Linnæus University

Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Project report:

June – September 2012

Application of statistical methods in rational design of biomaterials

Part II

AIM

Developing a QSAR (Quantitative Structure–Activity Relationship) model capable of predicting immunological properties of biopolymers, based on their physicochemical characteristics.

BIOPOLYMERS

The polymers were synthesized using (where applicable) a 20:80 monomer – crosslinker ratio[1] combinations of various chemicals (Table I).

TABLE I POLYMER COMPOSITION

Polymer	Monomer	Crosslinker
P1	MAA	DAP
P2	MAA	DVB
P3	MAA	EGDMA
P4	IPAAm	EGDMA
P5	Styrene	EGDMA
P6	HEMA	EGDMA
PS	Styrene	–
PVC	Vinyl chloride	–
Glass	–	–

QSAR

The QSAR (Quantitative Structure–Activity Relationship) methodology is based on the assumption that there is a close connection between the structure of a molecule and its physicochemical properties and biological activity. It is possible to quantify that relationship by means of chemometric methods, using theoretical and empirical descriptors (sometimes referred to as 'predictor' variables, X) and experimental endpoint values, the response variable, Y.

The resulting equation, or *model*, mathematically describes the relationship between molecular and biological properties of compounds. It is also capable of predicting the properties of whole new groups of substances, on the condition that they are structurally similar to the ones the model was based on.

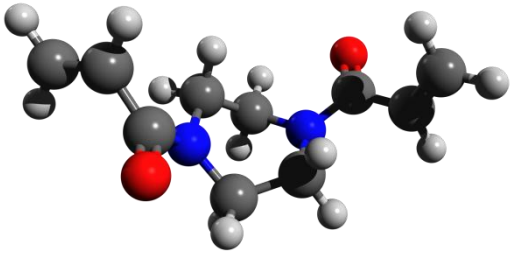
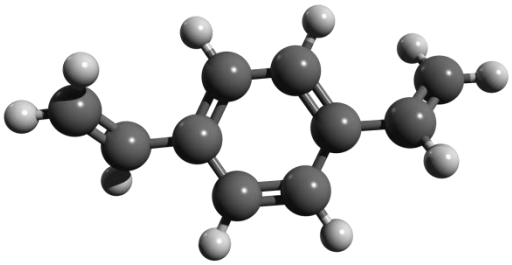
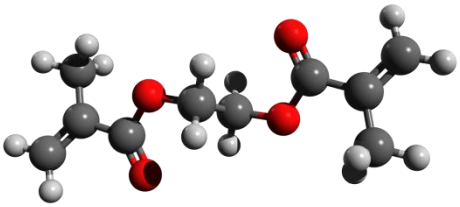
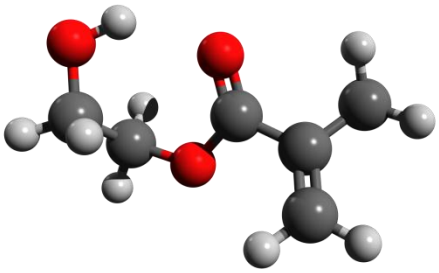
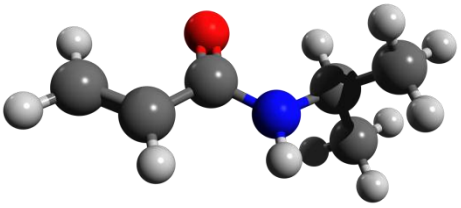
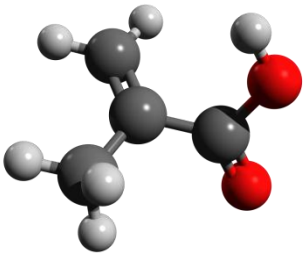
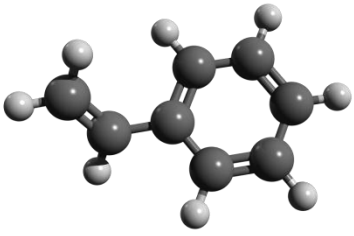
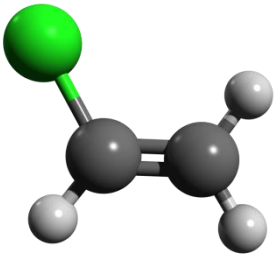
DESCRIPTORS

Two types of descriptors were utilized: experimental and computational. Five experimental descriptors had been previously determined for polymers P1-P6: particle surface area, particle pore diameter, swelling, specific swelling and contact angle (Drop-snake method). Their values were used in the analysis without any transformation. Since no experimental descriptor were available for PVC, PS and glass particles, those samples were omitted during the modeling process.

In order to obtain molecular descriptors for the biopolymers, 3D structures of all monomers and crosslinkers were built with the help of ACD-ChemSketch (Table II). Subsequent geometry

optimization was performed using MOPAC2012 computational package via gabedit 2.4.4. All calculations were done at the PM6 level of precision.

