Using Surrogate Modeling in the Prediction of Fibrinogen Adsorption onto Polymer Surfaces

Jack R. Smith,*,† Doyle Knight,‡ Joachim Kohn,† Khaled Rasheed,§ Norbert Weber,†
Vladyslav Kholodovych,‖ and William J. Welsh‖

Department of Chemistry and Chemical Biology and the New Jersey Center for Biomaterial, Rutgers, The State University of New Jersey, New Brunswick, New Jersey 08854, Department of Mechanical and Aerospace Engineering, Rutgers, The State University of New Jersey, New Brunswick, New Jersey 08903, Department of Computer Science, The University of Georgia, Athens, Georgia 30602, and Department of Pharmacology, University of Medicine & Dentistry of New Jersey (UMDNJ), Robert Wood Johnson Medical School and the Informatics Institute of UMDNJ, Piscataway, New Jersey 08854

Received January 6, 2004

We present a Surrogate (semiempirical) Model for prediction of protein adsorption onto the surfaces of biodegradable polymers that have been designed for tissue engineering applications. The protein used in these studies, fibrinogen, is known to play a key role in blood clotting. Therefore, fibrinogen adsorption dictates the performance of implants exposed to blood. The Surrogate Model combines molecular modeling, machine learning and an Artificial Neural Network. This novel approach includes an accounting for experimental error using a Monte Carlo analysis. Briefly, measurements of human fibrinogen adsorption were obtained for 45 polymers. A total of 106 molecular descriptors were generated for each polymer. Of these, 102 descriptors were computed using the Molecular Operating Environment (MOE) software based upon the polymer chemical structures, two represented different monomer types, and two were measured experimentally. The Surrogate Model was developed in two stages. In the first stage, the three descriptors with the highest correlation to adsorption were determined by calculating the information gain of each descriptor. Here a Monte Carlo approach enabled a direct assessment of the effect of the experimental uncertainty on the results. The three highest-ranking descriptors, defined as those with the highest information gain for the sample set, were then selected as the input variables for the second stage, an Artificial Neural Network (ANN) to predict fibrinogen adsorption. The ANN was trained using one-half of the experimental data set (the training set) selected at random. The effect of experimental error on predictive capability was again explored using a Monte Carlo analysis. The accuracy of the ANN was assessed by comparison of the predicted values for fibrinogen adsorption with the experimental data for the remaining polymers (the validation set). The mean value of the Pearson correlation coefficient for the validation data sets was 0.54 ± 0.12. The average root-mean-square (relative) error in prediction for the validation data sets is 38%. This is an order of magnitude less than the range of experimental values (i.e., 366%) and compares favorably with the average percent relative standard deviation of the experimental measurements (i.e., 17.9%). The effects of each of the user-defined parameters in the ANN were explored. None were observed to have a significant effect on the results. Thus, the Surrogate Model can be used to accurately and unambiguously identify polymers whose fibrinogen absorption is at the limits of the range (i.e., low or high) which is an essential requirement for assessing polymers for regenerative tissue applications.

1. INTRODUCTION

Protein adsorption to surfaces is critical in cell attachment and, therefore, is a major determinant of the suitability of materials for medical implant applications, especially those involving tissue regeneration.[1] In particular, the protein fibrinogen is known to participate in processes leading to blood clotting.[2] Thus, fibrinogen adsorption can have a major effect on the performance of implant materials. Though protein adsorption can be studied using a variety of methods, most in vivo/vitro experimental methodologies do not provide details concerning the molecular-scale properties of biomaterials relevant to protein/surface interactions. Moreover, first principles calculations are currently computationally intractable. Therefore, the use of semiempirical models (Surrogate Models) in conjunction with experiment is both a necessary and timely approach.[3,4] Such models are invaluable in the design of medical devices whose functions depend on controlling cell-material interactions at the device surface. For many implant applications such as tissue engineered skin,[5] extracorporeal circulation devices for cardiac surgery,[6] and antibacterial adhesion materials,[7] more candidate materi-

* Corresponding author phone: (732)445-4351; fax: (732)445-5006, e-mail: jasmith@soemail.rutgers.edu. Corresponding author address: Department of Chemistry and Chemical Biology, 610 Taylor Rd, Piscataway, NJ 08854-8087.

† Department of Chemistry and Chemical Biology and the New Jersey Center for Biomaterial, Rutgers, The State University of New Jersey.

‡ Department of Mechanical and Aerospace Engineering, Rutgers, The State University of New Jersey.

§ The University of Georgia.

‖ Robert Wood Johnson Medical School and the Informatics Institute of UMDNJ.

