Separation Science and Technology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713708471

High Throughput Determination and QSER Modeling of Displacer DC-50 Values for Ion Exchange Systems

Jia Liu[a]; Ting Yang[a]; Asif Ladiwala[a]; Steven M. Cramer[a]; Curtis M. Breneman[b]
[a] Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, Troy, New York [b] Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, New York

To cite this Article Liu, Jia , Yang, Ting , Ladiwala, Asif , Cramer, Steven M. and Breneman, Curtis M.(2006) 'High Throughput Determination and QSER Modeling of Displacer DC-50 Values for Ion Exchange Systems', Separation Science and Technology, 41: 14, 3079 — 3107
To link to this Article: DOI: 10.1080/01496390600894822
URL: http://dx.doi.org/10.1080/01496390600894822

Separation Science and Technology, 41: 3079–3107, 2006

ISSN 0149-6395 print/1520-5754 online
DOI: 10.1080/01496390600894822

High Throughput Determination and QSER Modeling of Displacer DC-50 Values for Ion Exchange Systems

Jia Liu, Ting Yang, Asif Ladiwala, and Steven M. Cramer
Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, Troy, New York

Curtis M. Breneman
Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, New York

Abstract: In this paper, the displacer concentration required to displace 50% of proteins bound in batch adsorption systems, DC-50, was employed as a means of ranking high-affinity, low molecular weight displacers for ion-exchange systems. A relatively large data set of cationic displacers with varying chemistries were evaluated with two proteins on two strong cation exchange resins in parallel batch screening experiments. Using this methodology, a significant number of high affinity displacers were identified that could displace both proteins at relatively low concentrations. In addition, the DC-50 data was used in concert with molecular structural information of the displacers to produce predictive quantitative structure-efficacy relationship (QSER) models based on a support vector machine (SVM) regression approach. The resulting models were well correlated and the predictive power of these models was demonstrated. Examination of the features selected in these models provided insight into the factors influencing displacer efficacy in cation exchange systems. These results demonstrate the utility of a combined DC-50/QSER approach to identify and design high-affinity displacers for ion-exchange displacement chromatography.

Keywords: Ion exchange, displacement chromatography, high throughput screening, DC-50, QSER

Received 26 October 2005, Accepted May 2006

Address correspondence to Steven M. Cramer, Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180. Tel.: (518)276-6198; Fax: (518)276-4030; E-mail: crames@rpi.edu

INTRODUCTION

Ion-exchange displacement chromatography has attracted significant attention as a powerful technique for the purification of biomolecules in biotherapeutic downstream processes mainly due to the high product concentration, purity, and yield that this technique can produce at high column loadings (1–4). While conventionally high molecular weight displacers (MW $>$ 2000), such as polyelectrolytes were used as displacers in ion-exchange systems (4, 5), it has also been demonstrated that low molecular weight displacers (MW $<$ 2000) can be used as effective displacers for bioproduct purification (6).

Low MW displacers have significant operational advantages as compared to large polyelectrolyte displacers. If there is any overlap between the displacer and the protein of interest, these low MW materials can be readily separated from the purified protein during post-displacement processing using standard size-based purification methods (e.g. size exclusion chromatography, ultrafiltration). The salt dependent adsorption behavior of these low MW displacers greatly facilitates column regeneration. Finally, the use of low MW displacers enables the operation of displacements in the selective displacement mode which results in elution of the weakly retained proteins in the induced salt gradient, displacement of the bioproduct of interest, and closely associated impurities, and the desorption of the more strongly retained impurities after breakthrough of the displacer front. A variety of low-molecular-mass displacers have been identified including protected amino acids (7), dendrimers (8), antibiotics (9), phloroglucinol based salts (10) and aminoglycoside-polyamines (11).

Despite these advances, the design of the low molecular mass high affinity displacers for the purification of highly retained biomolecules remains a challenge. The identification of appropriate displacer candidates has relied on the determination of their dynamic affinity based on the steric mass action (SMA) parameters of potential displacers (12). Shukla et al. (13–15) synthesized pentaerythritol based displacers and assessed their efficacy on a variety of cation exchange resins with different backbone chemistries in order to study the relationship between displacers' molecular structural characteristics and their efficacy. It has been shown that those displacers, consisting of pentaerythritol-bearing trimethylammonium groups, benzene rings, and heptyl or cyclohexyl groups had different affinities and selectivities on cation-exchange stationary phases. Tugcu et al. (10) synthesized a homologous series of displacers based on either triazine or phloroglucinol. It was demonstrated that aromaticity/hydrophobicity is very important in increasing the affinity of displacers for anion-exchange resins regardless of their backbone chemistry. The results also indicated that a benzene ring is superior to a triazine ring in increasing the affinity of these anionic displacers. In addition, the data indicate that the location of an aromatic ring in the core enables the molecule to approach the stationary phase in a flat geometry, thereby increasing the number of charges interacting with the stationary phase.

While these approaches yielded important information on column performance of the potential displacers, the determination of SMA parameters is time consuming. It becomes very important to develop new high throughput assays that can quickly identify the appropriate displacer leads from a large number of displacer candidates. A lot of study has been done and a batch displacement assay was then developed as a high throughput screening (HTS) technique for the rapid identification of potential displacer molecules. In this technique, "percentage of protein displaced", i.e. the % bound protein displaced by the 10 mM displacer solution, was determined to indicate displacer affinity. Furthermore, these HTS results have been used to develop quantitative structure-efficacy relationship (QSER) models (16), which can be helpful in displacer efficacy prediction and high affinity displacer design (17, 18). Although the HTS technique based on the "percentage of protein displaced" enabled quick screening of displacers, it was soon discovered that the "percentage of protein displaced" can measure displacer affinity based on only one fixed displacer concentration, which made high affinity displacers tend to lump together in the response factor used in those works. In an effort to better distinguish the relative efficacies of various high affinity displacers, a new response factor "the initial displacer concentration required for the elution of 50% bound protein" (DC-50) was developed and used successfully for a more accurate affinity analysis of a particular synthesis high affinity low molecular weight displacer library (11).

In this paper, a large number of commercially available displacer candidates were screened in parallel and ranked according to their DC-50 values obtained for two proteins (horse cytochrome C and chicken egg lysozyme) on two different types of stationary phases (Source 15S and SP Sepharose HP). According to the definition of DC-50, the displacers with the highest affinity (lowest DC-50 values) on each stationary phase were identified. The HTS data was then used to build predictive quantitative structure-efficacy relationship (QSER) models.

THEORY

QSER Modeling and SVM Regression Models

In order to develop informative QSER models, a combination of Molecular Operating Environment (MOE) descriptors, electron density-based Transferable Atom Equivalent (TAE) quantum mechanics descriptors as well as molecular fragment (FRAG) descriptors were employed. A SVM sparse regression algorithm was applied in a feature selection mode to determine a subset of relevant molecular property descriptors from this combined set of descriptors for each of the training sets involved in the bootstrapping procedure. Prior to subjective feature selection within the sparse SVM regression algorithm, an objective feature selection process was initially employed to remove invariant

descriptors, as well as those descriptors showing a variance of greater than 4 times standard deviation (4σ) of the mean value. Subsequently, nonlinear SVM regression models were built using the most relevant descriptors as determined through the linear SVM sparse regression algorithm.

Data Set and Implementation

Experimental DC-50 values for each of the displacers were obtained by carrying out high throughput screening experiments as described in the experimental section. The structures of the displacer molecules were obtained from the supplier's website (http://www.sial.com) and were drawn in SYBYL 6.5 (Tripos Inc., St.Louis, MO), after which the energy minimization of the structures was done using the MMFF94 force field. Molecular Operating Environment (Chemical Computing Group Inc., Montreal, Canada) software was used to obtain MOE molecular descriptors. The RECON 5.2 package (developed in house by C. M. Breneman and N. Sukumar, RPI, Troy, NY, 2000) was employed for generating the TAE descriptors for the displacers. The regression analysis was carried out by using an in-house SVM program, developed independently in the Department of Mathematics at Rensselaer Polytechnic Institute (19).

MOE Descriptors

The MOE program from the Chemical Computing Group Inc. provides for the calculation of molecular properties via a "QuaSAR" descriptor module within the MOE package. The descriptors generally include 2D descriptors, using only atomic properties and connectivity information of each molecule, in addition to shape-independent 3D descriptors and some pharmacophoric descriptors. The ACD/PK_a DB package (Advanced Chemistry Development Inc., Toronto, ON, Canada) was employed to compute the pK_a values for the charge centers on each displacer molecule. The pK_a values were then used to assign the charges on the displacer molecules at the pH of the experiments. MOE descriptors were computed using the appropriate charges on the displacers at the pH of the experiment.

TAE/RECON Descriptors

Based on the AIM theory, Breneman introduced the concept of "Transferable Atom Equivalents" (TAEs) (20, 21), which are composed of atomic electron density fragments bounded by interatomic zero-flux surfaces. The RECON program reads molecular structure information from one of several file formats and then reconstructs the electronic properties of the molecular surface from those of the atomic fragments. The distributions of electronic properties on molecular surfaces may then be quantified to give a large variety of numerical QSER descriptors. The advantage of constructing a large

complicated molecule from small fragments is to reduce the computational cost significantly. The electronic property distributions of about 300 displacers may be computed in ~3s on a single-headed 1.7-GHz Linux workstation.