TABLE II OPTIMIZED STRUCTURES OF MONOMER AND CROSSLINKER MOLECULES

	
N,N-diacryloylpiperazine, **DAP**	divinylbenzene, **DVB**
	
ethylene glycol dimethacrylate, **EGDMA**	2-hydroxyethyl methacrylate, **HEMA**
	
N-isopropyl acrylamide, **IPAAm**	methacrylic acid, **MAA**
	
styrene	vinyl chloride

Following that, molecular descriptors were generated for with employing Dragon6 software. Four blocks of descriptors were chosen as the basic set: constitutional indices, functional group

counts, molecular properties and P_VSA-like descriptors. Their initial number was reduced from 262 to 78 (Table III) – those with constant, near constant, missing or null values were discarded.

As the descriptors had been calculated for single monomer and crosslinker molecules, their weighted average values (D_w) were used for modeling. According to the polymer composition ratios, weight $w_1 = 0.2$ was assigned to monomer descriptors (D_m) and weight $w_2 = 0.8$ to crosslinker descriptors (D_c).

$$D_w = w_1 \times D_m + w_2 \times D_c$$

TABLE III DESCRIPTOR LIST

No.	Name	Description
1	PSA	particle surface area
2	PD	particle pore diameter
3	SW	swelling
4	SPSW	specific swelling
5	CA	contact angle (Drop-snake method)
6	MW	molecular weight
7	AMW	average molecular weight
8	Sv	sum of atomic van der Waals volumes (scaled on Carbon atom)
9	Se	sum of atomic Sanderson electronegativities (scaled on Carbon atom)
10	Sp	sum of atomic polarizabilities (scaled on Carbon atom)
11	Si	sum of first ionization potentials (scaled on Carbon atom)
12	Mv	mean atomic van der Waals volume (scaled on Carbon atom)
13	Me	mean atomic Sanderson electronegativity (scaled on Carbon atom)
14	Mp	mean atomic polarizability (scaled on Carbon atom)
15	Mi	mean first ionization potential (scaled on Carbon atom)
16	nAT	number of atoms
17	nSK	number of non-H atoms
18	nBT	number of bonds
19	nBO	number of non-H bonds
20	nBM	number of multiple bonds
21	SCBO	sum of conventional bond orders (H-depleted)
22	RBN	number of rotatable bonds
23	RBF	rotatable bond fraction
24	nDB	number of double bonds
25	nH	number of Hydrogen atoms
26	nC	number of Carbon atoms
27	nO	number of Oxygen atoms
28	nHet	number of heteroatoms
29	H%	percentage of H atoms
30	C%	percentage of C atoms
31	N%	percentage of N atoms
32	nCsp3	number of sp3 hybridized Carbon atoms
33	nCsp2	number of sp2 hybridized Carbon atoms
34	P_VSA_LogP_1	P_VSA-like on LogP, bin 1
35	P_VSA_LogP_2	P_VSA-like on LogP, bin 2
36	P_VSA_LogP_4	P_VSA-like on LogP, bin 4
37	P_VSA_LogP_5	P_VSA-like on LogP, bin 5
38	P_VSA_LogP_7	P_VSA-like on LogP, bin 7
39	P_VSA_m_1	P_VSA-like on mass, bin 1
40	P_VSA_m_2	P_VSA-like on mass, bin 2

41	P_VSA_m_3	P_VSA-like on mass, bin 3
42	P_VSA_v_1	P_VSA-like on van der Waals volume, bin 1
43	P_VSA_v_2	P_VSA-like on van der Waals volume, bin 2
44	P_VSA_v_3	P_VSA-like on van der Waals volume, bin 3
45	P_VSA_e_1	P_VSA-like on Sanderson electronegativity, bin 1
46	P_VSA_e_2	P_VSA-like on Sanderson electronegativity, bin 2
47	P_VSA_e_5	P_VSA-like on Sanderson electronegativity, bin 5
48	P_VSA_p_1	P_VSA-like on polarizability, bin 1
49	P_VSA_p_2	P_VSA-like on polarizability, bin 2
50	P_VSA_p_3	P_VSA-like on polarizability, bin 3
51	P_VSA_i_2	P_VSA-like on ionization potential, bin 2
52	P_VSA_i_3	P_VSA-like on ionization potential, bin 3
53	P_VSA_s_2	P_VSA-like on I-state, bin 2
54	P_VSA_s_3	P_VSA-like on I-state, bin 3
55	P_VSA_s_4	P_VSA-like on I-state, bin 4
56	P_VSA_s_6	P_VSA-like on I-state, bin 6
57	SPAM	average span R
58	MEcc	molecular eccentricity
59	SPH	spherosity
60	ASP	asphericity
61	PJI3	3D Petitjean shape index
62	L/Bw	length-to-breadth ratio by WHIM
63	nCp	number of terminal primary C(sp3)
64	nCconj	number of non-aromatic conjugated C(sp2)
65	nR=Cp	number of terminal primary C(sp2)
66	nR=Ct	number of aliphatic tertiary C(sp2)
67	nHAcc	number of acceptor atoms for H-bonds (N,O,F)
68	Uc	unsaturation count
69	Ui	unsaturation index
70	Hy	hydrophilic factor
71	AMR	Ghose-Crippen molar refractivity
72	TPSA(NO)	topological polar surface area using N,O polar contributions
73	TPSA(Tot)	topological polar surface area using N,O,S,P polar contributions
74	MLOGP	Moriguchi octanol-water partition coeff. (logP)
75	MLOGP2	squared Moriguchi octanol-water partition coeff. (logP^2)
76	ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)
77	ALOGP2	squared Ghose-Crippen octanol-water partition coeff. (logP^2)
78	SAtot	total surface area from P_VSA-like descriptors
79	SAacc	surface area of acceptor atoms from P_VSA-like descriptors
80	Vx	McGowan volume
81	VvdwMG	van der Waals volume from McGowan volume
82	VvdwZAZ	van der Waals volume from Zhao-Abraham-Zissimos equation
83	PDI	packing density index

