_n \quad (2)$$

where x_i is the ith independent molecular variable and b_1, b_2, ..., b_n are the variables coefficients. The first latent variable accounts for most of the variance, and consecutive latent variables account for relatively smaller amounts of variance. In addition, the latent variables in a model are orthogonal to each other.

RESULTS AND DISCUSSION

This work examines the selectivity of a wide range of proteins in two stationary phase materials that led to the derivation of quantitative structure retention relationship (QSRR) models for the prediction of protein retention in these systems. In particular, linear gradient experiments were carried out for a variety of proteins in two stationary phase materials, and their retention times and molecular properties were used to derive predictive GA/PLS QSRR models.

Experiments. The proteins employed in this study were chosen for their relatively high pI, their crystal structure availability in the Protein Data Bank,[29] and their diversity of structure. If available, the pI of the proteins was obtained from the literature; otherwise, the theoretical pI was determined using EXPASY tool.[36] Linear gradient chromatography was carried out on 22 different proteins in two cation exchange stationary-phase materials (FF Sepharose and Source 15S). The average retention times for each of the proteins in the two chromatographic systems are shown in Figure 1. The proteins are presented in the order of increasing retention in the FF Sepharose material. As seen in the figure, the elution order observed in the FF Sepharose material is not the same as that observed with the Source 15S. Furthermore, the retention times are, in general, lower in the Source 15S than in the Sepharose material, with the exception of four proteins in the data set. Although Sepharose is a hydrophilic agarose based material, Source 15S is a potentially more hydrophobic divinylbenzene-based resin. These results illustrate that two cation-exchange resins possessing different backbone and linker chemistries can have significant selectivity differences. In fact, it is exactly these types of selectivity differences that often necessitate the screening of a wide variety of potential chromatographic materials for a given bioseparation problem; therefore, accurate computational screening of separation materials would facilitate the selection of proper chromatographic conditions and speed up development processes.

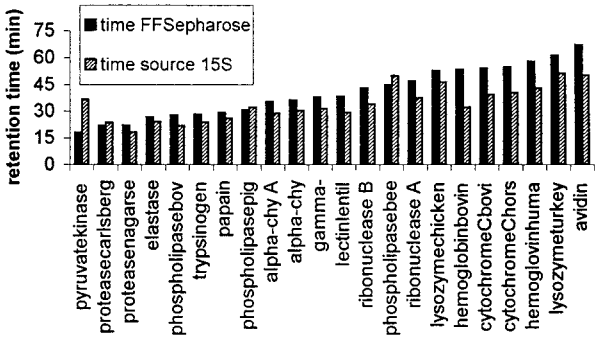

Figure 1. Plot comparing the average protein retention times in FF Sepharose and Source 15S stationary phase materials. Experimental conditions: linear gradient (buffer A, 50 mM sodium acetate, pH 5; buffer B, 50 mM sodium acetate and 600 mM sodium chloride, pH 5); linear gradient slope of 20 mM Na per column volume; 30-μL injections; 0.5 mL/min flow rate; monitored at 280 nm.

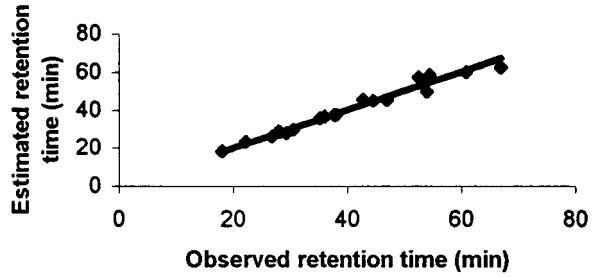

Figure 2. Plot correlating the experimental protein retention times in the training set vs their estimated values obtained by the FF Sepharose QSRR model.

Modeling. QSRR models were generated for each stationary phase as described in the Model Generation section. These models were produced using the experimental chromatographic data together with MOE and TAE molecular descriptors computed for a training set of 20 proteins. TAE electron density distributions for the proteins were reconstructed using an approximation of neutral side chains, and MOE descriptors were computed using estimated charge states based on estimated side-chain pK_a values. In each case, TAE electron-density-derived descriptor generation for all proteins required less than 5 min on an SGI Octane workstation. MOE descriptor generation required comparable CPU resources. Once suitable descriptor sets were available, GA/PLS modeling was performed to reduce the number of descriptors from 337 to 19 (FF Sepharose) and 8 (Source 15S) and to make predictions based on the best model. Plots of the predicted versus experimental retention times are presented in Figures 2 and 3. As seen in the Figures, the retention times provided by the QSRR models were in good agreement with the chromatographic experimental data. The QSRR models derived using two latent variables produced cross-validated correlation coefficients for the Sepharose and Source-15S models of 0.938 and 0.907, respectively. These results are compelling in that they demonstrate that protein retention can be accurately represented using molecular descriptors obtained from crystal structure geometries.