10.1021/ci0499774 CCC: $27.50
Published on Web 05/07/2004

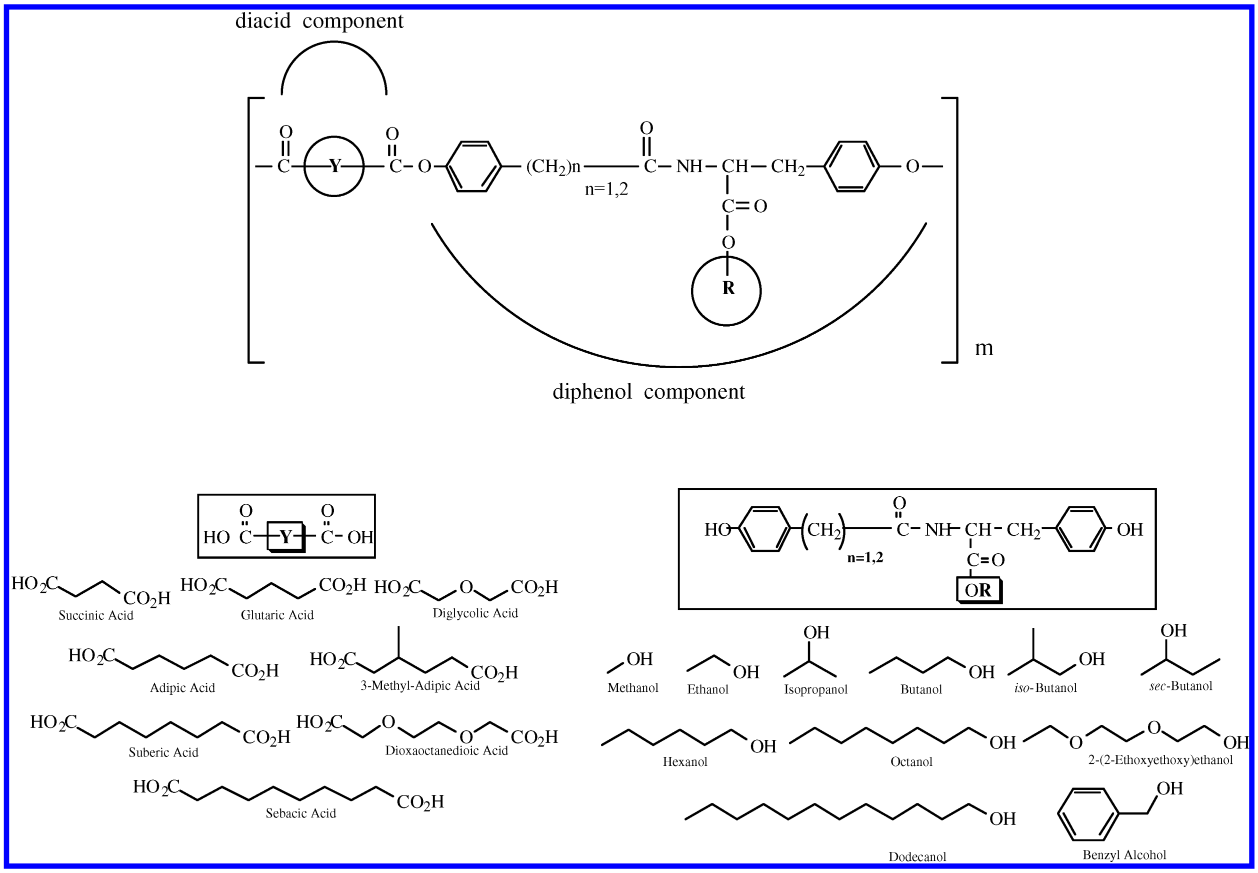


Figure 1. Library of 112 polyarylates is synthesized using 14 tyrosine-derived diphenols and 8 diacids. Polymers are strictly alternating copolymers consisting of a diacid (DA) and a diphenol (DP) component varied at Y and R, respectively. The number of methyl groups in the DP component is also variable.

als are available than can reasonably be tested in detail. Computational modeling of relevant aspects of the cell-material interactions will, therefore, be a valuable tool in making rational rather than intuitive selections of candidate materials for detailed studies.

We present a Surrogate Model for the prediction of fibrinogen adsorption on polymer surfaces for a library of tyrosine-derived biodegradable polymers (polyarylates). Previous studies have shown that small changes in chemical composition of a substratum can have a significant effect on protein adsorption and cell growth.[8,9] We explored these changes systematically using a library of 45 structurally related polyarylates[10] to develop semiempirical models capable of predicting fibrinogen adsorption as a function of chemical composition and polymer physical properties. The techniques used are well established: Artificial Neural Network (ANN) models based on multivariate statistical analysis and nonlinear regression[11] in conjunction with molecular modeling and machine learning methodology. The model inputs are quantifications of molecular scale structural properties (descriptors) of the polymers, and the output of the ANN is a quantitative prediction of the amount of fibrinogen which should adsorb to the polymer surface given these molecular scale aspects.

2. METHODOLOGY/BACKGROUND

Polyarylate Library and Sample Preparation. The polyarylate library has been developed by Fiordeliso et al.[12] to allow the systematic exploration of the effects of polymer structural and chemical properties on protein adsorption and cell proliferation/growth. Experimental studies of polyarylates have shown that several laboratory-measurable quantities correlate to cell proliferation[13] and to fibrinogen adsorption,[14] though the complexity of these correlations necessitates further analysis. Despite the fact that these polymers are structurally related, they show an impressive and reproducible variation in fibrinogen adsorption of over 360%. This made the polyarylate library ideal for our purposes, i.e., calibrating and testing a model to predict fibrinogen adsorption based on polymer structure.

Briefly, the total library of 112 polyarylates is synthesized using 14 tyrosine-derived diphenol (DP) and 8 diacid (DA) components. All polyarylates are strictly alternating copolymers. Note that the polyarylates all have a similar basic structure (Figure 1), apart from modifications at Y and R and at one of the CH_2 groups on the backbone of the DP component. The composition of the polymer backbone at Y and the pendent chain at R depend on the acid and alcohol precursors (Figure 1), respectively. Eleven DP components were synthesized with two CH_2 groups in the variable portion of the DP backbone ("n" in the figure is equal to 2). These were labeled according to their precursor alcohols: methyl, ethyl, butyl, hexyl, octyl, 2-(2-ethoxyethoxy)ethyl, dodecyl, isopropyl, isobutyl, *sec*-butyl, and benzyl. Three DP components were synthesized with one CH_2 group in the variable portion of the DP backbone ($n = 1$). These are named for their acid precursors: R = ethyl, hexyl, and octyl. DA component variations included the following: succinate, glutarate, diglycolate, adipate, 3-methyl-adipate, suberate, dioxaoctanedioate, and sebacate.

Table 1. Polymer Nomenclature and Numbering Scheme[a]

no.	pendent	diacid	no.	pendent	diacid
1	DTiB	sebacate	24	DTB	glutarate
2	HTH	sebacate	25	HTE	adipate
3	DTO	glutarate	26	DTsB	adipate
4	DTO	sebacate	27	DTM	methyl adipate
5	HTH	adipate	28	DTB	adipate
6	HTH	suberate	29	DTB	succinate
7	DTO	adipate	30	DTE	adipate
8	DTBn	sebacate	31	DTH	succinate
9	DTO	suberate	32	DTsB	glutarate
10	DTH	adipate	33	DTBn	methyl adipate
11	DTiB	adipate	34	DTM	suberate
12	DTH	suberate	35	DTBn	adipate
13	DTBn	suberate	36	DTH	diglycolate
14	DTH	methyl adipate	37	DTO	diglycolate
15	DTH	glutarate	38	DTM	adipate
16	DTM	sebacate	39	DTiP	methyl adipate
17	DTB	suberate	40	HTE	methyl adipate
18	HTE	suberate	41	DTE	glutarate
19	DTsB	suberate	42	DTsB	methyl adipate
20	DTiB	succinate	43	DTE	methyl adipate
21	DTiP	adipate	44	HTE	succinate
22	DTO	succinate	45	DTB	diglycolate
23	DTB	methyl adipate			

[a] Note polymers are numbered in order of increasing (measured) fibrinogen adsorption.

Diphenols are named according to the following abbreviations. DTR stands for “desaminotyrosyl tyrosine alkyl ester,” indicating two CH_2 groups in the backbone. R is the pendent chain which takes on the different identities indicated in Figure 1. Similarly, HTR stands for “hydroxyacetic acid-tyrosine alkyl ester,” indicating one CH_2 group in the backbone. The 45 polymers used in this experiment are named according to the convention in Table 1. Note that 45 out of the total 112 in the library were selected specifically because they adhered well to the polypropylene plates used in the adsorption experiment (see below). The polymer numbering convention illustrated in Table 1 will be used throughout this text.

The solvent casting procedure has been described in detail previously.[14] Briefly, polymers were dissolved in methylene chloride (5% (w/v)). Next, the polymer solutions were filtered through 0.45 μm PTFE filters (Whatman Inc., Clifton, NJ). Then, individual polypropylene microtiter wells on the plates were filled with test polymer solutions. To evaporate the methylene chloride, the plates were kept at a temperature of 50 °C for 3 h in a drying oven. This process generated millimeter-thick and macroscopically smooth polymer films inside the wells.

