Available online at www.sciencedirect.com
SCIENCE DIRECT
Polymer 46 (2005) 4296–4306
polymer
www.elsevier.com/locate/polymer

Predicting fibrinogen adsorption to polymeric surfaces in silico: a combined method approach

Jack R. Smith[a], Vladyslav Kholodovych[b], Doyle Knight[c], Joachim Kohn[a], William J. Welsh[b,*]

[a]The New Jersey Center for Biomaterials, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ, USA
[b]University of Medicine and Dentistry of New Jersey (UMDNJ), Robert Wood Johnson Medical School, The Informatics Institute of UMDNJ, Piscataway, NJ 08854, USA
[c]Department of Mechanical and Aerospace Engineering, Rutgers, The State University of New Jersey, New Brunswick, NJ 08903, USA

Available online 7 April 2005

Abstract

We present an improved semi-empirical (surrogate) model for the prediction of fibrinogen adsorption to the surfaces of polymers in a combinatorial library. The most important novel features of this model vis a vis our previous method is that it accurately predicts fibrinogen adsorption to a group of 20 polymers based on their structure alone, i.e. without using any experimental data for these 20 polymers. This implies that the model predictions can be generated prior to synthesis of these structures and their adsorption affinities can be evaluated entirely in silico. Modeling is accomplished by combining several more traditional computational methods in a 'hybrid' approach. The technique is used to systematically eliminate experimental inputs (air–water contact angle and glass transition temperature) from an existing artificial neural network model in favor of inputs that are derived mathematically from two-dimensional representations of polymer structure. We use partial least squares (PLS) regression to select structure-based molecular descriptors that are subsequently used to generate computational models for the subsequent prediction of protein adsorption by artificial neural networks (ANNs). The model provides accurate predictions of fibrinogen adsorption to polymeric surfaces using only adsorption data from a small representative subset of the polymer library. This work represents a major step toward the goal of generating virtual polymer libraries for rational design/optimization for properties relevant to biomedical applications.

Keywords: Biomaterials; Molecular modeling; Artificial neural network

1. Introduction

1.1. Fibrinogen adsorption and the performance of biomedical materials

Protein adsorption to surfaces is thought to be extremely relevant to biological/immunological response [1] in particular, and critical to cell attachment in general. The performance of materials in biomedical applications is highly dependent on the surface adsorption affinity for specific proteins [2]. Tailoring the fibrinogen adsorption affinity of biomedical implant materials, for example, is a necessity as this protein is known to participate in processes leading to blood clotting [3]. Doing this in a manner less costly than trial-and-error requires a method to predict the fibrinogen adsorption affinity for biomedical candidate materials based on their structure.

1.2. Modeling of protein adsorption and in silico materials evaluation using ANN

The most comprehensive way to predict protein adsorption onto a polymer surface is to explicitly calculate the energy of adsorption of individual fibrinogen molecules onto the implant material using either ab initio, atomistic or molecular mechanics methods. Indeed, recent progress has been made in evaluating the relative energies of adsorption of individual protein residues onto the surfaces of self-assembled monolayers [4–7]. However, such computations for entire protein molecules on the kinds of complicated surfaces likely to be used in biomedical applications are computationally intractable at the present time.

Therefore, we previously developed a semi-empirical approach for the prediction of fibrinogen adsorption to polymer surfaces and have demonstrated its usefulness

* Corresponding author.

0032-3861/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2005.03.012

within a combinatorial library of tyrosine-derived biodegradable polymers [8–10]. The computational procedure is inspired by quantitative structure activity relationship (QSAR) [11] model protocols developed by researchers in the pharmaceutical industry. QSAR is used primarily for designing small-molecule compounds with optimized bioactivity. From QSAR we borrow quantifications of molecular structure called 'molecular descriptors' to represent the polymers. The strategy allowed us to predict fibrinogen adsorption to polymers without explicitly representing the process in all its complexity. This was done by using adsorption data for a subset of polymers, mathematical descriptors and experimental measurements of physiomechanical properties (Fig. 1). In the 'training' phase, we used non-linear optimization to correlate (phenomenologically) calculated polymeric structural descriptors and two experimental descriptors [glass transition temperature (T_g) and air–water contact angle (θ) measurements] with measured protein adsorption for a subset of polymers in the library (Fig. 1). The subset of polymers used for training is referred to as the 'training set'. The correlations were then used, along with the same descriptors calculated or measured for the new structures, by an artificial neural network (ANN) model to predict the protein adsorption to polymers within the same family but outside the model training set in the 'validation' phase. Polymers whose fibrinogen adsorption was predicted based on their descriptors and the training set fibrinogen adsorption data belonged to the 'validation set'. We used an ANN instead of more traditional multivariate linear regression techniques because it has the capability to describe even nonlinear relationships between descriptors and experimental data [12].

The accuracy of the ANN fibrinogen adsorption model was sufficient to distinguish between the highest and lowest fibrinogen adsorbing polymers. Such a model permits the polymer scientist to focus attention only on the most promising candidate materials in lieu of exhaustive synthesis and experimental testing of all candidate polymers. This is important because polymer synthesis techniques have advanced to such a stage that the number of compounds that can be synthesized far exceeds the number that can be tested or evaluated in vivo and even in vitro at reasonable cost [9]. Thus, utilization of such surrogate (semi-empirical) models has the potential to speed the pace of biomedical materials development while saving considerable time and resources as well as to make rational optimization/design of materials for biomedical implants possible.