Support Vector Machine Modeling

The regression analysis was carried out by using an in-house SVM program (19). There are numerous benefits of using SVM modeling, including an effective avoidance of overfitting, which improves its ability to build models using large numbers of molecular descriptors with relatively few experimental results in the training set. Instead of deriving a function f(x) that has the least deviation between predicted and experimentally-observed responses for all training examples, SVMs tend to minimize the generalization error boundaries so as to achieve a higher generalization performance. This generalization error bound is composed of the training error and a regularization term that controls the complexity of the hypothesis space. Therefore, this technique helps to control the complexity of the model and tends to minimize the risk of overfitting. Using the SVM regression approach, the predicted DC-50s within a distance ε from the actual experimental responses are not penalized for being erroneous. Only those prediction points beyond ε of the real response values are considered to contain modeling errors and are included in the "loss function". In QSER studies, the magnitude of the ε is set to be roughly equivalent to the experimental error in the measurement of DC-50s.

In this work, sparse SVM modeling using a linear hypothesis with l_1-norm regularization was applied as a subjective feature selection technique, since feature selection is actually a side effect of minimizing capacity in an SVM model (22). Prior to the SVM procedure, objective feature selection is performed without reference to the response variables by eliminating overly correlated descriptors, and those with zero variance. During the feature selection process, a series of linear models (20–40 bootstraps) containing optimal weight vectors with relatively few nonzero weights are generalized. The union of all descriptors with nonzero weights from each bootstrap becomes a single feature set for the final nonlinear model construction. In order to get more robust and general predictive results in the final model, multiple QSER models with the same subset of features are generated. Therefore, instead of using a single model which can be easily and heavily affected by chance correlations, the average of all model predictions is used as our final prediction results. This kind of debiasing technique is referred to as "bagging" in the statistical analysis field (23).

Model Significance

In order to validate the models generated using these techniques and to minimize the risk of overfitting, a Y-scrambling (24) method was

employed. In this methodology, a number of parallel models are obtained based on generating the best models for the original set of descriptor vectors with randomly ordered *Y*-data (i.e., the DC-50 responses), and the significance of the real R^2/Q^2 (coefficient of determination for training/coefficient of determination for the external test points) are evaluated through comparison of the distribution of Q^2 values of the scrambled response data with the model generated using the original (real) data. Briefly, during *Y*-scrambling, the *X*-data (i.e., descriptors) for the training set are left intact, while the responses are permuted to appear in a different order. During this process, the responses remain numerically the same, but their positions are shifted between cases by random shuffling. A QSPR model is then created using each set of permuted *Y*-data, and R^2 and Q^2 parameters are computed for the training and test sets, respectively. These "permuted" values may then be compared with estimates of R^2 and Q^2 of the "real" model to get a first indication of the significance of the latter values. In general, it is expected that QSPR models for the scrambled responses will have low R^2 and Q^2 values. However, in some cases, high Q^2 (or high R^2) values may be obtained due to a chance correlation or structural redundancy of the training set (25). Therefore, this process is often repeated 50–100 times, and by establishing an equivalent number of parallel QSPR models it is possible to obtain reference distributions for the R^2/Q^2 of the model. Based on the distributions of the "permuted" R^2/Q^2, the statistical significance of the model may be quantitatively ascertained.

EXPERIMENTAL

Materials

Bulk strong cation exchange stationary phase Source 15S and HP Sepharose SP were donated by GE Healthcare (Uppsala, Sweden). Strong cation exchange stationary phase Toyopearl SP-650S was purchased from Tosoh Bioscience (Montgomeryville, PA). Horse cytochrome C and chicken egg white lysozyme, sodium phosphate (dibasic) and sodium phosphate (monobasic) were purchased from Sigma (St. Louis, MO, USA).

Potential displacer molecules: 1-(2-aminoethyl)piperidine, 1,2-diaminocyclohexane, 1,4-bis(3-aminopropyl)piperazine, 2-(2-aminoethylamino)ethanol, 2,2-dimethyl-1,3-propanediamine, 3,3-diamino-N-methyl-dipropylamine, 4,7,10-Trioxa-1,3-tridecanediamine, 5-amino-1,3,3-trimethyl cyclohexane, diethylenetriamine, N-(2-aminoethyl)-1,3-propanediamine, N,N'-bis(2-aminoethyl)-1,3-propanediamine, N,N'-bis(3-aminopropyl)-1,3-propanediamine, N-methyl-1,3-propanediamine, N,N'-diethyl-1,3-propanediamine, N,N-diethylcyclohexylamine, pentaethylenehexamine and Tris(2-aminoethly)-amine were purchased from Aldrich (Milwaukee WI, USA). Agmatine sulfate, amikacin sulfate, apramycin sulfate, bekanamycin sulfate, histamine,

L-arginine methyl ester dihydrochloride, L-lysine methyl ester dihydrochloride, neomycin sulfate, paromomycin sulfate, spermine, spermidine, and streptomycin sulfate were purchased from Sigma. 1,4,8,11-tetraazacyclotetradecane and piperazine hydrochloride were purchased from TCI America (Portland, OR). Expell SP1TM, a proprietary displacer, which has both aromatic and quaternary ammonium functionalities was donated by SACHEM (Austin, TX). Pentaerythrityl-(cyclohexyldimethylammonium iodide)$_4$ (PEDMACyMI$_4$), Pentaerythrityl-(benzyldimethylammonium chloride)$_4$ (PEDMABzCl$_4$), Pentaerythrityl-(trimethylammonium iodide)$_4$ (PETMA$_4$), and Pentaerythrityl-tetramine (PENH$_2$-HCl) were synthesized in Department of Chemistry at Rensselaer Polytechnic Institute (13–15) and their chemical structures are presented in Fig. 1.

Apparatus

HTS experiments were carried out on 96-well Multiscreen®-HV Durapore® membrane-bottomed plates (Millipore, Bedford, MA). The supernatants from the wells after equilibration with the displacer were recovered using a vacuum manifold (Millipore). The distribution of the stationary phase was carried out using an Eppendorf Repeater Plus® pipette. Supernatant analysis was carried out using a Perkin Elmer HTS 7000 plus plate reader and HTSoft® 2 software using a polystyrene 96-well plates for cytochrome C and 96-well quartz plate for lysozyme. An Eppendorf 8-channel Finpipette® (50–300 μl) pipette was used for the distribution of equal amounts of supernatant into the analysis plates.

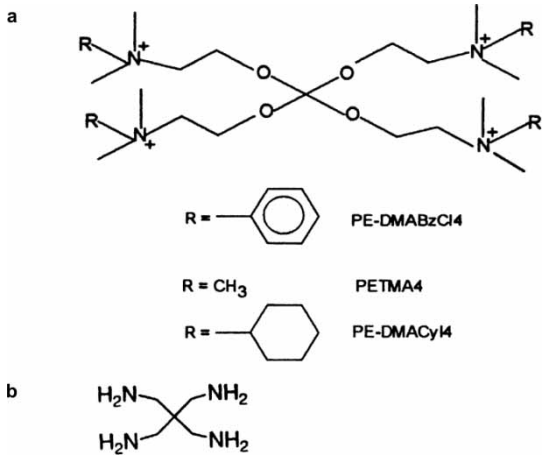


Figure 1. Chemical structures of synthesized displacers. (a) structures of PEDMABzCl$_4$, PETMA$_4$ and PEDMACyI$_4$ (b) structure of PENH$_2$-HCl.

PROCEDURE

Thirty-five displacers with different chemistries were chosen to investigate the relationship between displacer structure and efficacy in cation exchange systems using the HTS technique (11). The bulk stationary phase (HP Sepharose SP or Source 15S) was first washed twice with deionized water and then three times with the buffer, (50 mM phosphate, pH 6.0) and was allowed to equilibrate for 2 hours. After gravity settling of the stationary phase, the supernatant was removed and 3.0 ml of the remaining stationary phase slurry was equilibrated with 36 ml containing 5 mg/ml of the protein (horse-heart cytochrome C or chicken egg lysozyme) in 50 mM phosphate buffer, pH6.0, at 20°C. The protein was equilibrated with the resin for 5 hours to attain complete equilibration (11, 17, 26), during which the stationary phase was allowed to gravity settle. Upon settling, the supernatant was removed and the protein content in the supernatant was determined using absorbance detection at 340 or 280 nm, using a plate reader. The mass of the protein adsorbed on the stationary phase was determined by mass balance.

For the screening experiments, 120 μl of different initial concentrations of a displacer solution was added to 10 μl aliquots of the stationary phase slurry with bound protein. (Note that a different displacer molecule and concentration were employed for each vial to enable parallel screening.) The system was equilibrated for 5 hours. After equilibrium was achieved, the supernatant was removed and the protein content was determined by absorbance detection at 340 or 280 nm, using a plate reader. This entire procedure was carried out in parallel in 96-well membranebottomed plates to enable rapid screening of potential displacer candidates. The percent protein displaced was calculated for each aliquot based on protein mass balance and the data were plotted as a function of the initial displacer concentration for DC-50 determination as described in result and discussion section.