MODELED PROPERTIES

Two types of biological responses were used as endpoints: levels of proteins adsorbed on polymer surface and concentrations of immunomarkers[2] induced in blood while in contact with polymer particles – 56 characteristics in total.

Endpoints with less than three sample measurements per polymer as well as those with unequal number of measurements were discarded. The rest was divided into four blocks in accordance to the experimental method used and type of immunological response (Table IV).

Due to the large range of variable values (the difference between the largest and smallest values amounted to 3 orders of magnitude), a logarithmic transformation of the data was conducted:

$$y_t = \log_{10}(y)$$

TABLE IV GROUPING OF MODELED PROPERTIES. GREYED-OUT AREAS INDICATE DISCARDED ENDPOINTS

		Block A																			Block B					
hir-plasma	Totalprot	a2-M	C4BP	C1q	Fib	C4	C3	C5	Factor H	IgG	FXII	Haptogl	C1INH	Factor I	Vn	Transferrin	HSA	AT	Hemopexin	ApoAIV	ApoAI	FXIIa-C1 INH	FXIaC1 INH	KalI-C1 INH	FXIIa-AT	FXIa-AT	KalI-AT
Block C																								Block D			
C3a	C5a	TCC	Plt loss %	Gran loss %	utan add	med C5aRA	TNF-a	IFN-g	IL-17	IL-6	IL-1ra	IL-10	IL-8	Eotaxin	IP-10	MCP-1	MIP-1a	MIP-1b	IL-9	PDGF	FGF	GM-CSF	VEGF	C3a	TCC	Plt loss %	TAT

METHODS

A hybrid GA-MLR technique was used to develop the model. All the chemometric calculations were performed with the PLS_Toolbox 6.7 in combination with Matlab 7.11 (R2010b).

GENETIC ALGORITHM

Genetic algorithm variable selection is a technique that helps identify a subset of the measured variables that are, for a given problem, the most useful for a precise and accurate regression model. Given an X-block of predictor data and a Y-block of values to be predicted, one can choose a random subset of variables from X and, through the use of cross-validation, determine the root-mean-square error of cross validation ($RMSE_{CV}$) obtained when using only that subset of variables in a regression model. Genetic algorithms use this approach iteratively to locate the variable subset (or subsets) which gives the lowest $RMSE_{CV}$.

(Matlab) PLS_Toolbox 6.7:

PLS Workspace Browser >> Analysis Tools >> Other >> GA variable selection
SIZE OF POPULATION: 92
WINDOW WIDTH: 1
% INITIAL TERMS: 30
TARGET MIN/MAX: 0/8
PENALTY SLOPE: 0.001
MAX GENERATIONS: 200

% AT CONVERGENCE: 90
MUTATION RATE: 0.01
CROSSOVER: double
REGRESSION CHOICE: MLR
CROSS-VALIDATION: Contiguous
OF SPLITS: 5
OF ITERATIONS: 1
REPLICATE RUNS: 5

MULTIPLE LINEAR REGRESSION

Multiple linear regression attempts to model the relationship between two or more explanatory variables (X) and a response variable (y) by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for n explanatory variables x_1, x_2, ... , x_n is defined to be

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

This line describes how the mean response μ_y changes with the explanatory variables. The observed values for y vary about their means y and are assumed to have the same standard deviation σ. The fitted values b_0, b_1, ..., b_n estimate the parameters β_0, β_1,…,β_n of the population regression line.