1.3. Limitation of previous model: experimental descriptor inputs required

Despite the success of our original fibrinogen adsorption model [10], we were unable to eliminate both the experimental inputs, T_g and θ, without substantial degradation of accuracy. This means that, in order for the ANN model to generate a prediction of fibrinogen adsorption to a specific polymer, that polymer has to be synthesized and then its T_g and θ have to be measured. While this model is still useful in avoiding unnecessary and expensive protein adsorption testing of non-optimal candidate polymers, it would respond better to the current needs of the biomedical materials research community if it could evaluate material performance entirely in silico, i.e. prior to synthesis. Then, suitable substitutes for T_g and θ in the form of calculated descriptors must be found.

1.4. Modeling of biological response using partial least squares regression

Recently, we have also demonstrated the capability of one of the more traditional QSAR approaches to model a relatively complicated phenomenon—cellular response in the form of the metabolic activity (MA) of fetal rat lung fibroblasts (FRLF) exposed to the surfaces for several hours. We used partial least squares regression (PLS) along with principal component analysis (PCA) to predict FRLF MA [13]. As with the ANN models, PLS models are trained on both experimental and molecular descriptor data and then

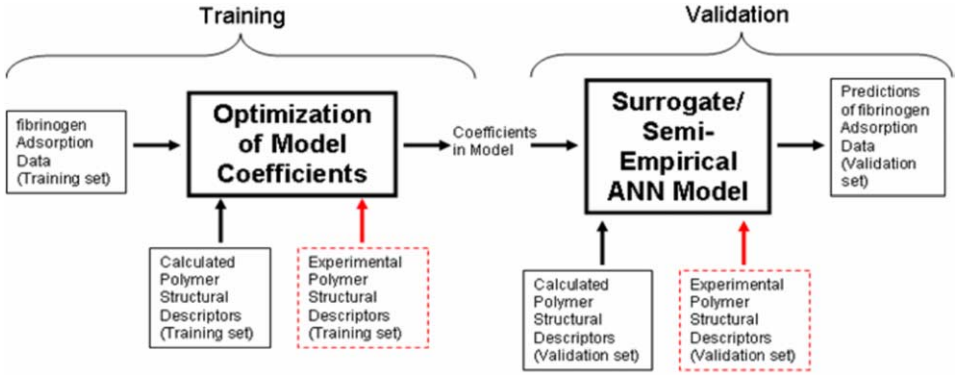

Fig. 1. Semi-empirical modeling using ANN. Our goal is to eliminate the experimental descriptor inputs (dashed boxes) for both the training and validation sets.

used to predict the former for novel or untested structures. The major differences between this and the ANN surrogate modeling discussed above rest in the computational details and the nature of the experimental data being predicted. The PLS models have shown a degree of accuracy sufficient for rational design in terms of cellular response, and the training methods are substantially less computationally expensive than those for ANN. In addition, the descriptor selection procedure in PLS is able to ascertain the aggregate significance of sets of descriptors by accounting for synergistic effects between them. While this is also theoretically possible with the descriptor–selection routine applied in the ANN approach [10], it is considerably less straightforward in that context.

1.5. The rationale for a combined method approach

Our goal in the present study is to eliminate the experimental inputs (T_g and θ) from the original ANN model for fibrinogen adsorption affinity without reducing the accuracy of prediction. First we show that this is not possible using the current single method (ANN) approach. Next, we present an approach that combines aspects of our ANN models with PLS in order to maximize the advantages of both. In this approach, PLS is used for feature extraction precisely because it can account for synergistic effects among collections of descriptors in a way that our ANN approach does not. Subsequently, the ANN is used in conjunction with the features, or descriptor sets, extracted by PLS in order to create a model of fibrinogen adsorption affinity that can account for non-linear relationships between polymer structural properties and this target property. Finally, we show that the combined approach generates a model that requires no experimental descriptor inputs yet is as accurate as the original model.

2. Methodology/background

2.1. The library of tyrosine-derived polyarylates

The ‘polyarylates’ are a series of structurally related polymers derived from monomers consisting of a tyrosine-derived diphenol and a diacid (Fig. 2). In total, there are 112 polymers resulting from all possible choices of diacid and pendent [14–16].

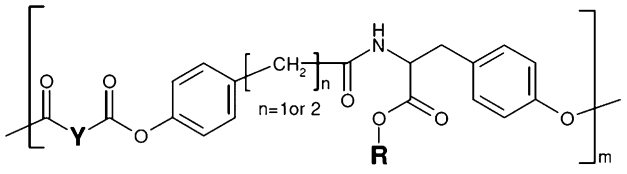

Fig. 2. Library of 112 polyarylates derived from 14 tyrosine-derived diphenols and 8 diacids. Polymers are strictly alternating copolymers consisting of a diacid (DA) and a diphenol (DP) component varied at *Y* and *R*, respectively. Commonly used pendant chains are ethyl, butyl, hexyl, octyl, and dodecyl esters, and diacids are succinic, glutaric, adipate, etc. The entire list of pendant chains and diacidic components can be found elsewhere [13]. The number of methyl groups in the DP component is also variable.