The descriptor sets for each model are shown in Tables 1 and 2, along with their loadings in each of the two latent variables.

(34) Livingstone, D. *Data Analysis for Chemists*; Oxford Science Publications: Oxford; 1995, Chapter 7.
(35) Wold, S. *Technometrics* **1978**, *20*, 397−405.
(36) http://www.expasy.ch/

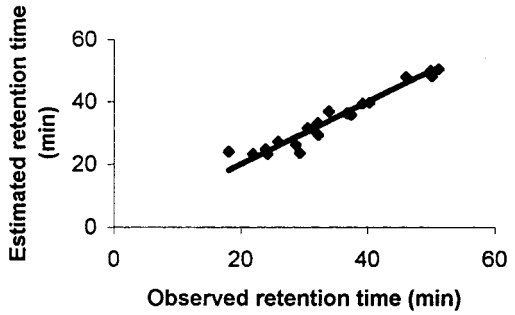

Figure 3. Plot correlating the experimental protein retention times in the training set vs the estimated values obtained by the Source 15S QSRR model.

Table 1. FF Sepharose QSRR Model Variables and Their Loadings

descriptor	dim 1	dim 2
SIKIA	0.17	−0.46
EP9	−0.21	0.18
EP10	−0.23	0.25
PIP1	−0.04	−0.29
PIP20	−0.14	−0.25
DIPOLEY	0.00	−0.09
E-OOP	−0.29	0.14
FASA-	−0.12	−0.48
FASA-H	0.04	−0.46
KIER2	−0.28	0.09
LOGP(O/W)	0.28	−0.04
PEOE-VSA3	−0.28	−0.02
PEOE-VSA5	−0.12	0.19
PEOE-VSA-5	−0.30	0.08
PEOE-VSA1	−0.30	0.08
PEOE-VSA3	−0.29	0.05
SMR-VSA4	−0.28	−0.01
SMR-VSA6	−0.29	0.07
STD-DIM2	−0.27	−0.05

Within PLS modeling, descriptor loadings provide indications of the importance of each raw descriptor used in the model as they appear within each latent variable. Although interpretation of descriptor patterns is inherently difficult, some of the most important descriptors may be directly attributed to specific types of noncovalent interactions. Such an analysis follows. As shown in Table 1, the Sepharose retention model was dominated by four terms: SIKIA, FASA-, FASA-H and PEOE-VSA-5. These descriptors are correlated with several of the electronic factors involved in the intermolecular interactions responsible for column retention. SIKIA describes the presence and strength of Bronsted basic sites, and FASA- provides an indication of the fraction of the solvent-accessible molecular surface area bearing a negative electrostatic potential. Such surface areas would be expected to correlate with Bronsted or Lewis basicity, as well as provide an indication of molecular dipolarity. FASA-H represents the fraction of solvent-accessible molecular surface area that is hydrophobic in nature and appears in the model with a negative coefficient, indicating that greater hydrophobicity inhibits retention, as expected. PEOE-VSA-5 is the amount of surface area of a molecule on which highly negative electrostatic potential is found when PEOE[32] atom-centered point charges are used in the calculation. This descriptor is expected to provide information similar to FASA-, but with a greater representation of the magnitude of the molecular basicity. This combination of descriptors and loadings indicates that Bronsted bases, such as ionizable amino groups, benefit from retention; because the descriptor values for the proteins are negative, the loading is also negative. The loading of FASA- (particularly in dimension 2) indicates that retention is hindered by molecular surface regions of negative potential, which is consistent with the idea that this negatively charged stationary-phase resin interacts poorly with negative molecular surface regions. Table 1 also shows that hydrophobicity hinders retention

Table 2. Source 15S QSRR Model Variables and Their Loadings

descriptor	dim 1	dim 2
DIPOLEX	−0.12	0.68
FASA	0.39	0.18
FASA-P	0.11	0.37
PEOE-VSA3	−0.40	0.06
PEOE-VSA5	−0.19	0.59
PEOE-VSA-FPPOS	0.48	−0.04
Q-VSA-FPPOS	0.48	−0.04
SMR-VSA4	−0.40	0.09