RESULTS AND DISCUSSION

In this paper, we used 35 displacer candidates of varying chemistries for the high throughput screening experiments. As described in the experimental section, the "percentage of protein displaced" was determined from the batch experiments and the data was employed for the determination of the displacer concentration required to displace 50% of proteins bound in batch adsorption systems, DC-50. In order to illustrate this approach, the data for three displacers are presented in Fig. 2. As seen in the figure, a curve through the % protein displaced versus displacer concentration data was employed to determine the DC-50 value. This approach was used to investigate displacer efficacy for a wide range of displacers. The DC-50 values

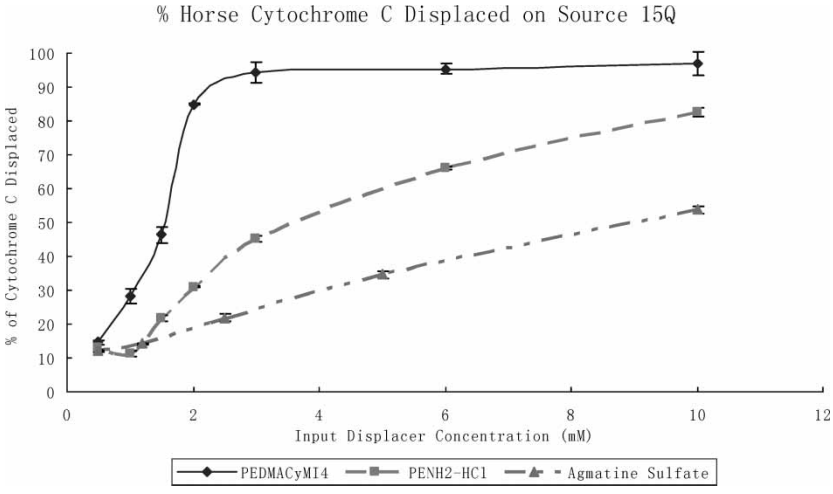


Figure 2. Determination of DC-50 values from percentage protein displaced data for Horse Cytochrome C displaced on Source 15S. The lines in the graph are for visualization only. DC-50 values: PEDMACyMI4, 1.6 mM; $PENH_2$-HCl, 3.9 mM; Agmatine Sulfate, 9.5 mM.

based on cytochrome C and lysozyme displacement were determined for each of the displacers on Source 15S and SP Sepharose HP cation-exchange resins at pH 6 and the results are presented in Table 1.

As seen in the table, the displacers examined in this study exhibited a wide range of DC-50 values. A major advantage of using the DC-50 response factor as compared to the percent protein displaced values previously employed (17) is that the current approach can distinguish between very high

Table 1. DC-50 data of cytochrome C and lysozyme on Source 15S and SP Sepharose HP

	Cytochrome C (mM)		Lysozyme (mM)	
	Source	Sepharose	Source	Sepharose
Pentaerythrityl-(trimethylammonium iodide)$_4$	1.3	1.2	3.6	2.4
Neomycin Sulfate	1.4	1.2	3.3	2.3
Pentaerythrityl-(cyclohexyldimethylammonium iodide)$_4$	1.6	1.3	N/A	N/A
Streptomycin sulfate	1.8	3.5	6.5	7.4
Spermine	1.9	1.7	6.5	4.9

(*continued*)

Table 1. Continued

	Cytochrome C (mM)		Lysozyme (mM)	
	Source	Sepharose	Source	Sepharose
Pentaerythrityl-(benzyldimethylammonium chloride)$_4$	2	1.4	4.3	4.3
SP1TM	2.1	4.1	4.5	4.8
Paromomycin sulfate	2.2	3.1	7.8	6
Amikacin sulfate	2.3	2.7	7.3	8.3
Apramycin sulfate	2.5	3.2	6.4	6
Spermidine	3	3.6	12.2	12
Bekanamycin (kanamycin sulfate B)	3.1	3.7	5.3	6
1,4-Bis(3-aminopropyl)piperazine	3.4	4.2	18.5	17.1
N,N bis(3-aminopropyl) 1,3-propanediamine	3.5	2.3	4.6	3.9
Pentaethylene hexamine	3.6	2.8	5.9	5.5
Pentaerythrityl-tetramine	3.9	3.8	21.8	22.3
3,3-diamino-N-methyl-dipropylamine	4.8	3.8	14.2	11.5
Tris(2-aminoethyl)amine	6.1	5.3	20.8	19.2
N,N bis(2-aminoethyl) 1,3-propanediamine	6.2	4.9	N/A	14.5
N-(2-Aminoethyl)-1,3-propanediamine	6.3	7.2	22.2	25.1
Agmatine sulfate	9.5	12.9	33.4	44.2
1,4,8,11-Tetraazacyclotetradecane	10.1	12.6	36.9	49.3
L-Arginine methyl ester dihydrochloride	10.2	16.2	38.4	48.2
5-Amino-1,3,3-trimethylcyclohexanemethylamine	10.3	14	47.4	N/A
L-Lysine methyl ester dihydrochloride	10.5	23.8	39.5	42.7
Diethylenetriamine	11.4	14.4	45.1	49.7
2,2-Dimethyl-1,3-propanediamine	11.8	18.3	47.1	49.5
4,7,10-Trioxa-1,13-tridecanediamine	12	19.3	N/A	N/A
Histamine	12.6	24.1	47.2	49.9
N-Methyl-1,3-propanediamine	14.2	16.7	40.7	48.2
N,N-Diethyl-1,3-propanediamine	14.3	13.8	45.3	49.6
1-(2-Aminoethyl)piperidine	14.8	18.7	N/A	N/A
2(2-Aminoethylamino)ethanol	15.8	22.1	49.8	49.8
Piperazine	21.5	23.6	N/A	N/A
1,2-Diaminocyclohexane	21.6	22.2	49.3	N/A

affinity displacers. In previous work (17), high affinity displacers all resulted in similar % protein displaced values when 10 mM displacer concentrations were employed. In contrast, when the DC-50 analysis is carried out, high affinity displacers (e.g. neomycin sulfate and paromomycin sulfate) can be readily distinguished (Table 1).

The data in Table 1 indicate that high affinity displacers may be able to be used as effective displacers at low concentrations. For example, a significant number of displacers exhibited very low DC-50 values for cytochrome C. It is expected that the use of low displacer concentrations could result in dramatic improvements in yield and purity for various applications in bioprocessing and proteomics.

The DC-50 data are also presented in Figs. 3 and 4 to examine the effects of displacer and/or resin chemistry on displacer efficacy in more detail. The DC-50 displacer data are presented in increasing order of the data on the Source 15S material. In general, dendrimeric molecules (PETMA4, PEDMACyMI4, etc.) had relatively higher affinities as compared to linear amines (N,N'-Bis(2-aminoethyl)-1,3-propanediamine, diethylenetriamine, etc.). The aminoglycosides (neomycin sulfate, paromomycin sulfate, etc.) were also found to be high affinity displacers for both proteins and resin systems. It was observed that dramatic changes in the affinities of these molecules can be brought about by relatively minor changes in the chemistry. For example, a change from the ethyl in N,N'-Bis(2-aminoethyl)-1,3-propanediamine to the propyl in N,N'-Bis(3-aminopropyl)-1,3-propanediamine resulted in a shift of the DC-50 value from 6.2 to 3.5 for cytochrome C on Source 15S.

The results also indicate that displacer affinities are often not generic and that displacers with specific structural characteristics may have increased efficacies for certain protein/resin combinations in agreement with some observations from our group in the past (17, 18). For example, some displacers showed significant changes in the affinity ranking when employed on different stationary phase materials. When comparing diethylenetriamine and L-lysine methyl ester, the DC-50 s are comparable on the Source material for displacing cytochrome C but are significantly different on the Sepharose resin for this protein. In another example, spermidine and N,N bis(3-aminopropyl) 1,3-propanediamine both have high affinities for cytochrome C displacement but spermidine is more effective on Source 15S while N,N bis(3-aminopropyl) 1,3-propanediamine is clearly more effective on SP Sepharose HP. Several other examples can be readily seen in the table where the relative DC-50 value for displacing cytochrome C is markedly different on the sepharose and source materials.

The effect of the protein on the DC-50 values was also pronounced. For example, while 1,4-Bis(3-aminopropyl)piperazine and bekanamycin had comparable DC-50 values for displacing cytochrome C on both resins, the DC-50 value for displacing lysozyme was significantly higher for 1,4-Bis(3-aminopropyl)piperazine as compared to bekanamycin, again on both resins. These

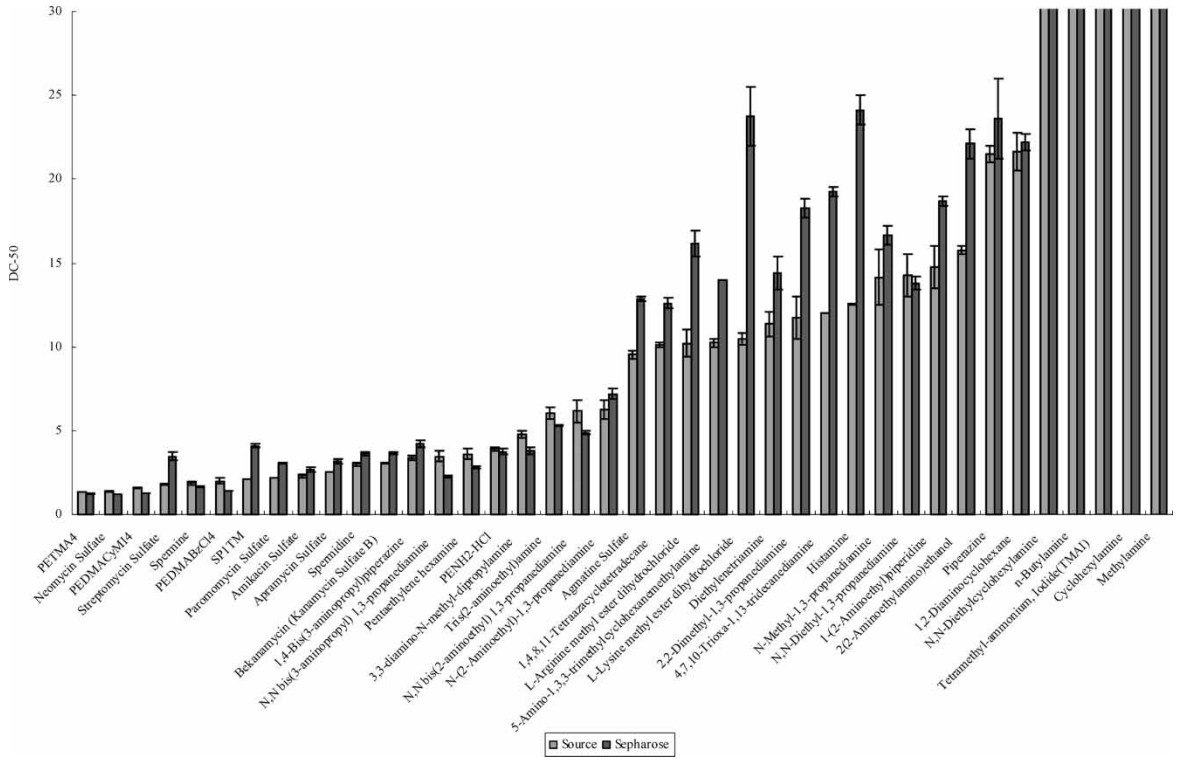


Figure 3. Comparison of cytochrome C DC-50 data on Source 15S and SP Sepharose HP. Data is presented in increasing order of the data on the Source 15S.