For this study, polymers were solvent-cast in a procedure that has been described in detail previously [16]. 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, USA). Then,

Table 1
List of 40 polyarylates used to build and validate the ANN models, together with values of fibrinogen adsorption amount (relative to polypropylene control) and standard deviation across four independent measurements

Pend.	Diacid	Fibrinogen adsorption (% PP control)	STDEV (% PP)
DTiB	Sebacate	56.60	9.96
DTO	Glutarate	64.80	12.96
DTO	Sebacate	66.00	13.46
HTH	Adipate	76.20	18.36
HTH	Suberate	77.16	13.97
DTO	Adipate	78.30	13.00
DTBn	Sebacate	80.30	16.22
DTO	Suberate	82.19	13.48
DTiB	Adipate	88.00	16.02
DTH	Suberate	91.68	17.24
DTBn	Suberate	92.10	13.35
DTH	Glutarate	96.75	20.61
DTM	Sebacate	99.14	22.50
DTB	Suberate	105.06	17.44
HTE	Suberate	107.69	15.40
DTsB	Suberate	108.71	21.52
DTiP	Adipate	121.76	8.89
DTO	Succinate	121.97	22.93
DTB	Methyl adipate	122.06	19.53
DTB	Glutarate	123.38	18.63
HTE	Adipate	125.19	19.28
DTsB	Adipate	125.35	22.19
DTM	Methyl adipate	125.70	28.03
DTB	Adipate	127.12	27.46
DTB	Succinate	129.36	18.76
DTE	Adipate	131.21	13.38
DTH	Succinate	131.93	22.96
DTsB	Glutarate	132.32	11.64
DTBn	Methyl adipate	138.98	17.65
DTM	Suberate	139.09	19.89
DTBn	Adipate	142.16	22.89
DTH	Diglycolate	142.29	17.36
DTO	Diglycolate	142.60	28.52
DTM	Adipate	142.69	15.13
DTiP	Methyl adipate	142.85	18.00
HTE	Methyl adipate	146.01	20.44
DTE	Glutarate	151.44	20.44
DTsB	Methyl adipate	153.27	19.47
DTE	Methyl adipate	156.74	24.14
HTE	Succinate	182.15	29.14

individual polypropylene micro titer wells on the plates were filled with test polymer solutions. In order 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 mm-thick and macroscopically smooth polymer films inside the wells.

2.2. Experimental data: immunofluorescence assay

The immunofluorescence protocol has been reported previously [17]. Briefly, a 25 μL aliquot of human 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 non-specific antibody binding. Afterwards, 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, USA). Fluorescently labeled antibodies were then allowed to bind to the surface-adsorbed fibrinogen for 1.5 h at 37 °C. Following this, the micro wells were rinsed again with PBS and the final fluorescence measurements were performed. Fibrinogen adsorption to non-coated polypropylene wells was used as an internal control to normalize the fluorescence signals within different plates. This procedure was carried out for 40 different polyarylates and the result is given in Table 1.

2.3. Molecular descriptors

A total of 109 descriptors were calculated using either the molecular operating environment [18] or Dragon [19] commercial software packages. Only two-dimensional (i.e. conformation independent) descriptors were used in the present study because the conformation of the polymers in the experimental environment was unknown. Input to MOE or Dragon consists of the basic molecular structure derived from the chemical formulae. Values of the various molecular descriptors are calculated as described elsewhere [10]. The descriptors vary widely from rather simplistic atomic and bond counting schemes to those representing the projection of a particular property (e.g. partial charge) [20] onto the van der Waals surface of the molecule. Others are purely empirical parameters based on fitting linear functions of atomic contributions to large sets of experimental data for many different molecules [21].

2.4. ANN methodology

The mathematical details of the computational methods used in the ANN procedure have been published elsewhere [10]. Briefly, modeling of fibrinogen adsorption in the polyarylate library is accomplished in two stages. In stage I, 109 ‘descriptors’ are generated for each polymer. These quantify various molecular and structural properties. Subsequently, the significance of each descriptor with respect to the set of experimental bioresponse data is ascertained using a quantity borrowed from machine learning routines called the Information Gain (IG) [22,23]. In stage II (depicted in Fig. 1), the most significant descriptors, in conjunction with the experimental data, are used to construct an ANN model for half of the polymer set (selected at random) in order to predict bioresponse for the remaining half of the data set. The experimental data was divided in half, one half (i.e. the training set) to train the model and the other half (the test set) to validate the model. The ANN used was a three-layer perceptron with two hidden neurons utilizing a sigmoid function with an inverse length scale parameter (κ) equal to 0.1.

One of the more novel aspects of this modeling procedure was the use of a Monte Carlo (MC) approach in the generation of the final predicted values. The purpose is to include the experimental variation explicitly in the model. We have shown that this gives a more realistic assessment of the ability to generalize model-generated predictions [10]. Here we provide only an outline of the procedure. A sequence of 1000 computer-based (pseudo) experiments is 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 experimental standard deviation. 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. The final predictions, then, are the averaged predictions over 1000 MC pseudo experiments for each polymer.

2.5. Partial least squares regression and principal component analysis

Though PLS regression and PCA can be used in the surrogate modeling of protein adsorption data, here they are both employed as a pre-filtering method to enrich the ANN model [24–27]. PLS/PC analysis allows us to assess the significance of combinations of descriptor variables in a way that accounts for synergetic effects that might exist among what are otherwise apparently uncorrelated parameters. As the details of the general computational methodologies appear elsewhere [13], in this section we focus on application of the method as used in the current context [28–32].

First, the principal components for the target property data were determined using the full set of 109 descriptors. Then, several PLS models were generated to predict the target property using from one to four of the most significant principal components. The correlation coefficient for each model was calculated, the results were tabulated and models with r values less than 95% of the correlation coefficient of the best statistical model were discarded. Of the remaining models, the one with the fewest number of principal

components was chosen to represent the data. Then, all principal components included in the PLS model were manually analyzed and the total contribution of each individual descriptor to the PCs was normalized, plotted and ranked. Subsequently, a new PLS model was generated using only the set of descriptors that had loadings in the first principal component [13] greater than 0.1. In order to eliminate all redundant descriptors, the forced pruning method was applied. This procedure is very similar to the well-known leave-one-out technique when each member of the parameter set is eliminated from the iteratively repeating PLS analysis. The final model, then, had the fewest descriptors, the fewest principal components and a correlation coefficient not less than 90% of that of the original model. This set of PLS/PCA-selected descriptors were then employed to construct our ANN model.