Experimental Data: Immunofluorescence Assay. The development of a new microtiterplate-based technique for the simultaneous analysis of adsorbed proteins on multiple polymer samples using immunofluorescence has been reported previously.[14] Briefly, a 25 μL aliquot of fibrinogen in phosphate buffered solution (PBS) was incubated into polyarylate-coated wells on a 384-well polypropylene plate for 1.5 h at 37 °C, followed by rinsing with PBS. Wells were then incubated with 1% (w/v) bovine serum albumin in phosphate buffered solution (BSA-PBS) for 30 min at 37 °C in order to block nonspecific antibody binding. Afterward, the plates were rinsed with PBS and, subsequently, a measurement of the background signal was taken with the fluorescence reader (Spectra Max Gemini, Molecular Devices, Sunnyvale, CA). Fluorescently labeled antibodies were then allowed to bind to the surface-adsorbed fibrinogen for 1.5 h at 37 °C. Following this, the microwells were rinsed again with PBS, and the final fluorescence measurements were performed. Human fibrinogen adsorption to noncoated polypropylene wells was used as an internal control to normalize the fluorescence signals within different plates.

Physicomechanical and 2D Chemical Connectivity Descriptors. Certain laboratory-measurable quantities, such as the glass transition temperature (Tg), do depend on polymer structural properties in a reasonably well-understood way. Because of this, we have used two of these in this analysis to guide our interpretation of the influence of molecular-scale structural properties on protein adsorption. Tg is related to the degree to which relative chain motion occurs, which may dictate the type and quantity of functional groups that segregate to the surface of polyarylate films during or after solvent-casting. Protein adsorption to these surfaces is expected to depend strongly on the presence of certain functional groups at the surface. Also, the air−water contact angle (θ) was furnished as a macromeasure of polymer hydrophobicity. Hydrophobicity of polymer surfaces has been shown to correlate with protein adsorption when those surfaces are exposed to protein-rich serum.[2]

In addition to the above quantities, we also included 104 molecular-level descriptors of polymer chain structure. These two-dimensional, chemical-connectivity descriptors were not chosen for any particular, a priori, expected correlation to fibrinogen adsorption. Rather, they were chosen because all of them have been seen to successfully predict bioactivity of molecules in the type of Quantitative Structure Activity Relationship (QSAR) Analysis[15] commonly applied in the pharmaceutical industry.

These descriptors were calculated using the Molecular Operating Environment[17] commercial software package. Although MOE is able to calculate three-dimensional descriptors that depend on molecule conformation, only 2D, 1D, and 0D descriptors are used. Input to MOE includes basic molecular structure derived from chemical formulas and all MOE properties are calculations based on (relative) atomic coordinates.

Calculation assumptions and methodology are specific to each particular descriptor type and vary widely. However, all of the 104 descriptors used are fairly standard in QSAR analysis. Some are rather simple atomic accounting schemes. Examples include the Total Flexibility Index (TFI) (a count of the number of carbon and oxygen atoms in the variable regions of the polymer defined in Figure 1), “a_nH” (the number of hydrogen atoms), and counts of the number of double bonds. Others involve the projection of a particular property onto the van der Waals surface of the molecule. An example is the “positive van der Waals (VdW) surface area” which is a summation of the VdW surface areas of each of the atoms in the molecule which have nonnegative partial charges according to the Partial Equalization of Orbital Electronegativities (PEOE)[16] method. Others use PEOE to determine more explicit partial charge estimates. Finally, many are purely empirical parameters based on fitting linear functions of atomic contributions to large sets of experimental data for many different molecules. The most important of these is the logarithm of the octanol water partition coefficient.[23]

Perfect monomer and linear chain structure was assumed, and only chemical connectivity information is input into the structural model. The models represent chains of sufficient length such that the normalized value of each of the 102 descriptors varies less than 1% upon chain length doubling, where 1% error is an order of magnitude less than the error in the protein adsorption measurements. Molecules that contain 32 repeat units were observed to meet this criterion for all polymers wherein fibrinogen adsorption had been measured.

Decision-Tree Methodology. Once the descriptors had been obtained, they were ranked in order of their correlation to fibrinogen adsorption data using techniques commonly applied in machine learning routines. This step was important in three respects. First, including all descriptors in the ANN model would have been theoretically feasible but computationally impractical. Thus, it was necessary to develop a method to screen out all but the most relevant descriptors. Second, the relative weakness of the correlation of the QSAR descriptors and fibrinogen adsorption as well as the wealth of descriptor data made it nearly impossible to screen the descriptors manually. This was particularly true given the level of experimental uncertainty (17.9%, on average). Thus, it was absolutely necessary to do this using an automated, objective, and statistically valid method. Finally, the method developed, which allows the rapid screening of hundreds of molecular-level descriptors for correlations with bioresponse data, is of general utility in attacking the problem of bioresponse at a level beyond the phenomenological. This aspect of the method will be explored more deeply in future work.

The process of descriptor ranking was conducted as follows. The C5 Decision Tree routine[18] was used to calculate the information gain of each of the 104 descriptors with respect to fibrinogen adsorption. Information gain[19] (ig) is a measure of the ability of a particular attribute to classify a sample set. Simply put, the ig of descriptor is the decrease of the weighted average disorder of classification according to the target attribute after the set has been partitioned by that descriptor. In this case, the target attribute is fibrinogen adsorption.

To perform the calculation of the information gain for each descriptor, each polymer was classified according to the target attribute. The total range in fibrinogen adsorption was divided up into five equal "bins" of 20%, and the polymers were given integer classifications (1−5) according to the appropriate bin. The 20% value was chosen because it was comparable to the mean experimental uncertainty (i.e., 17.9%). Then, the ig of particular descriptor (D) for the entire set (E) of samples classified into the five, 20% subgroups is given by the following relation

$$\mathrm{ig}(E,D) = entropy(E) - \sum_{i=1,\ldots,n}\left(\frac{|E_i|}{|E|}\right) entropy(E_i) \quad (1)$$

where n = number of (user defined) different partitions of the descriptor in the sample set, E_i = the set of samples which have a descriptor value within partition i, $|X|$ = the number of samples in set X, and

$$entropy(E) = -\sum_{j=1,\ldots,5}\left(\frac{|E_j|}{|E|}\right)\log_2\left(\frac{|E_j|}{|E|}\right) \quad (2)$$

Note that the summation in (1) is over all descriptor values in the sample set, while the summation in (2) is over all five classes (i.e., fibrinogen adsorption "bins") in the sample set.