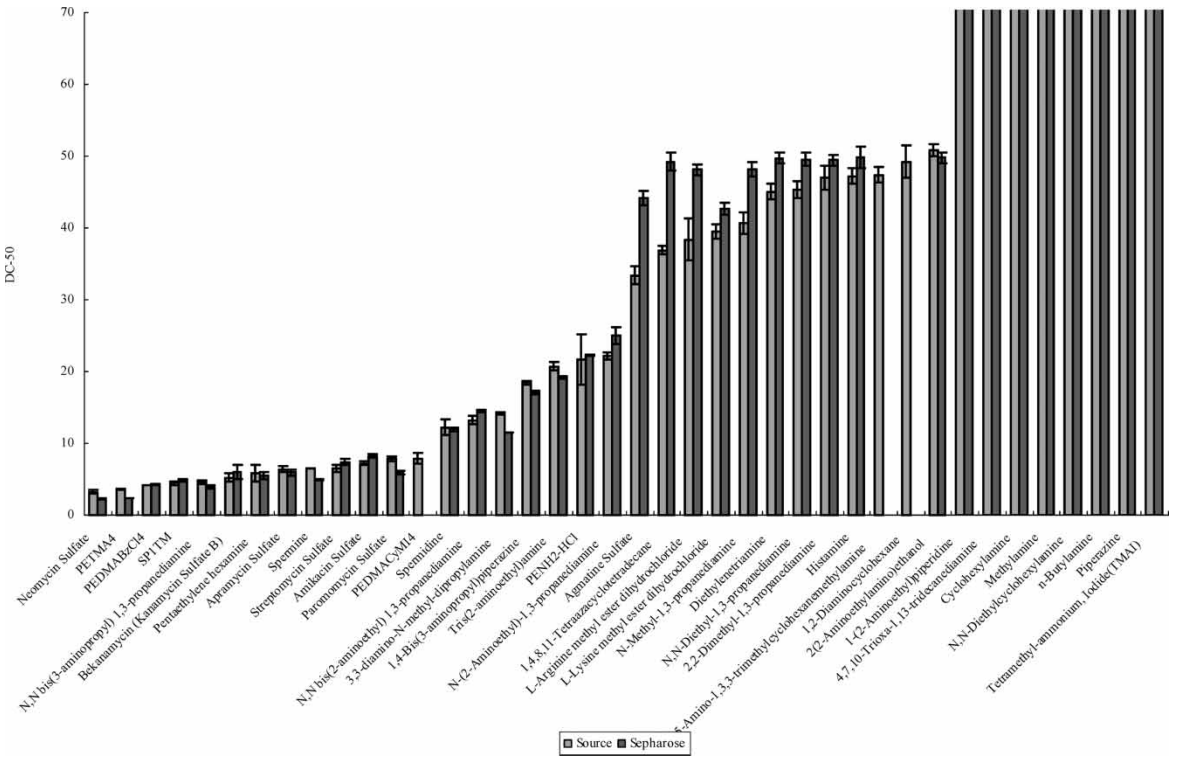


Figure 4. Comparison of lysozyme DC-50 data on Source 15S and SP Sepharose HP. Data is presented in increasing order of the data on the Source 15S.

results seem to indicate that this is a protein specific phenomenon rather than a resin dependent defect. Similar protein specific behavior can be observed for a variety of other displacers presented in the table and figures.

In order to better understand the relationship between displacers' efficacies and structures, QSER models were employed.

QSER Models and Y-Scrambling

QSER models were generated to predict the displacer efficacy for horse cytochrome C and lysozyme on two different resin materials, as well as to aid in identifying the structural properties that contribute to the efficacy of displacers in cation exchange systems. The responses employed in this study were the concentration of displacers which resulted in 50% of the adsorbed protein being displaced (DC-50). It is important to remember throughout this analysis that the lower the DC-50 value, the higher the affinity of the displacer. As described in the theory section, a wide variety of MOE, RECON, and FRAG descriptors were used to generate these models. The initial data set consisted of 4 experimental responses, 189 MOE descriptors, 208 FRAG descriptors, and 147 RECON descriptors. Descriptors having the same values for all of the displacers (invariant descriptors) as well as the descriptors showing a variance greater than 4 times standard deviation were removed from the data set as outliers. The removal of these descriptors as well as highly correlated "cousin" descriptors simplify the modeling process, reduce the risk of chance correlations, and enable interpretation of the models. The final dataset consisted of 4 responses and 341 descriptors that were used to generate 4 independent QSER models. This data set was subjected to SVM feature selection to give four independent feature sets, each corresponding to an individual response (i.e., displacement of one protein on a specific resin). Finally, the data set was divided into training and test sets for the purpose of model building, with 10% of the molecules (3 cases) in the test set and the remaining molecules in the training set. According to the principal component analysis (PCA) of the displacers' chemistry, apramycin sulfate, N,N-Diethyl-1,3-propanediamine and PEDMABzCl4 were arbitrarily chosen as the external test set to represent different displacer families in the training set as presented in Fig. 5.

Once developed, the QSER models were tested for their predictive ability for the external test set of displacers. The key descriptors in the final QSER models were then examined to determine the physicochemical phenomena that influence displacer efficacy of horse cytochrome C and lysozyme on different chromatographic resins. As part of the modeling process, graphic visualization plots (star plots) of the multiple bootstraps were generated to give a better understanding of relative descriptor importance of each descriptor within the different models. Figures 6a–6d show the correlation between the experimental and predicted results. The open circles represent the

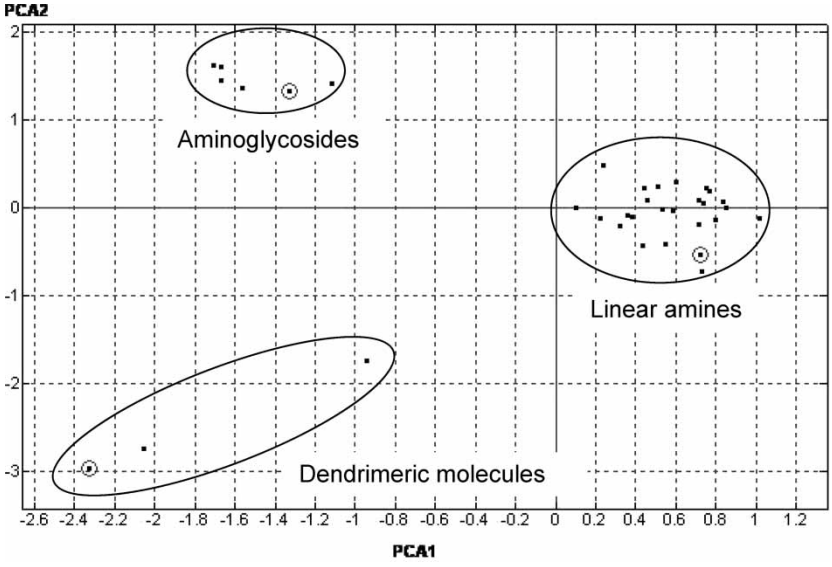


Figure 5. PCA analysis of the 35 displacers with 3 testset displacers circled.

"bagged" predictions for the training set molecules when left out in the validation set during bootstrapping procedure and the dark squares represent bagged predictions for the test set molecules. The error bars in these figures represent the standard deviation error range in the predicted DC-50 values. The prediction of test molecules is also presented in Figs. 7a–7d in a bar view fashion.

As seen in Figs. 6a-6d, the crossed-validated R^2 for four different models, cytochrome C and lysozyme on Sepharose SP and Source 15s are 0.974, 0.9649, 0.9476 and 0.9677, respectively. This indicates that the predicted DC-50 for these models are in good agreement with the experimental data. Furthermore, the DC-50 for the three test set molecules were successfully predicted by QSER models.