2.6. Model validation/evaluation

Each model is evaluated by calculating the 'correlation coefficient' r between the predictions and the experimental results. The coefficient r ranges from $-1 \le r \le 1$, where $r = 1$ is a perfect correlation and $r = -1$ represents a perfect inverse correlation. The mathematical definition of r is:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (1)$$

where n, number of polymers; x, predicted value; y, measured value.

All correlation coefficients are generated based on a comparison between the predicted values and the mean experimentally measured values without an accounting for the experimental variation. In the case where the model is the result of a series of predictions averaged over the course of a MC analysis, we report the correlation coefficient between the averaged predicted values and the experimental mean measured values. This is somewhat different than the number reported in our previous work [10], which was the average Pearson correlation coefficient over all pseudo-experimental models and the corresponding standard deviation of that quantity. Here we have chosen to report the correlation coefficients (r) of the averaged predictions because this allows a direct comparison with PLS/PCA results for which, at the time of writing, a MC analysis is not possible.

3. Results and discussion

3.1. Single model results: ANN

A summary of the results for our original model [10] for polyarylate fibrinogen adsorption affinity appears in Table 2 and the results for the validation set model generated in the 1000 step MC analysis (first row, Table 2) are plotted in Fig. 3. The three inputs to this model, the number of hydrogen atoms in the repeat unit, T_g and θ, were determined to be the most significant of the original set of 109 using the information gain criterion. Again, all of the predictions for the validation set (Fig. 3) are generated using only the fibrinogen adsorption data in the training set (i.e. the model is blind to the fibrinogen adsorption data in the validation set). Note that the model predictions are within experimental error for 75% of the validation set which is a level of accuracy sufficient for materials evaluation.

In an effort to replace the two measured inputs by calculated descriptors, we used the same methodology to generate models of both T_g and θ using the original set of 109 descriptors. We chose to use the five most important descriptors, as defined by the IG criterion, for each of these measurements in the ANN models. The statistical parameters are summarized in Tables 3a and 3b, while the validation set model predictions (1000 MC analysis, Tables 3a and 3b row 1) are plotted against the experimental data in Fig. 4.

First we note that the validation set correlation coefficients for the models T_g and θ models are 25 and 20% higher than for the fibrinogen adsorption model using these parameters as input (Table 2). This result was expected as the T_g and θ models predict physicomechanical properties that are far less complex than the physicochemical process of protein adsorption. Second, the most relevant descriptors (as defined by the IG criterion) to T_g and θ make intuitive sense. The T_g descriptors, for example, are the five descriptors in the original set of 109 that have some intuitive connection to chain flexibility. This is particularly obvious for the two most significant descriptors, the number of rotatable bonds and the Kier molecular flexibility index [33], but is also true for the number of single bonds as well as the latter two descriptors which both represent the shape or spatial extent of the molecule (Kier alpha modified shape index and third Kappa shape index [33]). The relevance of the most significant descriptor in the θ model, the number of secondary sp3 carbon atoms in the repeat unit, is less

Table 2
Statistics for fibrinogen adsorption model using the number of hydrogen atoms descriptor along with the two experimental inputs T_g and θ

Number of MC experiments	Training set fraction (%)	Training set correlation coefficient	Validation set correlation coefficient
1000	50	0.85	0.76
1000	100	0.80	na
1	100	0.80	na

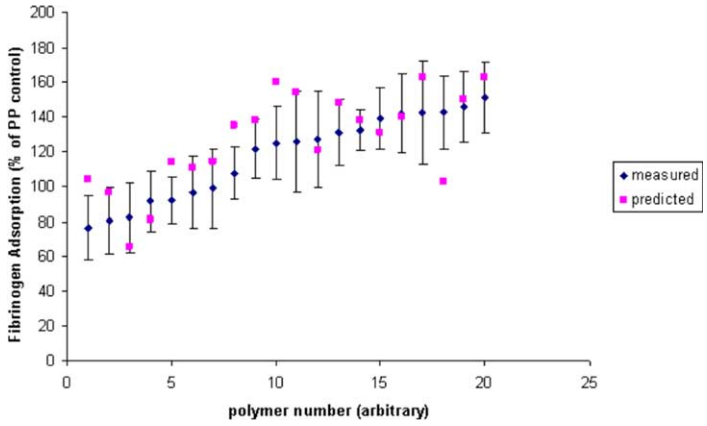

Fig. 3. Validation set results for 1000 MC fibrinogen adsorption model using descriptors in Table 2.