To take the experimental uncertainty into account in the descriptor ranking, a Monte Carlo approach was used. A sequence of 500 000 computer-based (pseudo) experiments was performed varying the value of fibrinogen adsorption measured for each polymer in a random fashion, but within a normal distribution defined by the experimental mean and standard deviation. Each pseudoexperiment yielded a single "most relevant" descriptor. This was the descriptor with the highest information gain as calculated via eq 1. In each pseudoexperiment a "dummy" descriptor, a random number between 0 and 1 generated for each polymer, was included. The purpose of the "dummy" descriptor was to allow the significance of each of the descriptors to be assessed by comparison to the significance of the random descriptor. The results for all 500 000 pseudoexperiments were tallied in a histogram. The three highest-ranking descriptors, defined as those with the highest counts in this histogram, were then selected as the input variables for the second stage, the Artificial Neural Network.

ANN Methodology. Overview and Motivation. Artificial Neural Networks (ANNs) have been applied successfully to various problems in molecular biology, especially to molecular sequencing applications.[20] They are essentially multivariate regression techniques that are particularly useful for pattern recognition. ANNs are ideal for situations in which experimental data are plentiful but a comprehensive theoretical framework is lacking. We chose an ANN for this application both because of this quality as well as its ability to tolerate experimental error. The network used in this case is a three-layer perceptron which is represented in Figure 2.

The input is a vector $\vec{x}$ of length $m + 1$ comprised of m descriptors (obtained from the Decision Tree analysis) plus a constant offset, and the output is a scalar prediction, z, of fibrinogen adsorption for the particular polymer represented by the descriptors. The vector of input variables is denoted by

$$\vec{x} = (x_o, x_1, \ldots, x_i, \ldots, x_m) \quad (3)$$

The additional term x_o is included to provide a constant offset and assigned the value $x_o = 1$. The vector of input variables for the kth polymer, $\vec{x}_k$, is denoted by

$$\vec{x}_k = (x_{k,o}, x_{k,1}, \ldots, x_{k,i}, \ldots, x_{k,m}) \quad (4)$$

The vector of hidden variables, $\vec{y}$, is denoted by

$$\vec{y} = (y_o, y_1, \ldots, y_i, \ldots, y_n) \quad (5)$$

where n is the number of hidden variables. The additional term, y_o, is included to provide a constant offset and assigned the value $y_o = 1$. The vector of hidden variables, $\vec{y}_k$, for the kth polymer is denoted by

$$\vec{y}_k = (y_{k,o}, y_{k,1}, \ldots, y_{k,i}, \ldots, y_{k,n}) \quad (6)$$

In Figure 2, $m = 3$ and $n = 2$.

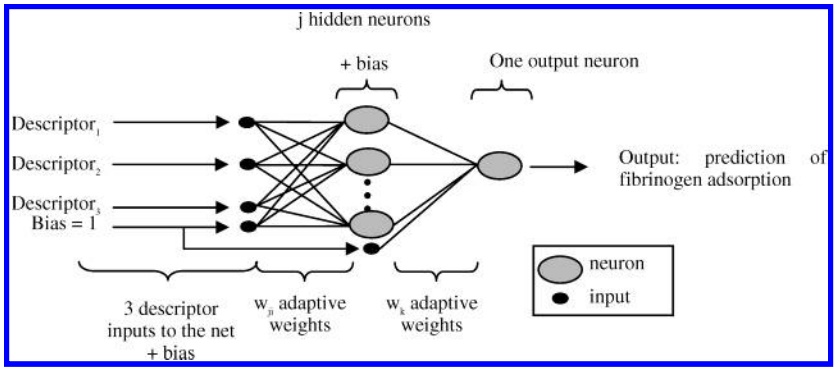


Figure 2. Schematic of ANN.

The hidden variables $\vec{y}_k$ for the kth polymer are calculated according to

$$y_{k,j} = f\left(\sum_{i=0}^{m} w_{j,i}{}^{0} x_{k,i}\right) \text{ for } k = 0,..., s \text{ - } 1 \text{ and } j = 1,n \quad (7)$$

where s is the number of polymers and the vector of weights $w_j{}^0$ for the jth hidden layer neuron is

$$w_j{}^0 = (w_{j,0}^0, w_{j,1}^0, ..., w_{j,i}^0, ..., w_{j,m}^0) \text{ for } j = 1,..., n \quad (8)$$

The function f defined by

$$f(\xi) = (1 + e^{-\kappa\xi})^{-1} \quad (9)$$

varies between 0 and 1 and is known as the *sigmoid function*. The quantity κ is user-specified. The output value for the kth polymer is denoted z_k and defined by

$$z_k = \sum_{j=0}^{n} w_j^1 y_{k,j} \quad (10)$$

where the vector of weights $\vec{w}^1$ for the output layers is

$$\vec{w}^1 = (w_0^1, w_1^1, ..., w_j^1, ..., w_n^1) \quad (11)$$

There are a total of $n(m+2) + 1$ unknowns in the ANN, namely, $n(m+1)$ values of $\vec{w}^0$ and $n + 1$ of $\vec{w}^1$. The weights $\vec{w}^0$ and $\vec{w}^1$ are obtained by minimizing the square error E defined by

$$E = \frac{1}{2}\sum_{k=0}^{s-1}(z_k - \hat{z}_k)^2 \quad (12)$$

for one-half of the polymers selected at random. This set is denoted the *training set*, and the remaining polymers are denoted the *validation set*. In (12), the experimental value for the kth polymer is denoted by $\hat{z}_k$ and the predicted value by z_k. The minimization is achieved using a Genetic Algorithm developed by Rasheed et al.[21]

The accuracy of the ANN is evaluated by comparison of the predicted and experimental fibrinogen adsorption for the remaining one-half of the polymers (the validation set). Two measures of accuracy are used. The first measure is the percent root-mean-square relative error E_{rms} defined by

$$E_{rms} = 100x\sqrt{\frac{1}{s_v}\sum_{k=0}^{s-1}\left(\frac{z_k - z_k}{\hat{z}_k}\right)^2} \quad (13)$$

where s_v is the number of polymers in the validation set. The second measure is the Pearson correlation coefficient

$$\rho = \frac{1}{\sigma\hat{\sigma}}\sum_{k=0}^{x_v-1}(z_k - \bar{z}_k)(\hat{z}_k - \bar{\hat{z}}) \quad (14)$$

where $\bar{z}$ and $\bar{\hat{x}}$ are the mean predicted and experimental values for the validation set, and σ and $\hat{\sigma}$ are the respective standard deviations, viz.,

$$\sigma = \sqrt{\frac{1}{s_v}\sum_{k=0}^{s_v-1}(z_k - \bar{z})^2} \quad (15)$$

There are three user-specified parameters in the ANN, namely (1) the number of descriptors m, (2) the number of neurons n in the hidden layer, and (3) the inverse length scale κ in the sigmoid function. The sensitivity of the predicted results to each of these parameters was assessed.