In addition to the external blind test set validation, *Y*-scrambling analyses were carried out as described in the theory section in order to validate the predictive nature of the QSER modeling and feature selection process. Accordingly, the *Y*-variables (i.e., the DC-50s in this study) were randomly shuffled for the displacers in the training set and an SVM model was constructed for this new dataset using the same procedure that was followed for developing the original models. This process was repeated 50 times, resulting in the generation of 50 different "scrambled" models for each response. Subsequently, the averages of the R^2 and Q^2 values of the scrambled models were compared with the R^2r/Q^2r (r stands for "real") values for the "real" model to obtain information about the validity of the model (Table 2). As seen in the table, the R^2r values were greater than R^2s

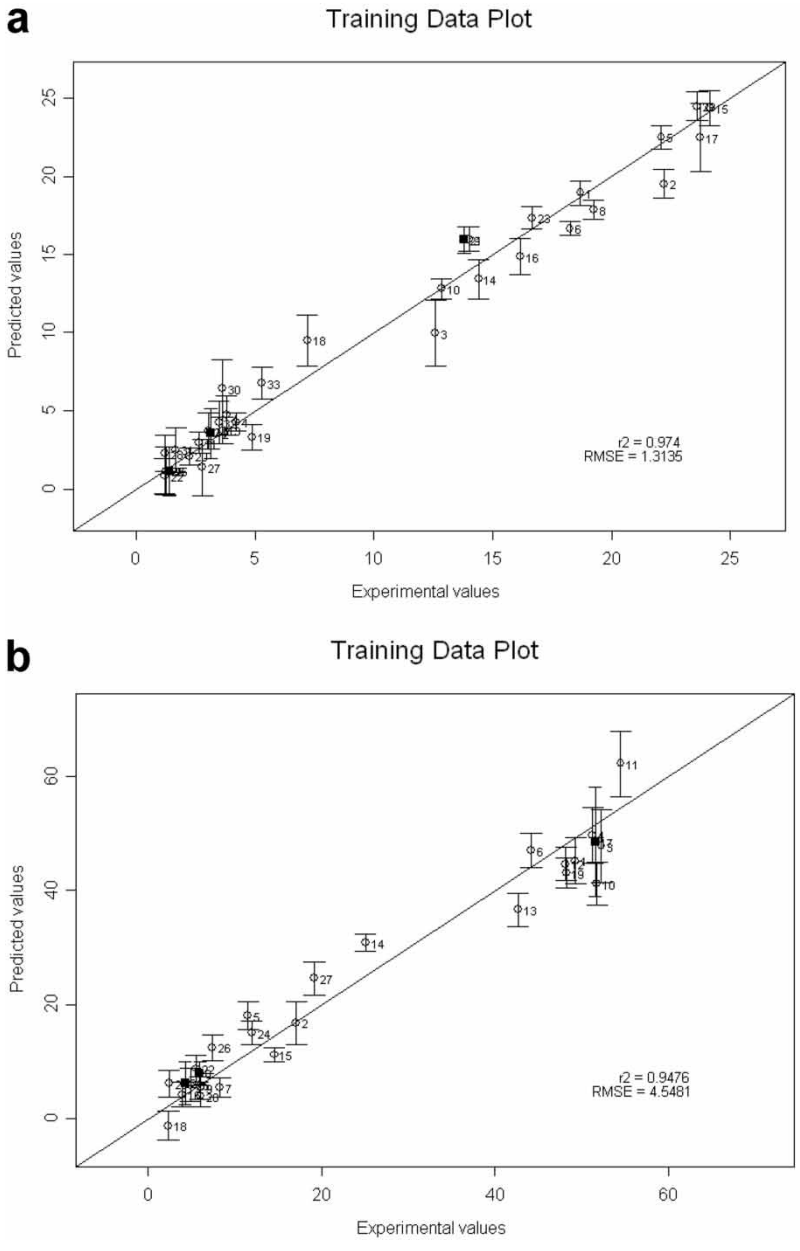


Figure 6. QSER models based on a support vector machine (SVM) regression approach for DC-50 for (a) horse cytochrome c on SP sepharose HP (b) chicken egg lysozyme on SP sepharose HP (c) horse cytochrome c on source 15S (d) chicken egg lysozyme on source 15S.

(continued)

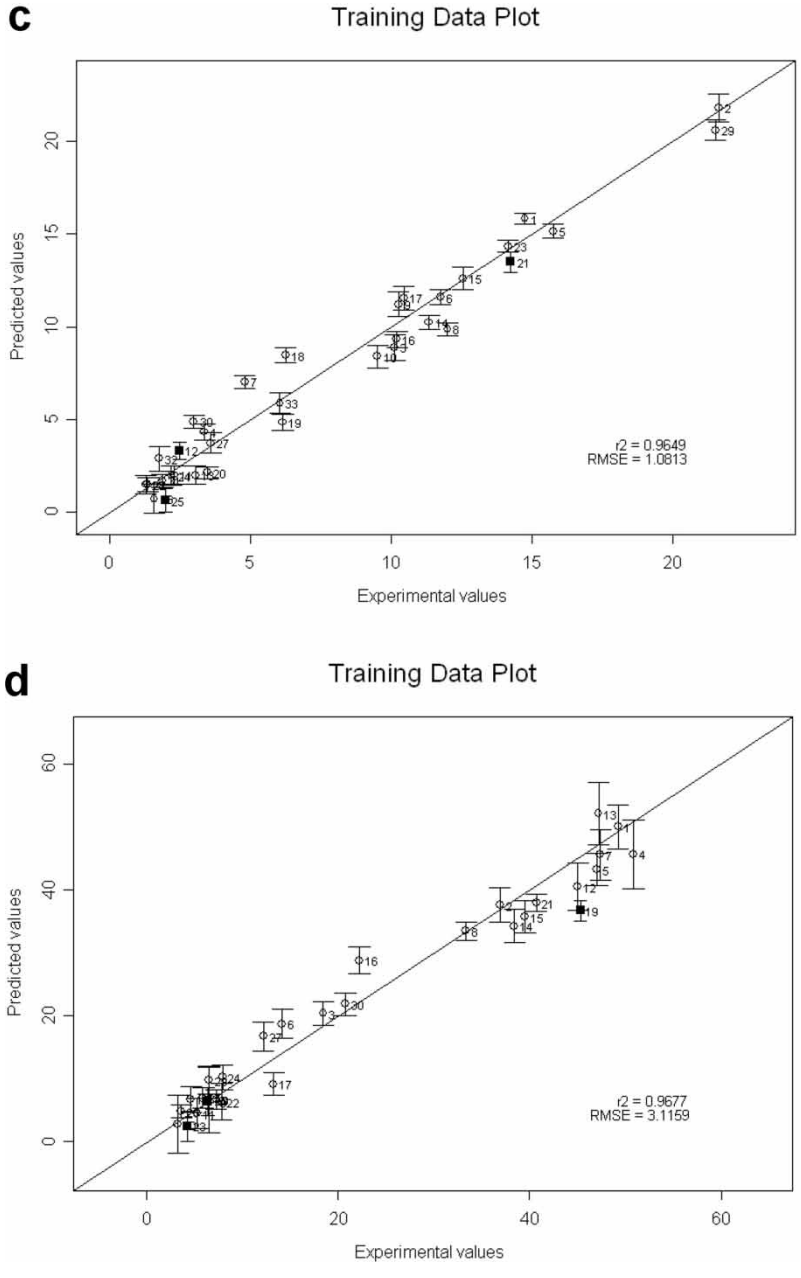


Figure 6. Continued.

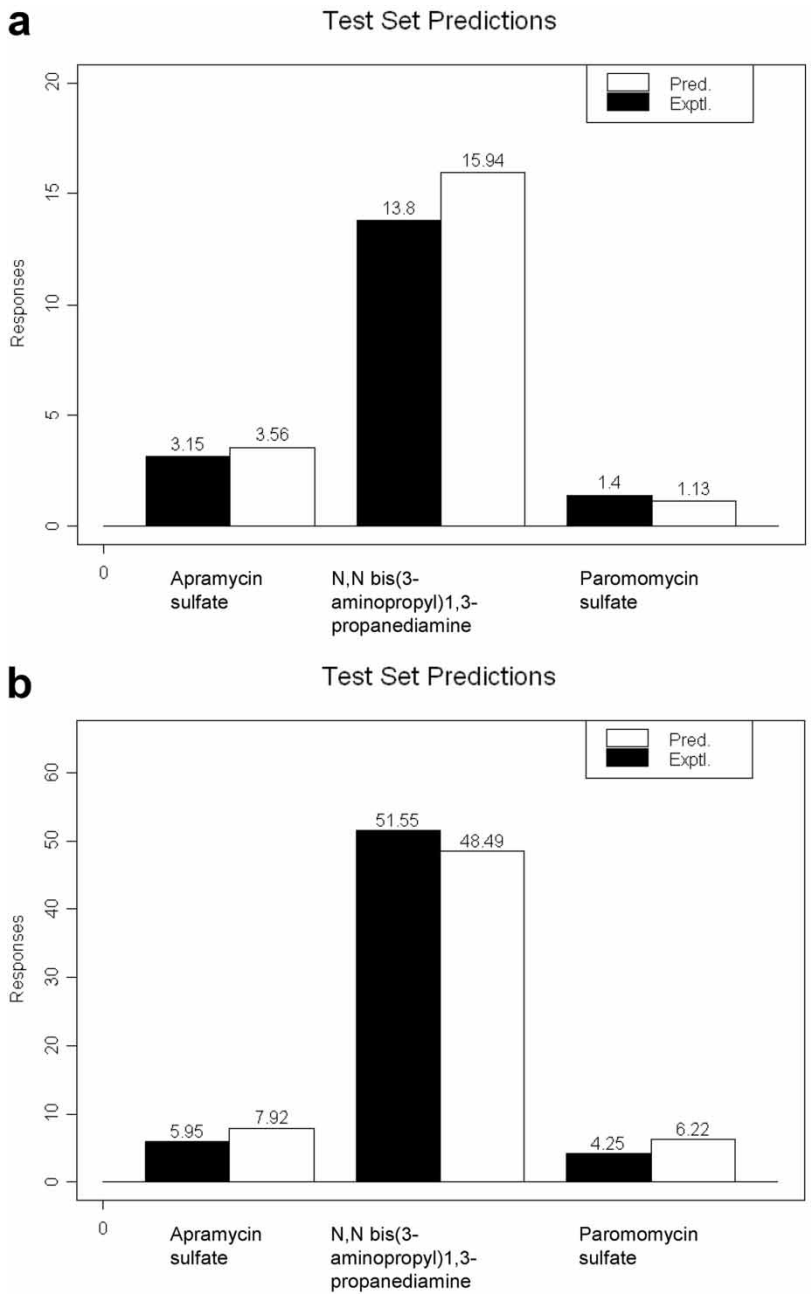


Figure 7. Prediction of the QSER models for DC-50 for external test set of molecules (a) horse cytochrome c on SP sepharose HP (b) chicken egg lysozyme on SP sepharose HP (c) horse cytochrome c on source 15S (d) chicken egg lysozyme on source15S.