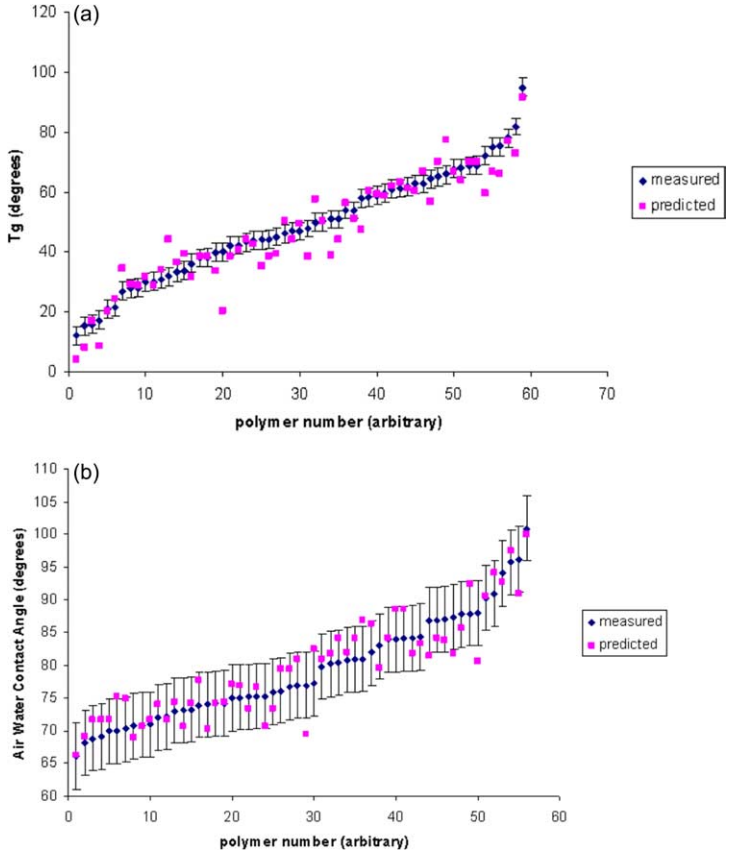

Fig. 4. (a) Validation set results for 1000 MC T_g model using descriptors in Table 3a. The average experimental error for the T_g measurement is $\pm 2°$. (b) Validation set results for 1000 MC θ model using descriptors in Table 3b. The average experimental error for the θ measurement is $\pm 5°$.

Table 3a
Statistics for T_g model using the 5 descriptors identified by the information gain criterion as being the most relevant: (1) the number of rotatable single bonds, (2) the Kier molecular flexibility index [33], (3) the number of single bonds, (4) 2-path Kier alpha modified shape index [33] and (5) the third kappa shape index [33]

Number of MC experiments	Training set fraction (%)	Training set correlation coefficient	Validation set correlation coefficient
1000	50	0.96	0.95
1000	100	0.96	na
1	100	0.95	na

Table 3b
Statistics for θ model using the 5 descriptors identified by the information gain criterion as being the most relevant: (1) the number secondary sp3 carbons, (2) the number of hydrogen atoms, (3) the hydrophilic factor [34], (4) the number of single bonds and (5) the sum of the van der Waals area of atoms in the molecule with negative partial charge [20]

Number of MC experiments	Training set fraction (%)	Training set correlation coefficient	Validation set correlation coefficient
1000	50	0.95	0.91
1000	100	0.94	na
1	100	0.93	na

Table 4
Statistics for fibrinogen adsorption model generated with T_g and θ replaced by calculated descriptors: (1) the number secondary sp3 carbons, (2) the number of hydrogen atoms, (3) the hydrophilic factor, (4) the number of single bonds and (5) the sum of the van der Waals area of atoms in the molecule with negative partial charge, (6) the number of rotatable single bonds, (7) the Kier molecular flexibility index, (8) 2-path Kier alpha modified shape index, and (9) third kappa shape index

Number of MC experiments	Training set fraction (%)	Training set correlation coefficient	Validation set correlation coefficient
1000	50	0.89	0.72
1000	100	0.82	na
1	100	0.80	na

intuitively obvious. However, the number of hydrogen atoms in the repeat unit, the hydrophilic factor [34], the number of single bonds descriptors can all be related to the overall chain hydrophobicity. The air–water contact angle θ corresponds to one of the best known measures of surface hydrophobicity.

Table 4 gives the statistics for the fibrinogen adsorption model that results when T_g and θ are replaced by the descriptors in their respective ANN models. The model has 9 inputs: the 'number of hydrogen atoms' descriptor from the original model along with the 8 unique descriptors from the T_g/θ set. The model predictions are compared with the experimental values in Fig. 5.

We note that this model is less accurate than the original (Fig. 3). This is reflected both in the decrease of the validation set correlation coefficient by 5.5% (Tables 2 and 4) and the fact that it predicts only 65% (Fig. 5) of data to within experimental error. The corresponding figure for the original model was 75%. Despite this, the resultant model represents a considerable advancement over the original in the sense that it does not include any experimental descriptors. Further, the input parameters are physically meaningful and many (like the number of rotatable bonds, the number of single bonds and the hydrophobicity) can be manipulated intuitively by the synthetic chemist. However, we would prefer a model that does not sacrifice accuracy and requires fewer descriptors. Such a model would be a more viable tool for in silico materials evaluation and rational materials design.

3.2. PLS descriptors

For the reasons outlined in the previous section, PLS is expected to yield descriptor sets with a higher aggregate predictive capability than the IG method used in the ANN analysis. Thus, we might expect that PLS/PCA will select a superior set of descriptors to replace the experimental quantities in the original fibrinogen adsorption model (Table 2, Fig. 3) and improve upon the results obtained with the IG analysis (Table 4 and Fig. 5).

The five significant descriptors according to the PLS/PC analysis for T_g and θ are given in Tables 5a and 5b. These are the descriptors that had the highest contributions to the first principal component for both models. Note that the T_g and θ descriptor sets in Tables 5a and 5b overlap to a much greater degree than those identified by the IG criterion (Tables 3a and 3b) and, in fact, to a much greater degree than might be expected given the difference in the molecular origins of the physiomechanical properties being modeled. Simply put, T_g depends strongly on molecule flexibility while θ on molecule hydrophobicity and these are conceptually distinct parameters. Indeed, this overlap implies two important points: (1) that there is redundancy in some of the descriptor definitions (i.e., the descriptors are not linearly independent), and (2) that there are synergistic effects between the descriptors that are unaccounted for by either intuition or individual examination as is performed in the IG analysis. The former point can be easily explained considering structural changes within the context of the polyarylate library. For example, the length of the aliphatic portion of the backbone ought to be related to both log $P(o/w)$ and the molecular density.[1]

It should also be noted that, while the descriptor sets identified by PLS/PCA (Tables 5a and 5b) are not identical with those identified by IG (Tables 3a and 3b), they are

[1] 'Density' or molecular density is defined as the molecular weight (calculated using atomic weights including implicit hydrogens) divided by the Van der Waals volume (calculated using a connection table approximation) [18].