(Matlab) PLS_Toolbox 6.7:

PLS Workspace Browser >> Analysis Tools >> REGRESSION >> MLR – Multiple Linear Regression
PREPROCESSING: none
CROSS-VALIDATION: contiguous block
OF DATA SPLITS: 6

CALIBRATION

To measure how well the model represents empirical data, determination coefficient R^2 and the root mean square error of calibration $RMSE_C$ is calculated. The closer the R^2 value is to 1 and the smaller $RMSE_C$, the better the model fitting.

$$RMSE_C = \sqrt{\frac{\sum(y^{exp} - y^{pred})^2}{n}}$$

where:
y^{exp} *– experimental values of the Y variable*
y^{pred} *– estimated values of the Y variable*
n *– total number of objects in the data set*

$$R^2 = 1 - \frac{\sum(y_i^{opred} - y_i^{exp})^2}{\sum(y_i^{exp} - \bar{y}^{exp})^2}$$

where:

y^{exp} *– experimental values of the Y variable*
y^{pred} *– estimated values of the Y variable*
$\bar{y}^{exp}$ *– mean experimental values of the Y variable*
n *– total number of objects in the data set*

CROSS-VALIDATION

Cross validation is a very useful tool that serves two critical functions in chemometrics - it enables an assessment of the optimal complexity of a model and allows an estimation of the performance of a model when it is applied to unknown data.

For a given data set, cross validation involves a series of experiments, each of which involves the removal of a subset of objects from a dataset (the test set), construction of a model using the remaining objects in the dataset (the model building set), and subsequent application of the resulting model to the removed objects. This way, each experiment involves testing a model with objects that were not used to build the model. A typical cross-validation procedure usually involves more than one sub-validation experiment, each of which involves the selection of different subsets of samples for model building and model testing.

The robustness of a model can be assessed by calculating the determination coefficient R^2_{CV} and the root mean square error of cross-validation $RMSE_{CV}$. The closer the R^2_{CV} value is to 1 and the smaller $RMSE_{CV}$, the better greater the flexibility (robustness) of the model.

$$RMSE_{CV} = \sqrt{\frac{\sum(y^{exp} - y^{pred\_cv})^2}{n}}$$

where:
y^{exp} *– experimental values of the Y variable*
y^{pred_cv} *– estimated values of the temporary excluded (cross-validated) sample*
n *– total number of objects in the data set*

$$R^2_{CV} = 1 - \frac{\sum(y_i^{pred\_cv} - y_i^{exp})^2}{\sum(y_i^{exp} - \bar{y}^{exp})^2}$$

where:
y^{exp} *– experimental values of the Y variable*
y^{pred_cv} *– estimated values of the temporary excluded (cross-validated) sample*
$\bar{y}^{exp}$ *– mean experimental values of the Y variable*
n *– total number of objects in the data set*

RESULTS

BLOCK A: PROTEIN SURFACE ADSORPTION LEVEL MEASUREMENTS

The GA-MLR found 16 unique models with 4 to 8 variables (Figure 1). $RMSE_{CV}$ ranged from 0.158 to 1.7×10^{10} logarithmic units.