(continued)

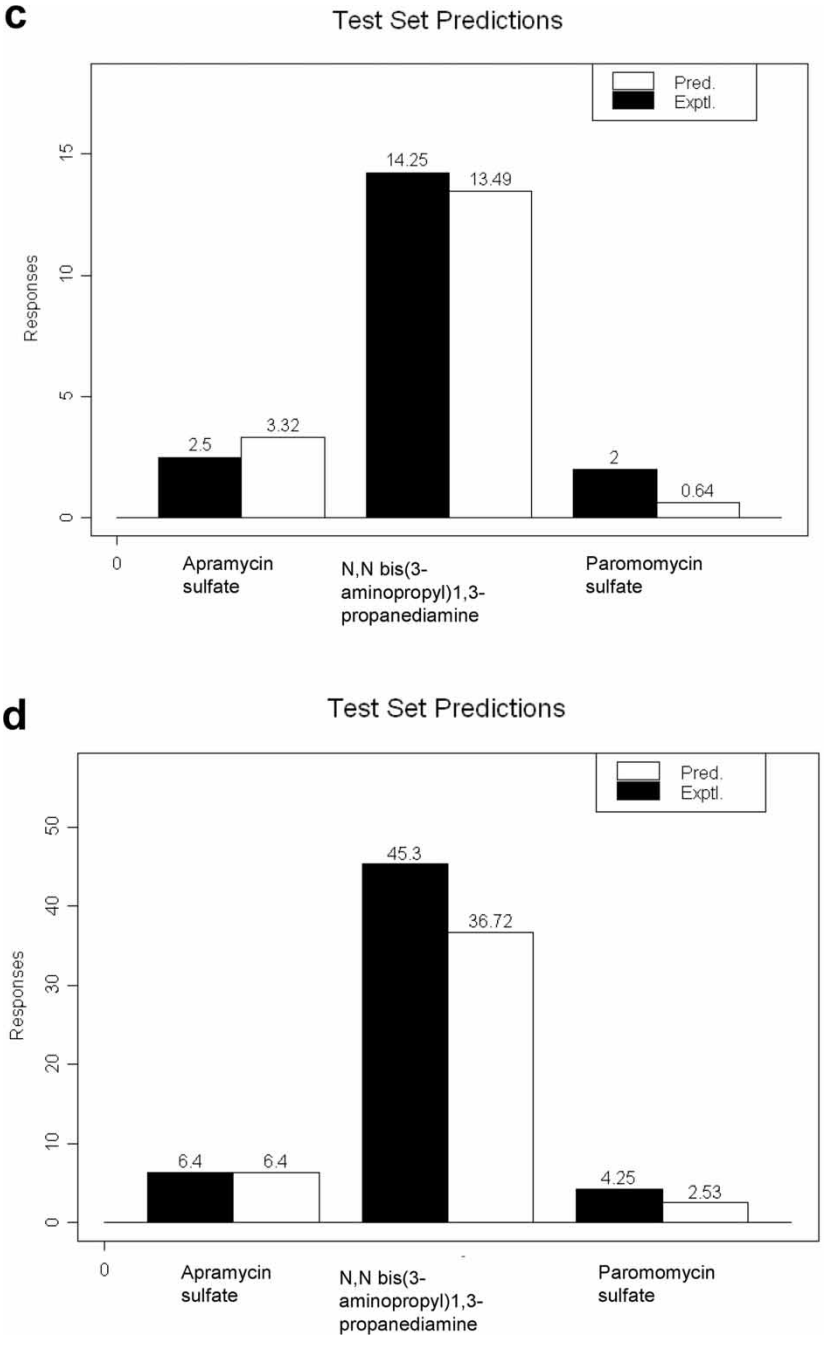


Figure 7. Continued.

Table 2. Model validation

Model	"Real" model		"Scarmbled" model		Probability
	R^2	Q^2	Avg.R^2	Avg.Q^2	$P(R^2s > R^2r)$
SEPH_CYTOCHROME C	0.974	0.94	0.62	−2.58	9.70%
SEPH_LYSOZYME	0.9476	0.98	0.66	−0.684	8.60%
SOURCE_CYTOCHROME C	0.9649	0.96	0.61	−1.18	6.10%
SOURCE_LYSOZYME	0.9677	0.92	0.64	−0.82	7.10%

(s stands for "scrambled") values by at least 0.2876 units in all cases, which clearly indicates the presence of a real "signal" in the original QSER models. At the same time, the average Q^2s values of the scrambled models were found to be significantly worse than the Q^2r value of the "real" model, demonstrating a clear deterioration of the predictive ability of the models upon scrambling.

While these QSER models are very well correlated with the experimental data and are capable of making good predictions, they also offer the opportunity to provide insight into the displacement phenomenon through direct 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 (different bootstraps) were evaluated. Each star plot was generated to evaluate the relative importance of each descriptor that was selected throughout each of the 40 bootstraps used for the creating the composite bagged model set. In these plots, each star corresponds to a specific descriptor, and the length of each ray represents the weight or importance of this descriptor in one of 40 bootstraps. For each star plot, the selected descriptors are ranked according to the sum of their ray radii for all bootstraps, so that the most significant descriptor with the highest positive weight appears in the upper left-hand corner, while the most significant negative contributor appears on the lower right-hand side. The order proceeds from left to right in a columnar fashion. The contribution of each descriptor to the model is quantified as the percentage of the total weight of the descriptor in terms of the total weight of all descriptors in the model. The star plots of these models are shown in Figs. 8a–8d. The contribution of the descriptors as identified by the star plots are listed in Table 3.

The relevant QSER descriptors selected in these models included shape, size, surface property, molecular fragment, and electron density-derived descriptors. The definitions of the most important descriptors are given in Table 4. As seen in the table, many descriptors have direct physical/

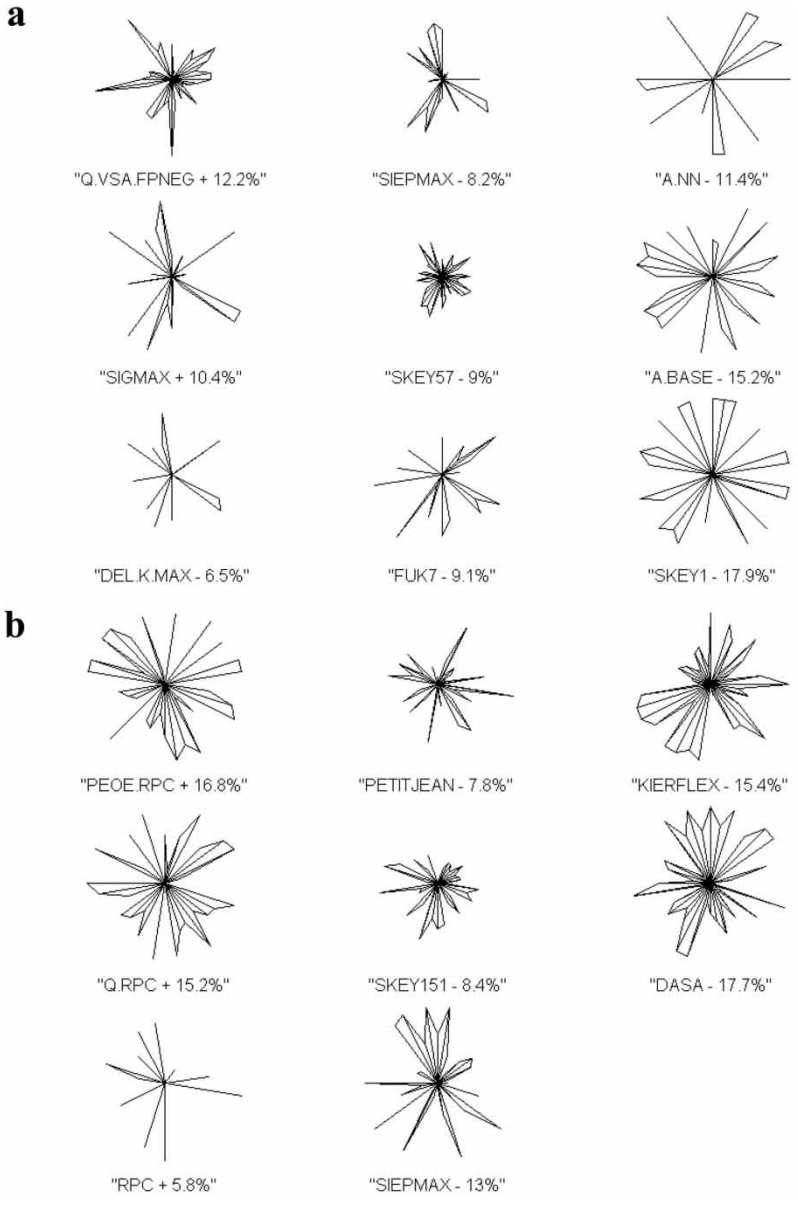


Figure 8. Star plots of important model descriptors selected in the QSER models based on the SVM regression approach for DC-50 for (a) cytochrome C on SP sepharose HP (b) lysozyme on SP sepharose HP (c) cytochrome C on source 15S (d) lysozyme on source15S.