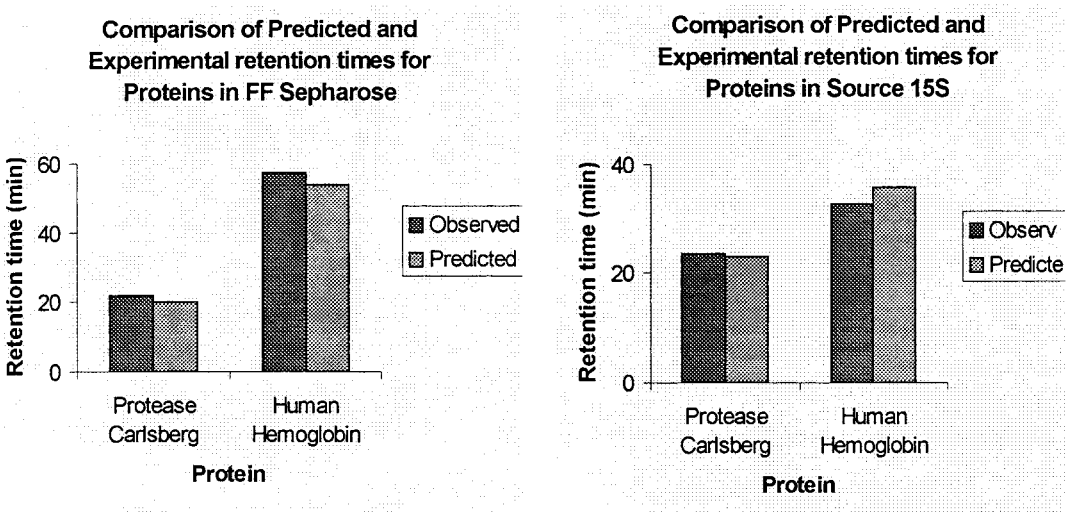

Figure 4. Plots comparing the experimental protein retention times in the test set and their predictions made by the QSRR models: (a) FF Sepharose and (b) Source 15S.

(FASA-H with a negative loading), as would be expected with such a polar stationary phase.

The interactions of the Source-15S stationary phase with the target proteins may be described by a QSRR model with three main features: polarity, hydrogen bonding and polarizability. These chemical effects are encoded into several descriptor types that are shown in Table 2. Molecular polarity is described by the DIPOLEX descriptor, which represents the component of a molecular dipole moment that is aligned with the first principal inertial axis of rotation, thus describing the charge distribution in each molecule within a common frame of reference. Polarity and dipolarity are represented by several electrostatic surface area descriptors that are also implicated in describing hydrogen-bonding behavior. These descriptors are: FASA+, FASA-P, PEOE-VSA-FPPOS, PEOE-VSA3 and PEOE-VSA5. The first three descriptors in the list relate the percentage of molecular surface area to specific electrostatic properties, and the latter two describe absolute surface areas of molecules bearing specific value ranges of Gasteiger-derived electrostatic potentials. PEOE-VSA3 and PEOE-VSA5 report two value ranges with moderate values of positive potential. The two types of surface descriptors (fractional vs absolute) are combined in the QSRR model to account for both molecular size and charge distribution effects. Positive ranges of molecular surface electrostatic potential are normally associated with Lewis acid sites or hydrogen bond donor regions. Molecular polarizability is also a significant part of the model and is described by the SMR-VSA4 index. This descriptor is obtained by summing atom-based molecular refractivity indices, as described by Crippen,[37] and provides a measure of the deformability of the molecular electron density.

The predictive power of both models was then evaluated using human hemoglobin and protease carlsberg as a test set. These proteins were not utilized in either descriptor feature selection or model generation. A comparison of the model predictions and the actual experimental results is shown in Figure 4. As seen in the figure, the models do an excellent job of predicting the retention of the proteins in the test set. This is quite remarkable in that it demonstrates that not only can these models correlate protein retention data (Figures 2 and 3), but they can also be successfully employed for the a priori prediction of protein retention (Figure 4). It is important to note that these models can encompass proteins possessing a variety of sizes, shapes, functionalities, and selectivity for these resins. The results presented here indicate that the models have been able to capture differences in the structures of the proteins as well as the various modes of interactions that they exhibited with each stationary phase material. Finally, it is interesting to note that even though protease carlsberg was one of the four proteins that had higher retention in the Source 15S material, the models were well-equipped to predict this selectivity reversal.

(37) Wildman, S. A.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

CONCLUSIONS

The results in this paper demonstrate the utility of (QSRR) models based on a genetic algorithm/partial least squares approach using chromatographic data in concert with molecular descriptors generated using protein crystal structure geometries. These models were able to correlate the retention data for a wide variety of proteins and were shown to predict selectivity reversals for proteins not included in the generation of the model. Future work in our laboratory will examine the utility of this technique for a wide variety of chromatographic systems and conditions.

ACKNOWLEDGMENT

This work has been funded in part by Amersham Pharmacia (Uppsala, Sweden) and a grant from the National Science Foundation (IIS-9979860).

Received for review July 17, 2001. Accepted August 14, 2001.

AC010797S