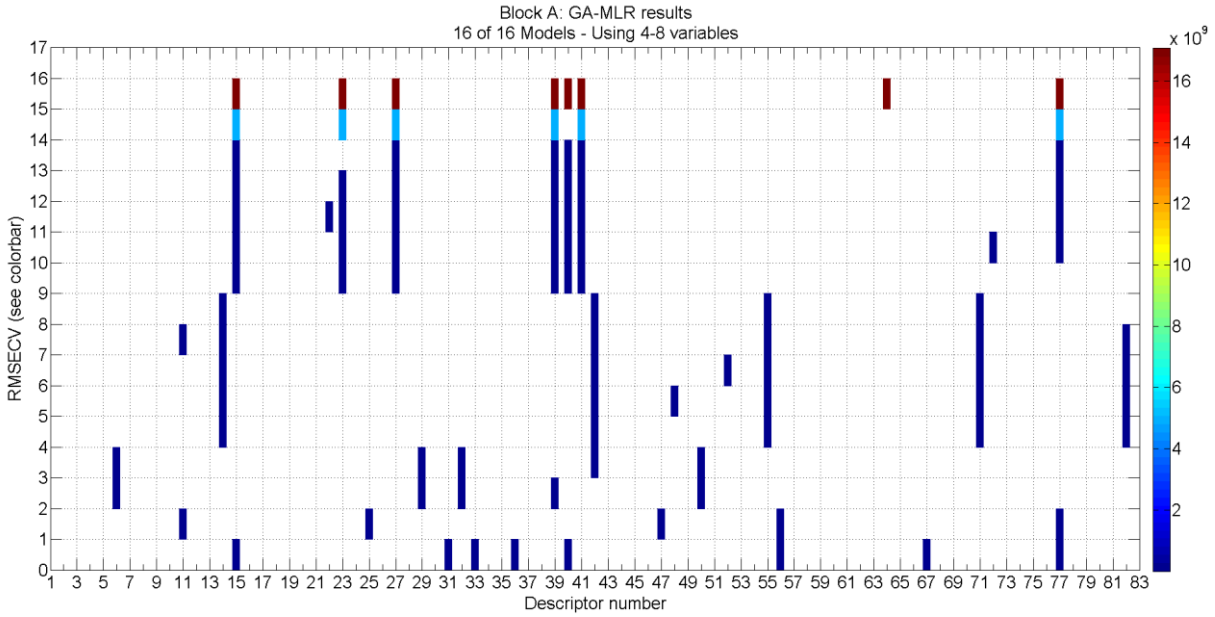


FIGURE 1. GA-MLR RESULTS FOR BLOCK A

Even though GA-MLR variable selection offers ready models, sometimes a manual selection based on the GA-MLR results with smallest $RMSE_{CV}$ may yield an even better set of descriptors. In this case, the final model contained six predictor variables (Figure 2). The green and purple colors mark positive and negative regression coefficients, respectively. The blue and red colors in the 'Model statistics' section represent the quality of the models – blue fields indicate models with RMSE and R^2 values close to optimal, red – models of very poor quality.

#	Descriptor	Regression coefficients																			
		a2-M	C4BP	C1q	Fib	C4	C3	C5	Factor H	IgG	FXII	Haptogl	C1INH	Factor I	Vn	Transferrin	HSA	AT	Hemopexin	ApoAIV	ApoAI
33	nCsp2	-0.001	1.112	0.629	0.075	0.778	0.746	0.478	0.673	0.397	0.397	0.169	0.101	0.513	-0.104	0.207	0.730	0.415	0.371	0.110	0.550
36	P_VSA_LogP_4	0.012	0.027	0.032	0.014	-0.025	-0.030	-0.013	0.034	0.045	0.025	0.016	0.006	-0.006	0.028	0.032	0.026	0.016	0.042	0.002	-0.065
40	P_VSA_m_2	-0.014	-0.138	-0.091	-0.021	-0.104	-0.092	-0.064	-0.088	-0.060	-0.067	-0.025	-0.037	-0.071	0.002	-0.035	-0.100	-0.071	-0.059	-0.043	-0.084
56	P_VSA_s_6	0.035	-0.037	0.016	-0.019	0.000	0.040	0.034	0.010	0.006	0.029	0.031	0.015	0.003	0.056	0.033	-0.007	0.046	0.022	0.024	0.051
67	nHAcc	-0.853	1.245	-0.151	0.320	1.132	0.197	-0.130	-0.189	-0.468	-0.516	-0.752	0.016	0.447	-1.700	-1.045	0.332	-0.718	-0.759	-0.279	0.360
77	ALOGP	-0.187	0.208	-0.004	0.160	0.125	-0.144	-0.172	0.010	-0.048	-0.039	-0.227	0.082	-0.008	-0.189	-0.200	0.185	-0.076	-0.049	0.056	0.019
Model statistics																					
RMSE	calibration	0.11	0.07	0.13	0.06	0.04	0.12	0.07	0.09	0.11	0.09	0.16	0.03	0.03	0.06	0.11	0.12	0.20	0.03	0.07	0.05
	cross-validation	0.64	1.21	0.58	0.30	0.97	0.87	0.62	0.53	0.58	0.44	0.67	0.08	0.54	1.00	0.71	0.66	0.65	0.55	0.19	0.52
R^2	calibration	0.88	0.97	0.92	0.66	0.99	0.92	0.97	0.95	0.92	0.95	0.84	0.99	0.99	0.94	0.92	0.82	0.79	0.99	0.92	0.98
	cross-validation	0.05	0.05	0.13	0.10	0.10	0.13	0.04	0.03	0.12	0.35	0.10	0.96	0.01	0.00	0.17	0.00	0.00	0.15	0.49	0.51

FIGURE 2. MLR MODELING RESULTS – BLOCK A

Those six descriptors were used to estimate each of the 20 biological endpoints, with the only difference being the regression coefficients. E.g.:

log(C1INH) = 0.101 **nCsp2** + 0.006 **P_VSA_LogP_4** - 0.037 **P_VSA_m_2** + 0.015 **P_VSA_s_6** + 0.016 **nHAcc** + 0.082 **ALogP**

log(a2-M) = -0.001 **nCsp2** + 0.012 **P_VSA_LogP_4** -0.014 **P_VSA_m_2** + 0.035 **P_VSA_s_6** - 0.853 **nHAcc** -0.187 **ALogP**

The quality of each individual model varied from very good (C1INH) to quite low. It is not surprising, since the GA-MLR method tries to find descriptor sets with lowest average $RMSE_{CV}$ – it will therefore, over-fit some endpoints and undercompensate its estimation for others.

As can be seen in Figure 2, for most of the proteins, the adsorption levels are proportional to the number of hybridized sp^2 atoms in the monomer/crosslinker molecules (nCsp2), sum of van der Waals surface area with low valence electron availability (P_VSA_s_6), and sum of van der Waals surface area with Ghose-Crippen logP in the range of [-0.25; 0) (P_VSA_LogP_4).