(continued)

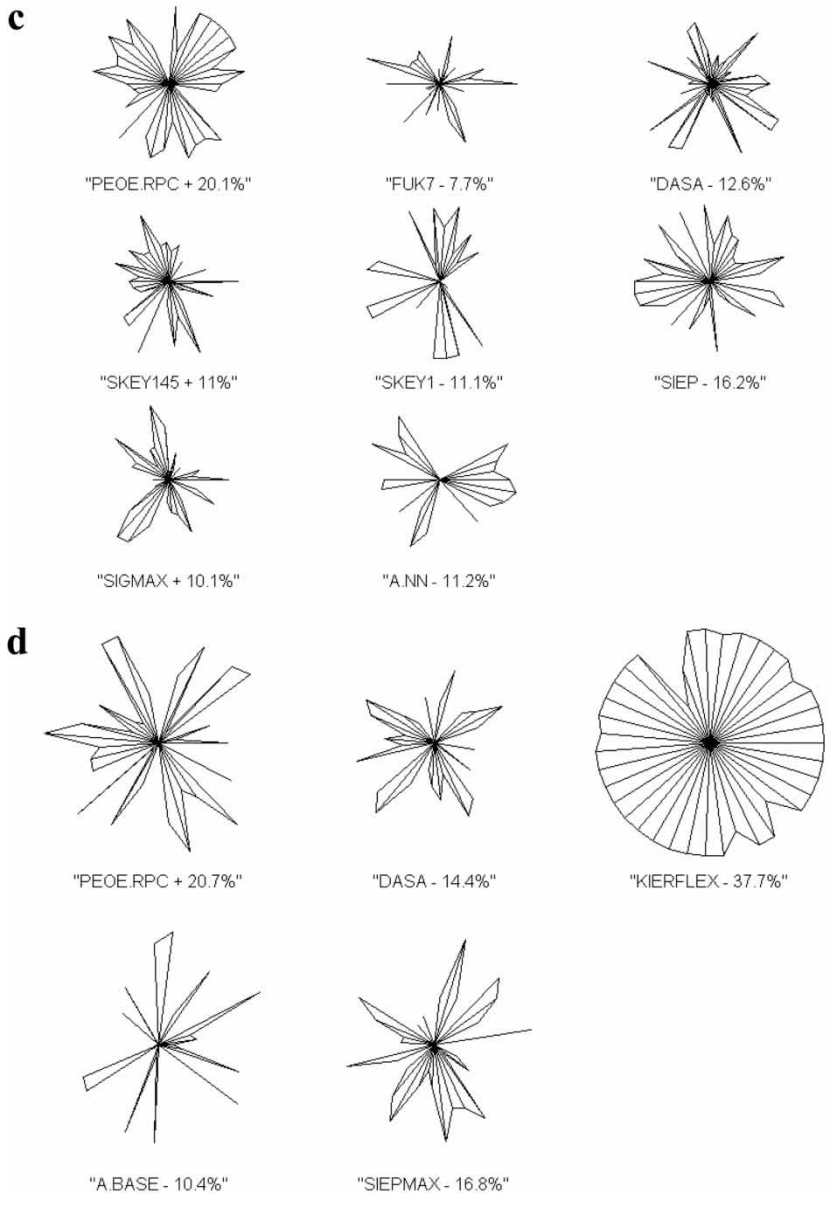


Figure 8. Continued.

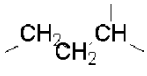
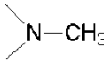
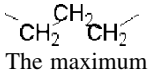
chemical significance and are relatively easy to interpret. For descriptors that are less intuitive, tools such as correlation matrices/plots and molecular surface visualization were employed to gain insight into the physicochemical information provided by these descriptors in the model.

Table 3. Countribution of the key descriptors as identified by the star plots for each model

Contribution to DC-50 of horse cytochrome C			
HP Sepharose SP		Source 15S	
Descriptor	% Contribution	Descriptor	% Contribution
Descriptors with POSITIVE contributions			
Q.VSA.FPNEG	12.2	PEOE.RPC	20.1
SIGMAX	10.4	SKEY145	11
		SIGMAX	10.1
Descriptors with NEGATIVE contributions			
SKEY1	17.9	SIEP	16.2
A.BASE	15.2	DASA	12.6
A.NN	11.4	A.NN	11.2
FUK7	9.1	SKEY1	11.1
SKEY57	9	FUK7	7.7
SIEPMAX	8.2		
DEL.K.MAX	6.5		
Contribution to DC-50 of lysozyme			
Descriptors with POSITIVE contributions			
PEOE.RPC	16.8	PEOE.RPC	20.7
Q.RPC	15.2		
RPC	5.8		
Descriptors with NEGATIVE contributions			
DASA	17.7	KIERFLEX	37.7
KIERFLEX	15.4	SIEPMAX	16.8
SIEPMAX	13	DASA	14.4
SKEY151	8.4	A.BASE	10.4
PETITJEAN	7.8		

As seen in Table 3, descriptors representing the charge distribution of displacer molecules such as PEOE.RPC, Q.RPC, and RPC were found to exist in nearly all of the models. RPC is associated with the relative positive partial charge (based on the MMFF94 force field) on the molecule and is mathematically defined as the ratio of the largest positive partial charge to the sum of all positive partial charges on the molecule. It is thus a measure of the distribution of the positive partial charge, with a higher value corresponding to a concentrated partial charge and a lower value indicating that the partial charge is more evenly distributed over the molecular surface. The positive contribution of these descriptors to the models indicates that the greater distribution of positive partial charge on the molecule surface is favorable for high displacer efficacy in cation exchange systems. Protein QSER study has also indicated that charge distribution is

Table 4. Descriptor definition

List and definition of important descriptors used for displacer efficacy modeling		
Descriptor	Type	Definition
Q.RPC+/PEOE.RPC +	MOE	Relative positive partial charge: the largest positive partial charge divided by the sum of the positive partial charge. Q_RPC+ is identical to RPC+ which has been retained for compatibility.
RPC +	MOE	Relative positive partial charge: the largest positive partial charge on an atom in the molecule divided by the sum of the positive partial charge on the molecule
Q.VSA.FPNEG	MOE	Fractional negative polar van der Waals surface area
SIGMAX	TAE	The maximum of the surface integral of the G electronic kinetic energy density (SIG) distribution
SKEY145	FRAG	
SKEY1	FRAG	N atom
SKEY57	FRAG	
A.NN	MOE	Number of nitrogen atoms
A.BASE	MOE	Number of basic atoms
SIEPMAX	TAE	The maximum of the surface integral of the electrostatic potential (EP) distribution
SIEP	TAE	Surface integral of electrostatic potential
FUK7	TAE	The 7th histogram bin of Fukui F+ function scalar value
DASA	MOE	Absolute value of the difference between water accessible area with positive partial charge and that with negative partial charge
KIERFLEX	MOE	Kier molecular flexibility index: (KierA1) (KierA2)/n [Hall 1991]
PETITJEAN	MOE	Value of (diameter − radius)/diameter
SKEY151	FRAG	
DEL.K.MAX	TAE	The maximum of the scalarized electron density gradient normal to molecular surface

desirable for increased affinity (27). The presence of Q.VSA.FPNEG in the QSER model of DC-50 for cytochrome C indicates that a negative polar surface area on the molecule is not favorable for cation exchange system which is of course to be expected. SKEY145, which has a positive contribution to the model for cytochrome C on Source 15S, enumerates the methylene and methine group spacing in the displacer structure. The displacers that have a higher occurrence of SKEY145 such as 1, 2-diaminocyclohexane, L-arginine/lysine methyl ester have relatively high DC-50 values. These displacers have relatively fewer ionized amines compared to the other displacers at the pH of the experiment (pH 6.0), and this again confirms that more charge on the molecule surface is favorable for displacement in ion exchange systems. SIGMAX, a descriptor representing the region with tightly held electrons (i.e., regions with low polarity) was found to have positive contributions. This indicates that molecules with lower polarity have fewer tendencies to interact with the stationary phase and thus the binding affinity is less strong. The descriptors associated with positive charges on the displacer molecules were found to exist in all of the models. A.BASE, the descriptor describing the number of basic atoms on the molecules was found to be important in the displacement of both cytochrome C and lysozyme. The presence and weight of this descriptor shows that the molecules with more positive charges have a higher tendency to bind to the resin, as would be expected in a cation exchange system. A.NN, SKEY1 as well as SKEY57 descriptors, associated with the number of nitrogens on the molecules, were found to have negative contributions in these models (since the lower the DC-50, the higher the affinity of the displacer). For the molecules included in our study, most of the partial positive charges were found on the amine groups, therefore all of A.NN, SKEY1 and SKEY57 essentially account of the number of positive charges present. The RECON-based SIEP/SIEPMAX descriptors represent the surface area of the molecule having a high electrostatic potential, i.e., regions with high positive partial charge. The negative contributions of these descriptors suggest that as the fraction of the polar positive surface area increases, the displacer efficacy increases. Again, this is consistent with the expected characteristics of cation-exchange chromatography where positively charged solutes are more highly retained. The DASA descriptor represents the difference between the positively and negatively charged surface areas of the displacer molecule. Since all of our displacers had a net positive charge at the pH of the batch screening experiments, this descriptor is associated with the relative abundance of positive charges on the displacer molecules. This descriptor was found to exist in displacement models for both cytochrome C and lysozyme, indicating that the net positive charge is again desirable for displacement. KIERFLEX and FUK7, descriptors correlated highly with DIAMETER (0.882) and WEIGHT (0.959) respectively (Fig. 9), were found to have negative contributions to these QSER models. This result suggests that the more complex descriptors were acting as surrogates for