Again we took the experimental uncertainty into account using a Monte Carlo approach. A sequence of 100 computer-based (pseudo) experiments was performed wherein the mean value of fibrinogen adsorption for each polymer was perturbed by a random number obtained from a normal distribution based upon the standard deviation. Then, for each experimental data set, an ANN was built using one-half of the experimental data set (selected at random but identical for all pseudo-experiments) for training.

3. RESULTS AND DISCUSSION

Decision Tree Analysis. The results for the best 10 descriptors, along with their definitions, are compiled in Table 2. The “Significance” column in the table gives the ratio between the number of times each descriptor had the highest ig over 500 000 MC pseudoexperiments to the number of times the random “descriptor” was found to have the highest ig.

Two results here are worthy of particular note. First, the experimental measure of surface hydrophobicity, θ (not shown in Table 2), has less than one-fifth the significance of logP(o/w) and SlogP. The latter are empirical estimates of hydrophobicity based only on structure of individual, single polymer chains. We note that θ was measured on spin-coated polymers, while the films on which fibrinogen adsorption was measured were solvent cast. It is, perhaps, the difference between surface structures of solvent cast and spin coated films that yields this result. This result will be explored in the next section using ANN methodology.

Table 2. DT Monte Carlo Descriptor Results for Fibrinogen Adsorption

descriptor name	definition	significance
Tg	experimentally measured glass transition temperature	823
a_nH	number of hydrogen atoms in the molecule	509
logP(o/w)	log of the octanol/water partition coefficient (including implicit hydrogens) calculated from a linear atom type model[22]	489
SlogP	log of the octanol/water partition coefficient (including implicit hydrogens, atomic contribution model)[23]	353
number of secondary C (sp3)	the number of secondary carbons in each monomer which are sp3	264
b_count	number of bonds (including implicit hydrogens)	255
density	molecular mass density: weight divided by van der Waals volume	151
PEOE_VSA+0	sum of van der Waals surface areas of all atoms with a partial charge in the range [0.00,0.05)[16]	93
number of ethers (alphatic)	the number of aliphatic ethers in the molecule	89
SMR_VSA5	sum of van der Waals surface areas of all atoms with molar refractivities between (0.485,0.56][23]	89

Second, it is also unexpected that Tg, a bulk (measured) film property, should be more significant than over a hundred other properties—including estimates and experimental measures of hydrophobicity. We speculate that this indicates the importance of individual chain mobility in the formation of surface structure. It has been shown that surface roughness can have a substantial impact on cell proliferation and growth, which is related to protein adsorption.[24]

The Tg results, in particular, yield inspiration for more descriptors. Currently, efforts are ongoing to include more molecular-scale quantifications relating to chain mobility and surface structure, specifically taking into account three-dimensional chain structure and conformation. These will be correlated to measurements of surface roughness provided by Scanning Electron Microscopy or Atomic Force Microscopy. While intuition can guide this process, the use of Decision Tree analysis will prove an invaluable tool in descriptor selection and refinement.

ANN Prediction of Fibrinogen Adsorption. Figure 3a,b summarizes the results for ANN prediction using the three most significant descriptors: Tg, a_nH, and logP(o/w). The number of hidden neurons in the ANN that generated these results was two, and $\kappa = 0.1$. As will be shown in the next section, a sensitivity analysis of ANN predictive capability on intrinsic variables demonstrates that this result is robust with respect to the choice of the number of hidden neurons, κ or even the number of input variables within the limits tested.

The results are averaged over 100 MC pseudoexperiments. The training and validation sets were chosen at random but were identical in each pseudoexperiment. The predictive capability of the ANN model is assessed in reference only to the validation set.

As can be seen from Figure 3b, the ANN predicts 70% of the validation set results to within experimental error. The root-mean-square (rms) error of prediction for the validation set was 38%. This is an order of magnitude less than the range of fibrinogen adsorption seen in the experimental data (i.e., 366%). It also compares favorably with the average percent relative standard deviation of the measurements (i.e. 17.9%). The average Pearson correlation coefficient was 0.54. The effect of experimental error on this value was deduced by examining the distribution of validation set correlation coefficients in the 100 pseudoexperiments in the Monte Carlo analysis. The variation in the correlation coefficient was ~22% or 0.12. We note that the results are not entirely independent of the choice of training set. A duplicate Monte Carlo analysis performed with a somewhat more representative (but randomly chosen) training set yielded a validation set correlation coefficient of 0.66 ± 0.09.

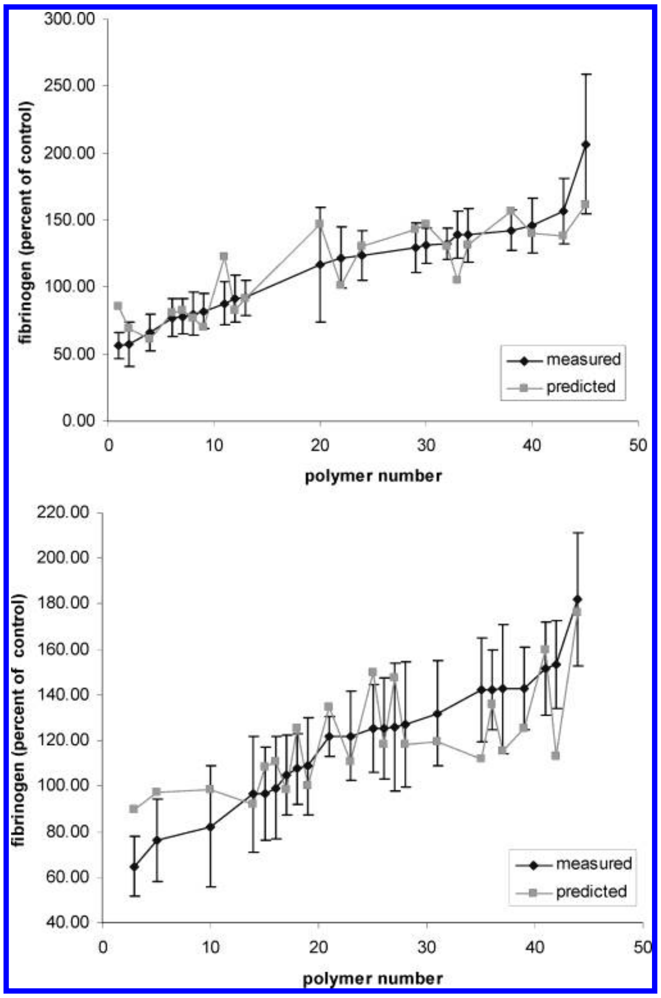


Figure 3. a. Training set for ANN built using three descriptors: Tg, a_nH, and logP(o/w). Note that the polymer numbering scheme is that used in Table 1. b. Validation set for ANN built using three descriptors: Tg, a_nH, and logP(o/w). Note that polymer numbering scheme is that used in Table 1.