The protein adsorption levels are inversely proportional to the number of acceptor atoms for H-bonds (nHAcc), sum of van der Waals surface area with atomic weight between in the range [10,12) and the Ghose-Crippen logP (ALogP).

BLOCK B: CONTACT ACTIVATION PROTEIN LEVELS

The GA-MLR found 12 unique models with 2 to 5 variables (Figure 3). $RMSE_{CV}$ ranged from 0.443 to 0.619 logarithmic units.

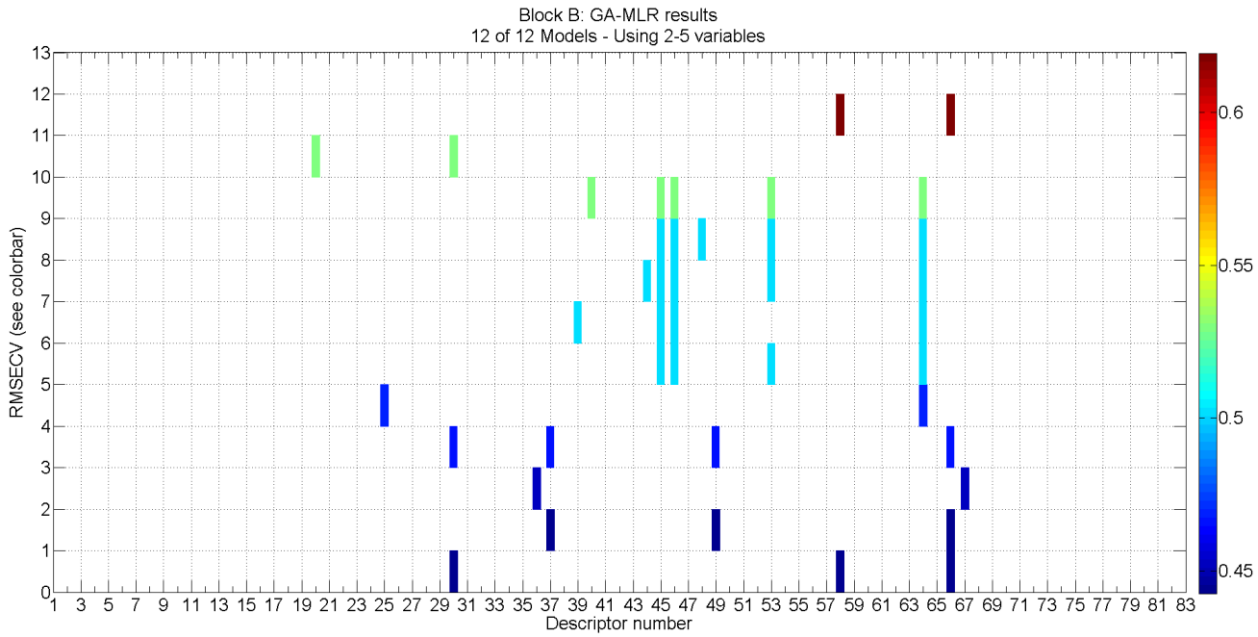


FIGURE 3. GA-MLR RESULTS FOR BLOCK B

The final model contained three molecular descriptors (Figure 4). There are no distinctive trends within the modeling block – the modeled values are split more or less evenly between direct and reverse proportion to the sum of van der Waals surface area with Ghose-Crippen logP in the range of [0; 0.25) (P_VSA_LogP_5), sum of van der Waals surface area with polarizability in the range of [0.4; 1) (P_VSA_LogP_2) and the number of aliphatic tertiary carbon atoms C~sp2 (nR=Ct).

#	Descriptor	Regression coefficients				
		FXIIa-C1 INH	FXIaC1 INH	FXIIa-AT	FXIa-AT	Kall-AT
37	P_VSA_LogP_5	-0.333	-0.248	0.196	0.075	-0.207
49	P_VSA_p_2	0.001	-0.030	0.023	0.003	-0.033
66	nR=Ct	4.357	3.992	-3.283	-1.185	3.363
Model statistics						
RMSE	calibration	0.838	0.183	0.513	0.412	0.399
	cross-validation	1.168	0.358	1.059	0.628	0.616
R^2	calibration	0.379	0.194	0.017	0.077	0.099
	cross-validation	0.025	0.135	0.003	0.000	0.005

FIGURE 4. MLR MODELING RESULTS - BLOCK B

BLOCK C: CONCENTRATIONS OF INFLAMMATION MEDIATORS

The GA-MLR found 19 unique models with 2 to 8 variables (Figure 3). $RMSE_{CV}$ ranged from 0.288 to 7.68×10^9 logarithmic units, which is expected when trying to building a universal model.