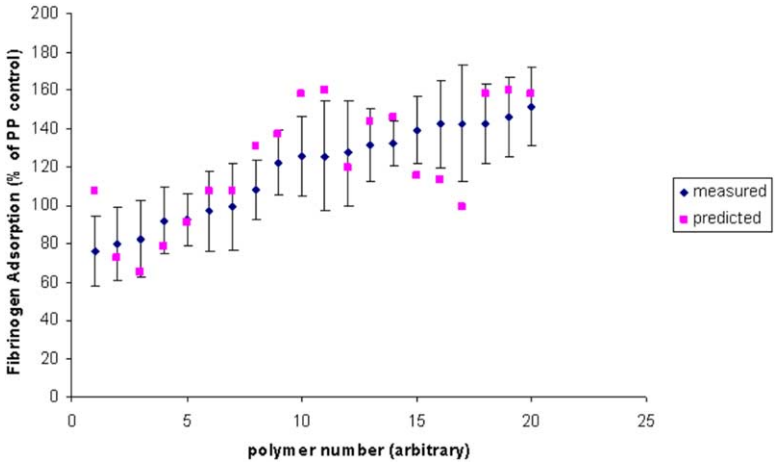

Fig. 5. Validation set results for 1000 MC fibrinogen adsorption model using descriptors in Table 4.

entirely consistent in terms of the physical properties being represented. For example, octanol/water partition coefficient reflects the hydrophobic/hydrophilic behavior of the molecule: a property strongly related to the hydrophilic factor in the IG model. In addition, the number of aliphatic ethers in the repeat unit corresponds to the sum of the van der Waals surface area of atoms in the molecule with negative partial charge in Table 5b. The molecular density, van der Waals volume and Kier index can be seen as important measures of the molecular shape complementarity as can the Kier molecular flexibility index, the number of single bonds, 2-path Kier alpha modified shape index, and the third kappa shape index from Table 5b.

To corroborate the above, we have shown that the set of unique descriptors identified by PLS contains effectively all of the physical information in the corresponding descriptors from the IG/ANN analysis. This was accomplished in the following manner. Altogether, 8 ANN models were built to predict each of the descriptors from IG/ANN analysis using the 8 descriptors supplied by the PLS/PCA analysis (Tables 5a and 5b). Each of the models was trained on half of the polyarylate structures (training set) and used to predict descriptor values for the other half (validation set). The results appear in Table 6. Note that the validation set correlation coefficient for each of the models was greater than 0.91 and the overall average was 0.96. This compares an average validation set correlation coefficient of 0.13 if eight random descriptor inputs are used to train the models in the same manner, indicating the high degree of predictive capability of this set of descriptors.

Table 7 gives the statistics for the fibrinogen adsorption model that results when T_g and θ are replaced by the 8 unique descriptors identified by PLS/PCA. The model has a total of 9 inputs, the ninth being the 'number of hydrogen atoms' descriptor from the original. The model predictions are compared with the experimental values in Fig. 6.

Note that the validation set correlation coefficient for this model is considerably improved over the value generated using the 8 IG descriptors along with the number of hydrogen atoms (Table 4 and Fig. 5). Indeed, the validation set correlation coefficient is the same as the original model containing T_g and θ (Table 2), yet all descriptors are computed (i.e. no experimental descriptors are used). Further, this model is able to predict 75% of data to within experimental error. This is more accurate than the IG descriptor model and equal in accuracy to the original. Thus, PLS/PCA descriptor selection has enabled us to generate an ANN model that is as accurate as the original, but uses no experimental descriptor inputs.

Despite the apparent success of this model, it is true that it represents a threefold increase in the number of descriptors over the original model (Fig. 3). In fact, some of these descriptors are much more relevant than others. For the both T_g and θ parameter sets in Tables 5a and 5b, the molecular density was the most important contributor following by the octanol/water partition coefficients and the number of aliphatic ethers. Models based only on these

Table 5a
The five most significant descriptors for T_g according to the PLS/PCA model

Significance rank	Molecular descriptor
1	The molecular density
2	Log of the octanol/water partition coefficient from the linear atom type model [35]
3	Log of the octanol/water partition coefficient calculated from the atomic model based on the corrected protonation state [36]
4	Molecular refractivity from the linear atom type model [21]
5	Sum of atomic van der Waals volume scaled on carbon atom

Table 5b
The five most significant descriptors for θ according to the PLS/PCA model

Significance rank	Molecular descriptor
1	The molecular density
2	Log of the octanol/water partition coefficient from the linear atom type model [35]
3	Number of aliphatic ethers in the repeat unit
4	The first kappa shape index of Kier [33]
5	Molecular refractivity calculated from the atomic model based on the corrected protonation state [36]