To test the predictive capability of the model built with the above descriptors, results were compared to those from an ANN trained to model the same data set using three random "descriptors". The random descriptors were three random numbers generated for each individual polymer. The ANN modeling protocol was identical for the "random" model. The results are presented in Figure 4a,b.

Figure 4a shows that the ANN can fit the training set reasonably well even using random inputs. In fact, the

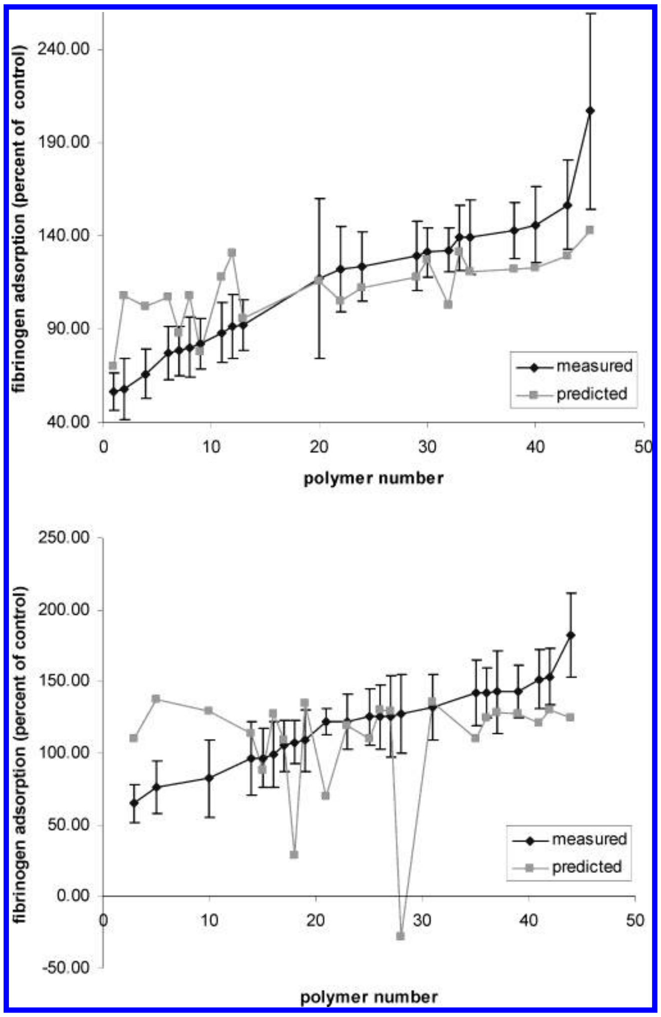


Figure 4. a. Training set results for ANN built using three "descriptors" which are simply random numbers generated for each polymer. Note that the polymer numbering scheme is that used in Table 1. b. Validation set for ANN built using three "descriptors" which are simply random numbers generated for each polymer. Note that polymer numbering scheme is that used in Table 1.

correlation coefficient for the random training set prediction is not terribly worse than the correlation coefficient for its counterpart ANN using the most significant descriptors (0.78 for the latter and 0.58 for the former). This shows the potential pitfall of using the entire data set to train the ANN. In these cases, successful models do not necessarily imply any physical/causal connection between input variables and output.

Despite this apparent "success" in fitting the training set with random variables, the resulting ANN using random inputs is essentially useless in predicting fibrinogen adsorption on polymers in the validation set. These results are presented in Figure 4b. Indeed, the average validation set correlation coefficient for the random case is essentially zero (0.08), an order of magnitude smaller than that for the descriptor-trained ANN. This strongly confirms the correlation between the selected descriptors and fibrinogen adsorption that was first predicted via the Decision Tree analysis.

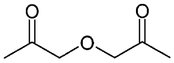
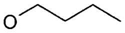
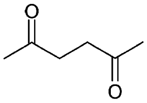
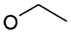
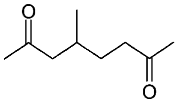
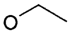
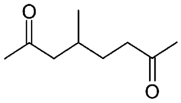
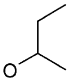
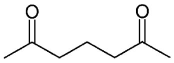
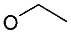
These results are, of course, expected. However, it should be emphasized that such a test provides a strong validation of the use of molecular-level polymer structural information in the model. Two of the three successful descriptors were the result of relatively simple molecular-level computations, while only the third was a macroscopic measurable quantity (Tg). Then, this analysis confirms the idea that one can take structural information provided by the synthetic chemist and use it to predict the absorption of specific proteins. This will dramatically reduce time needed for materials development in that it will allow the mapping of materials design space according to molecular-level properties. Only candidate materials in regions of design space found likely to have a relatively high biological activity will need to be selected for extensive testing.

A comparison of the top and bottom five polyarylates in terms of fibrinogen adsorption reveals sharp structural distinctions (Table 3). The most obvious influence arises from the choice of the pendant R group. Specifically, relatively high fibrinogen adsorption correlates with shorter R groups. Extension of the alkyl side chain (associated with the longer R groups) is expected to increase both the hydrophobicity of the polymer itself and the degree of steric hindrance it might encounter upon interacting with fibrinogen. Polymers with shorter R groups present a more cylindrical and less undulated surface to fibrinogen and, therefore, allow closer contact required to form strong van der Waals and hydrogen-bonding interactions. The influence of the Y group on fibrinogen adsorption again indicates the role of hydrophobicity and perhaps also chain flexibility. Shorter and, therefore, less flexible and less hydrophobic Y groups exhibit better fibrinogen adsorption. This conclusion is corroborated by the observation that DTB diglycolate (#45), which contains the most hydrophilic diacid group ($Y = -C(O)-CH_2-O-CH_2-C(O)-$), showed the highest measured fibrinogen adsorption by a considerable margin over the next best polymer HTE succinate (#44) in this series of polyarylates. The length of the diacid, together with the number of O atoms within it, will also determine the number of and spacing between potential hydrogen-bond donors along the polymer backbone. Although the precise role played by these donors with respect to interactions with the fibrinogen molecule is beyond the scope of the present study, the results summarized in Table 1 provide impetus to pursuing a detailed atomistic analysis of this phenomenon.

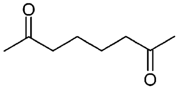
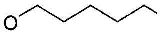
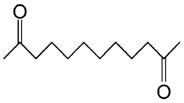
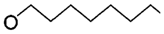
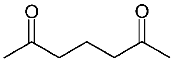
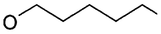
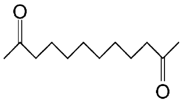
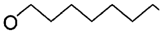
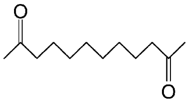
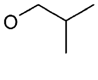
The corresponding values of the descriptors Tg, AWCA, a_nH, and logP(o/w) for the highest and lowest fibrinogen adsorbing polymers among this series can be interpreted in terms of the same factors as discussed above: hydrophobicity, hydrogen-bonding potential, chain flexibility, and steric accessibility. Enhanced fibrinogen adsorption is associated with higher values of Tg and lower values of a_nH and logP(o/w). Higher Tg values are typically indicative of a more rigid (less flexible) polymer chain and, all other things being equal, greater chain−chain interactions. Correspondingly, those polyarylates with shorter diacids (Y) and less extended pendant groups (R) will possess less flexible backbones and weaker chain-chain interactions. Lower logP values denote reduced hydrophobic character, again consistent with shorter diacids that contain a higher proportion of O atoms and with shorter pendant groups. Lower a_nH values further reflect the limited conformational flexibility, lower hydrophobicity, and shorter pendant groups (R) of polyarylates with improved fibrinogen adsorption. Taken together, these results provide useful clues for the rational design of polymers whose structural features lend themselves to optimal biological and materials properties.