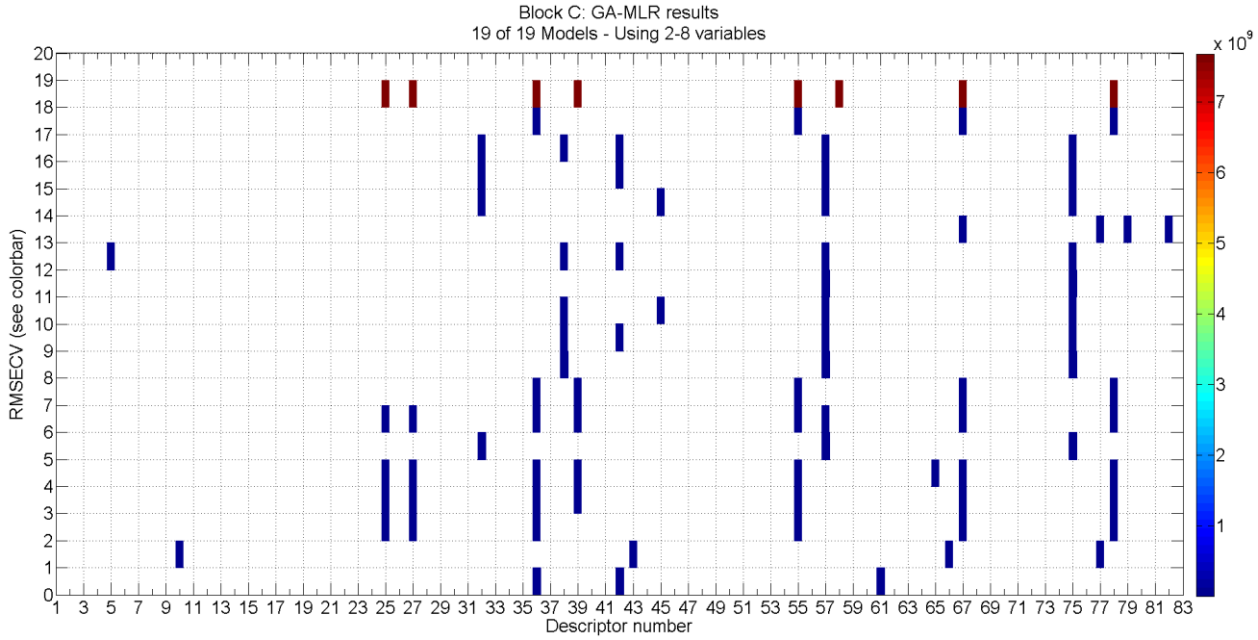


FIGURE 5. GA-MLR RESULTS - BLOCK C

All but one of the immunoprotein concentrations increase proportionally to the sum of atomic polarizabilities scaled on carbon atoms (Sp). The sum of van der Waals surface areas correspondent to van deer Waals volume in the range of [0.5; 1), the number of aliphatic tertiary carbon atoms C~sp2 (nR=Ct) and the squared Ghose-Crippen octanol-water partition coefficient (ALogP2) cause an increase and decrease of protein concentration in an equal amount of cases (Figure 6).

#	Descriptor	Regression coefficients																
		C3a	C5a	TCC	Plt loss %	Gran loss %	TNF-a	IFN-g	IL-6	IL-1ra	IL-10	IL-8	IP-10	MCP-1	MIP-1a	MIP-1b	IL-9	VEGF
10	Sp	0.504	0.114	0.095	0.013	0.266	0.107	0.397	0.078	0.102	0.112	0.102	0.709	-0.073	0.253	0.292	0.267	0.103
43	P_VSA_v_2	-0.076	0.019	0.027	0.025	-0.042	0.003	-0.073	-0.010	0.019	-0.006	0.029	-0.149	0.059	-0.045	-0.031	-0.050	0.011
66	nR=Ct	1.006	-0.617	-0.699	-0.485	0.746	-0.191	1.141	0.559	-0.587	-0.007	-0.516	2.361	-1.042	0.815	0.468	0.832	-0.355
77	ALOGP2	-0.415	-0.007	0.056	0.159	-0.279	0.089	-0.284	0.007	0.191	-0.012	0.262	-0.700	0.324	-0.180	-0.034	-0.195	0.086
Model statistics																		
RMSE	calibration	0.464	0.206	0.236	0.157	0.128	0.184	0.361	0.270	0.235	0.125	0.231	0.595	0.165	0.391	0.285	0.201	0.133
	cross-validation	1.874	0.558	0.448	0.268	0.282	0.249	1.768	0.318	0.291	0.479	0.287	3.617	0.442	0.603	0.796	1.274	0.320
R^2	calibration	0.422	0.763	0.660	0.204	0.800	0.158	0.366	0.590	0.053	0.251	0.434	0.396	0.545	0.179	0.351	0.471	0.192
	cross-validation	0.150	0.000	0.049	0.065	0.251	0.023	0.005	0.476	0.000	0.051	0.359	0.007	0.112	0.043	0.156	0.031	0.090

FIGURE 6. MLR MODELING RESULTS - BLOCK C

BLOCK D: BLOOD CHAMBER IMMUNOMARKER CONCENTRATION MEASUREMENTS

The GA-MLR found 15 unique models with 2 to 5 variables (Figure 7). $RMSE_{CV}$ ranged from 0.37 to 1.38 logarithmic units.