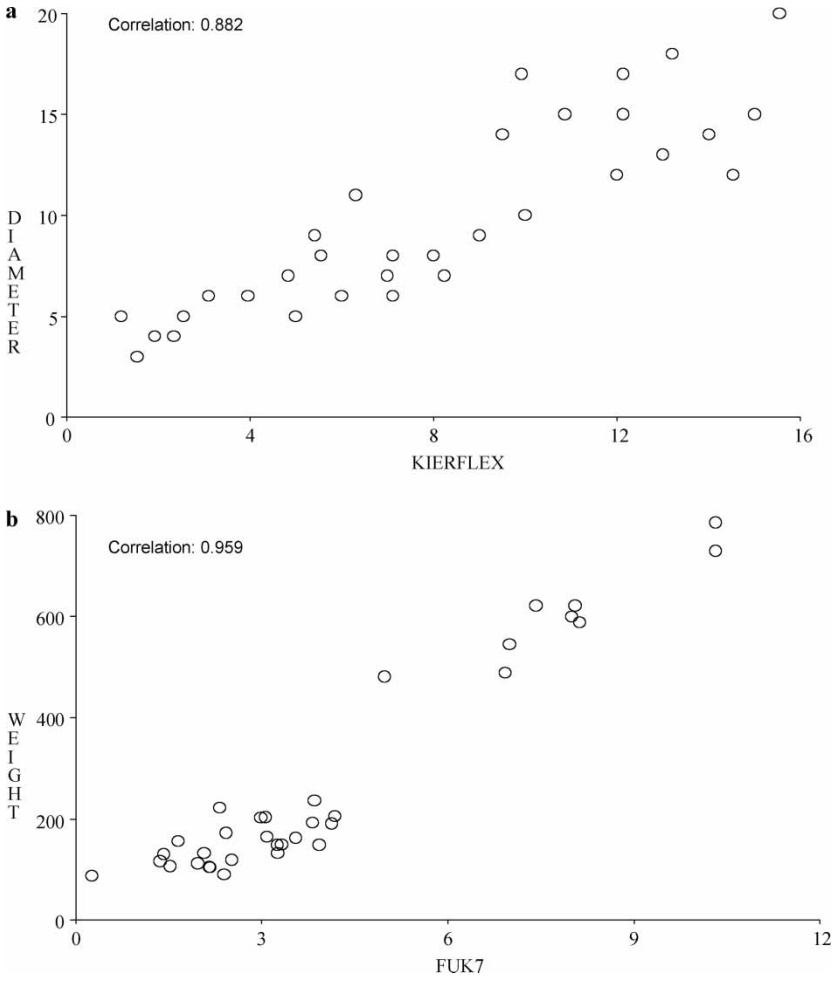


Figure 9. Correlation plots of (a) DIAMETER vs. KIERFLEX; and (b) WEIGHT vs. FUK7.

size effects—and this suggests that increasing the size of the molecules increases the displacer efficacy in cation exchange systems. PETITJEAN, another size/shape descriptor, has a negative contribution to the model of lysozyme. Based on the definition of this descriptor, the negative contribution in the model suggests that "flatter" displacer molecules will exhibit greater efficacy in displacing the proteins. Intuitively, this makes sense because a planar conformation of the displacer molecule will enable easy access of the charged groups on the displacer to the resin surface. Finally, SKEY151 describes a three-methylene spacing in a molecule. This result is consistent with previous observations in our laboratory that displacers such as

spermidine and spermine with three and four nitrogen atoms, respectively, are more effective displacers than those containing bridges of other sizes, such as diethylene triamine (13, 16).

CONCLUSION

In this paper, a modified high-throughput screening assay was developed to investigate displacers' efficacies over a wide range of concentrations. The resulting data were then successfully used to build QSER model for more detailed study of the relationship between displacer structure and efficacy. The screening results indicate that dendrimeric molecules and aminoglycosides have normally high and moderate affinities. Polyamines can have dramatic affinity differences with different structures. The utility of a synergistic DC-50/QSER methodology presented in this work was demonstrated to be a powerful tool for the identification and design of high-affinity displacers for ion-exchange displacement chromatography.

REFERENCES

1. Horvath, C., Nahum, A., and Frenz, J.H. (1981) High-performance displacement chromatography. *J. Chromatogr.*, 218: 365–393.
2. Kalasz, H. and Horvath, C. (1981) Preparative-scale separation of polymyxins with an analytical high-performance liquid chromatography system by using displacement chromatography. *J. Chromatogr.*, 215: 295–302.
3. Vigh, G., Varga-Puchony, Z., Szepesi, G., and Gazdag, M. (1987) Semi-preparative high-performance reversed-phase displacement chromatography of insulins. *J. Chromatogr.*, 386: 353–362.
4. Jayaraman, G., Gadam, S.D., and Cramer, S.M. (1993) Ion-exchange displacement chromatography of proteins: Dextran-based polyelectrolytes as high affinity displacers. *J. Chromatogr. A*, 630 (1–2): 53–68.
5. Torres, A.R. and Peterson, E.A. (1983) Ion-exchange displacement chromatography of proteins, using narrow-range carboxymethyldextrans and a new index of affinity. *Anal. Biochem.*, 130 (1): 271–282.
6. Tugcu, N., Deshmukh, R.R., Sanghvi, Y.S., Moore, J.A., and Cramer, S.M. (2001) Purification of an oligonucleotide at high column loading by high affinity, low-molecular-mass displacers. *J. Chromatogr. A*, 923 (1–2): 65–73.
7. Kundu, A., Vunnum, S., Jayaraman, G., and Cramer, S.M. (1995) Protected amino acids as novel low-molecular-weight displacers in cation-exchange displacement chromatography. *Biotechnol. Bioeng.*, 48 (5): 452–460.
8. Jayaraman, G., Li, Y., Moore, J.A., and Cramer, S.M. (1995) Ion-exchange displacement chromatography of proteins dendritic polymers as novel displacers. *J. Chromatogr. A*, 702 (1–2): 143–155.
9. Kundu, A., Vunnum, S., and Cramer, S.M. (1995) Antibiotics as low-molecular-mass displacers in ion-exchange displacement chromatography. *J. Chromatogr. A*, 707 (1): 57–67.

10. Tugcu, N., Park, S.K., Moore, J.A., and Cramer, S.M. (2002) Synthesis and characterization of high-affinity, low-molecular-mass displacers for anion-exchange chromatography. *Ind. Eng. Chem. Res.*, 41: 6482–6492.
11. Rege, K., Hu, S., Moore, J.A., Dordick, J.A., and Cramer, S.M. (2004) Chemoenzymatic synthesis and high-throughput screening of an aminoglycoside-polyamine library: identification of high-affinity displacers and DNA-binding ligands. *J. Am. Chem. Soc.*, 126 (39): 12306–12315.
12. Brooks, C.A. and Cramer, S.M. (1992) Steric mass-action ion exchange: displacement profiles and induced salt gradients. *AIChE J.*, 38 (12): 1969–1978.
13. Shukla, A.A., Bae, S.S., Moore, J.A., and Cramer, S.M. (1998) Structural characteristics of low-molecular-mass dispalcers for cation-exchange chromatography II. Role of the stationary phase. *J. Chromatogr. A*, 827: 295–310.
14. Shukla, A.A., Barnthouse, K.A., Bae, S.S., Moore, J.A., and Cramer, S.M. (1998) Synthesis and characterization of high-affinity, low molecular weight dispalcers for cation-exchange chromatography. *Ind. Eng. Chem. Res.*, 37: 4090.
15. Shukla, A.A., Barnthouse, K.A., Bae, S.S., Moore, J.A., and Cramer, S.M. (1998) Structural characteristics of low-molecular-mass dispalcers for cation-exchange chromatography. *J. Chromatogr. A*, 814: 83–95.
16. Mazza, C.B., Rege, K., Breneman, C.M., Sukumar, N., Dordick, J.S., and Cramer, S.M. (2002) High throughput screening and quantitative structure-efficacy relationship models of potential displacer molecules for ion-exchange systems. *Biotechnol. Bioeng.*, 80 (1): 60–73.
17. Rege, K., Ladiwala, A., Tugcu, N., Breneman, C.M., and Cramer, S.M. (2004) Parallel screening of selective and high-affinity displacers for proteins in ion-exchange systems. *J. Chromatogr. A*, 1033: 19–28.
18. Tugcu, N., Ladiwala, A., Breneman, C.M., and Cramer, S.M. (2003) Identification of chemically selective displacers using parallel batch screening experiments and quantitative structure efficacy relationship models. *Anal. Chem.*, 75 (21): 5806–5816.
19. Bi, J., Bennett, K., Embrechts, M., Breneman, C.M., and Song, M. (2003) *J. Mach. Learn. Res.*, 3: 1229–1243.
20. Breneman, C.M., Thompson, T.R., Rhem, M., and Dung, M. (1995) Electron-density modeling of large systems using the transferable atom equivalent method. *Comput. Chem.*, 19: 161–179.
21. Breneman, C.M. (1991) Transferable atom equivalents. Molecular electrostatic potentials from the electric multipoles of PROAIMS atomic basins. In *The Application of Charge Density Research to Chemistry and Drug Design*; Plenum Press, 357–358.
22. Bennett, K.P., Bi, J., Embrechts, M., Breneman, C., and Song, M. (2002) Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.*, (Special Issue on Feature Selection): (In press).
23. Breiman, L. (1996) *Machine Learning*, Vol. 24: 123–140.
24. Eriksson, L.J., Jaworska, A.P., Worth, M.T.D., Cronin, R., McDowell, M., and Gramatica, P. (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.*, 111: 1361–1375.
25. Tropsha, A.P., Gramatica, V.K., and Qsar, G. (2003) The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Comb. Sci.*, 22: 69–77.
26. Rege, K., Ladiwala, A., and Cramer, S.M. (2005) Multidimensional high-throughput screening of displacers. *Anal. Chem.*, 77 (21): 6818–6827.

27. Ladiwala, A., Rege, K., Breneman, C.M., and Cramer, S.M. (2003) Investigation of mobile phase salt type effects on protein retention and selectivity in cation-exchange systems using quantitative structure retention relationship models. *Langmuir*, 19: 8443–8454.