Table 3. Influence of the Length of Diacid Component (Y) and Pendent Chain (R) on the Adsorption of Fibrinogen[a]

Top 5 polyarylates in terms of fibrinogen adsorption.

Diacid component (Y)	Polymer #	Pendent chain (R) **	Fibrinogen Adsorption (%)	Tg (°C)	*AWCA* (°)	*a_nH*	*logP(o/w)*
	45	 C_4H_9	206.91	64	72	29	3.30
	44	 C_2H_5	182.15	73	68	23	2.90
	43	 C_2H_5	156.74	63	75	31	4.23
	42	 C_4H_9	153.27	56	76	35	5.31
	41	 C_2H_5	151.44	69	69	27	3.43

Bottom 5 polyarylates in terms of fibrinogen adsorption.

Diacid component (Y)	Polymer #	Pendent chain (R) **	Fibrinogen Adsorption (%)	Tg (°C)	*AWCA* (°)	*a_nH*	*logP(o/w)*
	5	 C_6H_{13}	76.20	32	84	35	5.72
	4	 C_8H_{17}	66.00	13	96	49	8.46
	3	 C_6H_{13}	64.80	32	86	39	6.25
	2	 C_8H_{17}	57.60	23	86	43	7.49
	1	 C_4H_9	56.60	33	83	41	6.61

[a] LogP(o/w) and a_nH values given per monomer. Note that polymer numbering scheme is that used in Table 1.

The effectiveness of the experimental measure of hydrophobicity (θ) was explored because the Decision Tree results discussed in the previous section were counterintuitive. The Monte Carlo analysis presented in the previous section was repeated with the calculated hydrophobicity descriptor, logP(o/w) replaced by the experimental measure. In this test, identical validation and training sets were used. The results are summarized in Table 4 and indicate that models #1 and #2 are equivalent in accuracy.

This suggests a rather surprising result: the calculation of logP(o/w) for each molecule is as accurate a representation

Table 4. Comparison of Monte Carlo ANN Study Experimental vs Empirical/Calculated Hydrophobicity Measures[a]

model no.	descriptors	av training set correlation coefficient	av validation set correlation coefficient
		(100 trials)	(100 trials)
1	Tg, a_nH, logP(o/w)	0.78 ± 0.06	0.54 ± 0.12
2	Tg, a_nH, θ	0.76 ± 0.06	0.50 ± 0.11

[a] The correlation coefficients represent averages over 100 pseudo-experiments.

Table 5. Results of an Analysis of ANN-Generated Fibrinogen Adsorption Prediction Sensitivity to ANN Intrinsic Parameters

no. of hidden neurons	*k*	no. of input variables	correlation coeff (validation set)
2	0.01	3	0.771
2	0.01	5	0.695
2	0.01	10	0.689
2	0.10	3	0.803
2	0.10	5	0.742
2	0.10	10	0.811
4	0.01	3	0.738
4	0.01	5	0.722
4	0.01	10	0.636
4	0.10	3	0.787
4	0.10	5	0.756
4	0.10	10	0.732
6	0.01	3	0.785
6	0.01	5	0.714
6	0.10	3	0.745
6	0.10	5	0.767

of film surface hydrophobicity as the experimentally measured θ. It may correlate to the inherent variability in the θ measurement (on order of 2−4%) which is, at least in part, due to the fluctuation of ambient conditions such as the humidity. In any event, this observation will save considerable time and effort in materials development within the library. In particular, the value of logP(o/w) will be used in place of the rather laboriously measured θ to predict the fibrinogen adsorption onto solvent-cast polyarylates in cases for which data is not already available.

Sensitivity of ANN Predictions to User-Defined Parameters in ANN. As mentioned in section II, there are three user-defined parameters in the ANN: the number of hidden neurons, κ and the number of input variables in the net. We tested the sensitivity of ANN prediction of fibrinogen adsorption to each of these variables. This was done using only mean fibrinogen adsorption values (i.e., there was no Monte Carlo analysis, but the training set was the same as used for the aforementioned Monte Carlo analysis), and the six most significant descriptors were used in order of their significance. The latter means that the inputs to a three-variable ANN were the three most significant descriptors (Tg, a_nH, logP(o/w)), while the inputs to a ten-variable ANN were the 10 most significant descriptors (all of the descriptors in Table 2). The results are summarized in Table 5. Overall, the correlation coefficient for the validation set is 0.74 ± 0.05, thereby indicating that the ANN results are insensitive to these user-defined parameters. A detailed discussion follows below.

Effect of Number of Hidden Neurons. The number of hidden neurons in the ANN (Figure 2) was chosen to be two, four, or six. The results showed that ANN predictive capability is rather insensitive to the number of hidden neurons. Specifically, changing the number of hidden neurons from 2 to 6, with all other parameters held constant, accounted for a maximum change in the correlation coefficient of 9.7%. This is less than half of the variation in the average correlation coefficient due to experimental error (22%). The maximum impact of number of hidden neurons occurred for the model with $\kappa = 0.1$ and 10 input variables with the number of hidden neurons increased from two to four.

Effect of Number of the Number of Inputs to the Net. The number of inputs to the neural net was chosen to be three, five, or ten. Increasing the number of input variables from three to ten, while holding all other variables fixed, decreased the correlation coefficient in nearly every case. In every case, the change was considerably less than that due to the experimental variation. Specifically, in the two cases examined with two hidden neurons ($\kappa = 0.01$ and 0.10), the maximum variation in the correlation coefficient was 11% and 8%, respectively. In the two cases examined with four hidden neurons, this change decreased the correlation coefficient by 14% ($\kappa = 0.01$) and 7% ($\kappa = 0.01$). In the two cases examined with six hidden neurons, the correlation coefficient decreased by 9% ($\kappa = 0.01$) and increased by 3% ($\kappa = 0.10$). In each instance, the variation was significantly less than the variation of the average correlation coefficient due to experimental error (22%).