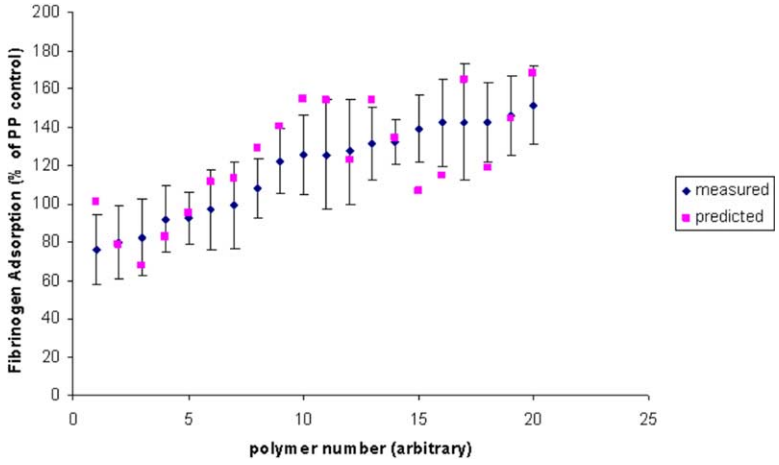

Fig. 6. Validation set results for 1000 MC fibrinogen adsorption model using descriptors in Table 7.

three descriptors could correlate 89% of the T_g and 91% of θ experimental results. Since we strive to create as simple model as possible, we eliminated all but the molecular density and the log of the octanol water partition coefficient from the model. The results appear in Table 8 and Fig. 7.

Note that dropping six descriptors from the model decreases the validation set correlation coefficient by only around 2.6%. This is an indication of the high predictive capability of the remaining descriptors, and further evidence of the synergistic effects between molecular density and the log of the octanol water partition coefficient. Further, this model is able to predict 75% of the experimental results to within experimental uncertainty. Therefore, its predictions are as accurate as the original model despite the fact that it includes no experimental descriptor inputs and has only three inputs.

4. Conclusions and future work

We have shown that it is possible to use an approach that combines artificial neural networks and partial least squares regression to generate models that accurately predict fibrinogen adsorption affinity based on structure alone, i.e. without the use of experimental descriptors. The present work, then, builds substantially on our previous results in

Table 6
Validation set correlation coefficients for ANN models generated using the eight unique descriptors identified by PLS (Tables 5a and 5b) to model the eight unique descriptors identified by IG

IG identified descriptor	Validation set correlation coefficient
The number of rotatable single bonds	0.92
2-Path Kier alpha modified shape index	0.94
Third kappa shape index	0.95
The number secondary sp3 carbons	0.95
The Kier molecular flexibility index	0.96
Sum of the van der Waals surface area of atoms in the molecule with negative partial charge	0.96
The number of single bonds	0.98
The hydrophilic factor	0.99

The average correlation coefficient for models using the same number of random descriptor inputs is 0.13.

Table 7
Statistics for fibrinogen adsorption model generated with T_g and θ replaced by calculated descriptors: (1) the molecular density, (2) the number of hydrogen atoms, (3) Log of the octanol/water partition coefficient from the linear atom type model, (4) the first kappa shape index of Kier, (5) the molecular refractivity calculated from the atomic model based on the corrected protonation state, (6) sum of atomic van der Waals volume scaled on carbon atom, (7) the number of aliphatic ethers, (8) Log of the octanol/water partition coefficient calculated from the atomic model based on the corrected protonation state, and (9) the molar refractivity from the linear atom type model

Number of MC experiments	Training set fraction (%)	Training set correlation coefficient	Validation set correlation coefficient
1000	50	0.86	0.76
1000	100	0.83	na
1	100	0.86	na

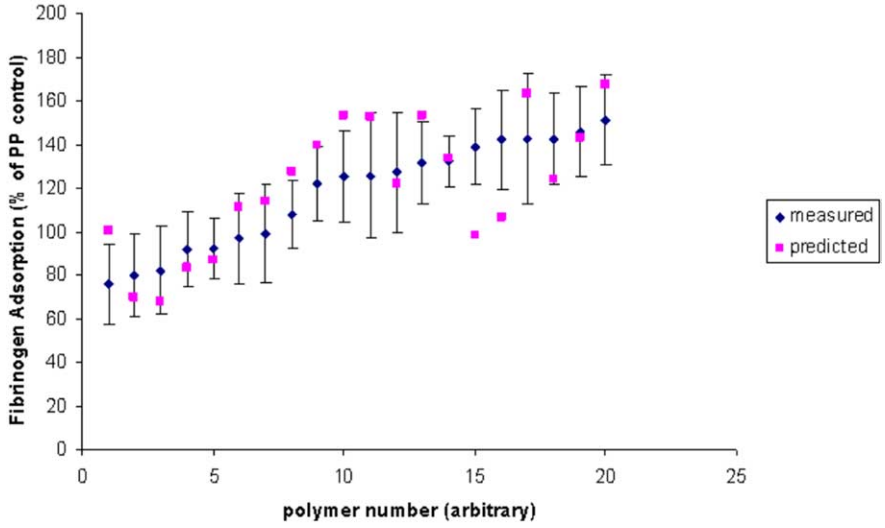

Fig. 7. Validation set results for 1000 MC fibrinogen adsorption model using descriptors in Table 8.

which we had used the two techniques in the combined method independently to accurately model biological response to polymeric materials. Here we have shown that using the two approaches in a complementary fashion generates a much more powerful model through using the advantages of both methods. Such an approach can be used to define a simulation protocol for generating models for other problems in biological response relevant to biomedical materials development such as the adsorption of proteins other than fibrinogen, cell response/proliferation, as well as immunological response (in the form of macrophage genotypic expression). Efforts in this direction are currently ongoing in our group.

Computational modeling approaches, which deploy readily calculated polymer descriptors instead of expensive and labor-intensive in vitro or in vivo measurements, can significantly reduce the costs and labor associated with identifying high-performance biomaterials for specific applications. Once the surrogate model has been built using experimental data for a specific biological response (e.g. fibrinogen adsorption) for a selected subset of polymers in a given library, we can use the model to evaluate the bioresponse for the remaining polymers entirely in silico, i.e. prior to synthesis. This not only enables researchers to take full advantage of the recent advances in combinatorial synthesis that allow the exploration of a fantastic amount of different materials, but also leads to the ability to design materials with optimized properties for biomedical applications. The application of methods of design optimization, well developed in other fields, may lead to great advances in biomaterials development.

Table 8
Statistics for fibrinogen adsorption model generated with T_g and θ replaced by calculated descriptors: (1) the molecular density, (2) the number of hydrogen atoms, (3) Log of the octanol/water partition coefficient from the linear atom type model

Number of MC exp.	Training set fraction (%)	Training set correlation coefficient	Validation set correlation coefficient
1000	0.5	0.85	0.74
1000	1	0.80	na
1	1	0.80	na

Acknowledgements

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

[1] Castner DG, Ratner BD. Surface Sci 2002;500:28–50.
[2] Magnani A, Peluso G, Margarucci S, Chittur KK. Protein adsorption and cellular/tissue interactions. In: Barbucci R, editor. Integrated biomaterials science. New York: Kluwer; 2002.
[3] Bloom AL, Forbes CD, Thomas DP, Tuddenham EGD. Haemostasis and thrombosis. Edinburgh: Livingstone; 1994.
[4] Latour RA, Rini CJ. J Biomed Mater Res 2002;60:564–77.
[5] Latour RA, Hench LL. Biomaterials 2002;23:4633–48.
[6] Wilson K, Stuart SJ, Garcia A, Latour RA. J Biomed Mater Res A 2004;69A:686–98.
[7] Basalyga DM, Latour RA. J Biomed Mater Res A 2003;64A:120–30.
[8] Smith J, Knight D, Kohn J, Rasheed K, Weber N, Abramson S. Proceedings of the materials research society, Boston, MA 2003.
[9] Smith J, Seyda A, Weber N, Knight D, Abramson S, Kohn J. Macromol Rapid Commun 2004;25:127–40.
[10] Smith J, Knight D, Kohn J, Rasheed K, Weber N, Kholodovych V, et al. J Chem Inf Comput Sci 2004;44(3):1088–97.
[11] Perkins R, Fang H, Tong W, Welsh WJ. Environ Toxicol Chem 2003; 22:1666–79.
[12] Hertz J, Palmer RG, Krogh A. Introduction to the theory of neural computation. vol. 1. Redwood City, CA: Addison-Wesley; 1991.
[13] Kholodovych V, Smith JR, Knight D, Abramson S., Kohn J, Welsh WJ. Polymer 2004;45:7367–79.
[14] Brocchini S, James K, Tangpasuthadol V. J Am Chem Soc 1997;119: 4553–4.
[15] Brocchini S, James KS, Tangpasuthadol V, Penharkar SM, Tong X. J Kohn proceedings of the society for biomaterials, New Orelans, LA 1997.
[16] Brocchini J, James K, Tangpasuthadol V, Kohn J. J Biomed Mater Res 1998;42:66–75.
[17] Weber N, Bolikal D, Bourke SL, Kohn J. J Biomed Mater Res 2004; 68A:496–503.
[18] Chemical Computing Group Inc. MOE (The Molecular Operating Environment), v. 2003.02. Montreal, Canada; 2003.
[19] Todeschini R, Consonni V, Mauri A, Pavan M. Dragon Web version. 3.0. Milano, Italy; 2003.
[20] Gasteiger J, Marsali M. Tetrahedron 1980;36:3219–22.
[21] Labute P. J Mol Graphics Mod 2000;18:646–77.
[22] Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.
[23] RuleQuest Research, P.L. C5, v. 30 NSW, Australia; 2000.
[24] Draper N, Smith H. Applied regression analysis. 2nd ed. NY: Wiley; 1981.
[25] Anderson T. An introduction to multivariate statistical analysis. 3rd ed. New York: Wiley; 2003.
[26] Wold S. PLS for multivariate linear modelling. In: deWaterbeemd HV, editor. Methods and principles in medicinal chemistry. Weinheim, Germany: Verlag; 1995.
[27] Wold S, Albano C, Dunn III WJ, Edlund U, Esbensen K, Geladi P, et al. Multivariate data analysis in chemistry. In: Kowalski BR, editor. Chemometrics: mathematics and statistics in chemistry. Dordrecht, The Netherlands: Reidel; 1984. p. 17–95.
[28] Tong W, Perkins R, Xing L, Welsh WJ, Sheehan DM. Endocrinology 1997;138:4022–5.
[29] Tong W, Perkins R, Chen Y, Welsh WJ, Lowis DR, Goddette DW, et al. J Chem Inf Comp Sci 1998;38:669–77.
[30] Jayatilleke P, Nair A, Zauhar R, Welsh WJ. J Med Chem 2000;43: 4446.
[31] Puri S, Chickos J, Welsh WJ. J Chem Inf Comp Sci 2002;42:209–14.
[32] Yu SJ, Keenan SM, Tong W, Welsh WJ. Chem Res Toxicol 2002; 15(10):1229–34.
[33] Hall LH, Kier LB. The molecular connectivity chi indexes and kappa shape indexes in structure–property relations. In: Boyd D, Lipkowitz K, editors. Reviews of computational chemistry New York, 1991. p. 367–422.
[34] Todeschini R, Consonni V, Mauri A, Pavan M. Dragon Web version. 3.0. Milano, Italy; 2003.
[35] Lin A. J Chemical Computing Group (http://www.chemcorp.com/Journal_of_CCG/Features/descr.htm).
[36] Wildman SA, Crippen GM. J Chem Inf Comput Sci 1999;39:868–73.