The general decrease in fit quality with the number of variables probably is due to two causes. First, as the number of variables increases so too does the possibility that the ANN will "overfit" the training set. This leads to a better training set correlation coefficient to the detriment of the validation set correlation coefficient (as was observed). Second, increasing the number of variables beyond three contributes additional variables of sharply decreasing significance. For example, the difference in significance between Tg and SMR_VSA5 (Table 2) is an order of magnitude. These results suggest that even descriptors with a moderate ig, e.g., 90 times more significant than random (SMR_VSA5), are not useful in this type of analysis.

Effect of the Shape of the Sigmoid Function. The effect of variation in the value of the constant κ in the sigmoid function of the neural net (3), essentially the sensitivity of the response of each neuron to an input, was also explored. If κ is too large, the output of a neuron will become saturated. This causes inaccurate or misleading results. For this reason, we performed each trial with two different values of κ, each differing by an order of magnitude (0.01 and 0.1). Additionally, κ values of 1.00 were also tried, but these generally resulted in neuron saturation and subsequent decrease in the correlation coefficient by up to 30% from values generated using neurons with $\kappa = 0.10$. The correlation coefficient increased as κ increased from 0.01 to 0.10 in seven out of eight cases. The maximum increase in the correlation coefficient occurred for the case of two hidden neurons and 10 input variables. However, even this change (18%) was less than the effect of experimental error. In the cases with three input variables, changing κ by an order of magnitude changes the correlation coefficient by less than 7%, less than one-seventh the effect of experimental uncertainty.

4. CONCLUSIONS AND FUTURE WORK

This research represents the first time that semiempirical computational models have been used to predict protein adsorption on biomaterials using entirely computed descriptors (input variables) based upon the polymer molecular structure. The model employed is novel in its use of molecular modeling, machine learning routines, and an Artificial Neural Network. Of particular importance was the methodology developed to use the information gain (ig) criterion in order to identify relevant descriptors. This technique will determine relevant properties of materials from large descriptor banks and will thus be a powerful tool in facilitating rational design of biomaterials. It was able to identify molecular-level empirical estimates of polymer hydrophobicity that were determined (via both the ig and the ANN analysis) to be as relevant as the experimental measure of air−water contact angle. These results will be applied directly in materials development.

Since computed polymer descriptors are less expensive to obtain than in vitro or in vivo measurements, the use of a computational modeling approach can significantly reduce the cost associated with identifying high performance biomaterials for specific applications. While we have demonstrated the feasibility of this approach for the prediction of fibrinogen adsorption only, the semiempirical models presented here can be extended to cover other important biological responses to materials, such as the adsorption of other relevant proteins, cell growth, migration, and differentiation of different cell types. These efforts are currently underway.

In addition, the bank of input parameters can be expanded to include more detailed molecular descriptors which reflect three-dimensional polymer structure and surface properties. This effort is also currently underway and involves the use of cutting-edge techniques in atomistic and meso-scale molecular-level modeling.

ACKNOWLEDGMENT

This work was supported by seed funds provided by Rutgers University, by NIH Grant R01 EB00286, and by “RESBIO - The National Resource for Polymeric Biomaterials” funded under NIH Grant P41 EB001046-01A1. Financial support was also provided by NIH Grant HL-60416 and the New Jersey Center for Biomaterials.

REFERENCES AND NOTES

(1) Magnani, A. et al. In *Integrated Biomaterials Science*; Kluwer Academic/Plenum: New York, 2002.

(2) Bloom, A. L. et al. In *Haemostasis and Thrombosis*; Livingstone: Edinburgh, 1994.

(3) Castner, D. G.; Ratner, B. D. Biomedical Surface Science: Foundations to Frontiers. *Surf. Sci.* **2002**, *500*, 28−50.

(4) Wold, S.; Hellberg, S. III W. J. D. Computer Methods for the Assessment of Toxicity. *Acta Pharmocol. Tox.* **1983**, 52 Suppl 2, 158−189.

(5) Hutmacher, D. W.; Vanscheidt, W. Matrices for Tissue-Engineered Skin. *Drugs Today (Barc)* **2002**, *38*, 113−133.

(6) Janvier, G. et al. Extracorporeal Circulation, Hemocompatibility and Biomaterials. *Ann. Thorac. Surg.* **1996**, *62*, 1926−1934.

(7) Kohnen, W.; Jansen, B. In *Handbook of Bacterial Adhesion*; Humana Press: Totowa, NJ, 2000.

(8) Chesmel, K. D.; Black, J. Cellular Responses to Chemical and Morphologic Aspects of Biomaterial Surfaces. I. A Novel in Vitro Model System. *J. Biomed. Mater.* **1995**, *29*, 1089−1099.

(9) Lee, J. H. et al. Interaction of Cells on Chargeable Functional Group Gradient Surfaces. *Biomaterials* **1997**, *18*, 351−358.

(10) Brocchini, S.; James, K.; Tangpasuthadol, V.; Kohn, J. A Combinatorial Approach for Polymer Design. *J. Am. Chem. Soc.* **1997**, *119*, 4553−4554.

(11) Hertz, J.; Palmer, R. G.; Krogh, A. In *Introduction to the Theory of Neural Computation*; Addison-Wesley Publishing Co.: Redwood City, CA, 1991.

(12) Fiordeliso, J.; Bron, S.; Kohn, J. In *Biomaterials in Solution, as Interfaces and as Solids*; VSP Publisher: Utrecht, The Netherlands, 1995.

(13) Brocchini, J. et al. Structure−Property Correlations in a Combinatorial Library of Degradable Biomaterials. *J. Biomed. Mater. Res.* **1998**, *42*, 66−75.

(14) Weber, N. et al. Small Changes in the Polymer Structure Influence the Adsorption Behavior of Fibrinogen on Polymer Surfaces: Validation of a New Rapid Screening Technique. *J. Biomed. Mater. Res.* **2004**, *68A*, 496−503.

(15) Seydel, J. K. In *QSAR and Strategies in the Design of Bioactive Compounds*; VCH: Weinheim, 1985.

(16) Gasteiger, J.; Marsali, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219.

(17) Chemical Computing Group Inc.; MOE (The Molecular Operating Environment), v. 2003.02; Montreal, Canada H3A 2R7, 2003.

(18) Research, P. L. R. C5.0, v. 5.0; St. Ives NSW 2075, Australia, 2002.

(19) Mitchell, T. M. In *Machine Learning*; McGraw-Hill: New York, 1997.

(20) Wu, C. H. Artificial Neural Networks for Molecular Sequence Analysis. *Comput. Chem.* **1997**, *21*, 237−256.

(21) Rasheed, K.; Hirsh, H.; Gelsey, A. A Genetic Algorithm for Continuous Design Space Search. *Artif. Intelligence Eng.* **1997**, *11*, 295−305.

(22) Labute, P. MOE LogP(Octanol/Water Model). unpublished.

(23) Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(24) Meredith, J. C. et al. Combinatorial Characterization of Cell Interactions With Polymer Surfaces. *J. Biomed. Mater. Res. A* **2003**, *66A*(3), 483−490.

CI0499774