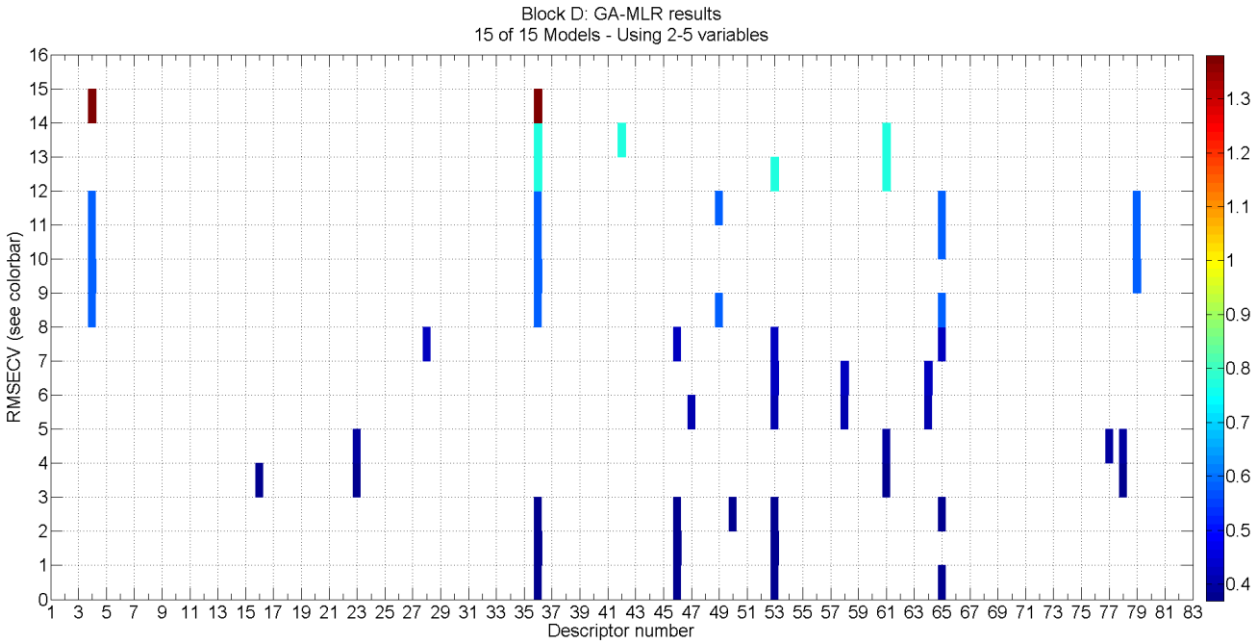


FIGURE 7.GA-MLR RESULTS - BLOCK D

As was the case with Block C, Block D endpoint concentrations are proportional to sum of atomic polarizabilities scaled on carbon atoms (Sp). The sum of van der Waals surface area with Ghose-Crippen logP in the range of [-0.25; 0) (P_VSA_LogP_4), sum of van der Waals surface area with Sanderson electronegativity in the range of [1; 1.1) (P_VSA_e_2) and number of terminal primary C(sp2) (nR=Cp) are ambiguous in their influence (Figure 8).

#	Descriptor	Regression coefficients		
		C3a	TCC	TAT
10	Sp	0.504	0.114	0.095
36	P_VSA_LogP_4	-0.076	0.019	0.027
46	P_VSA_e_2	1.006	-0.617	-0.699
65	nR=Cp	-0.415	-0.007	0.056
Model statistics				
RMSE	calibration	0.174	0.153	0.510
	cross-validation	1.644	0.429	2.585
R^2	calibration	0.572	0.567	0.739
	cross-validation	0.111	0.114	0.152

FIGURE 8. MLR MODELING RESULTS - BLOCK D

CONCLUSIONS

It is possible to create a general model predicting immunological properties of polymers based on their chemical descriptors. However it is not a simple task – the more endpoints are being modeled at once, the worse the accuracy of the predictions.

The general trend in all of the utilized descriptors seems to be pointing towards quantification of the electrostatic properties of the monomer/crossinker molecules as well as their hydrophobicity (logP). There are certain descriptor which featured in models more than once: nR=Ct, P_VSA_LogP_4, Sp – they might make a good starting point in future biopolymer design attempts.

Quite disappointingly, hardly any experimental descriptors were present in the GA-MLR results and none of them have been chosen in any of models.

Perhaps, in the future, an alternative modeling method might be more effective or at least a different method of descriptor calculations.

REFERENCES

1. Engberg, A. E. *et al.* Blood protein-polymer adsorption: Implications for understanding complement-mediated hemoincompatibility. *Journal of biomedical materials research. Part A* 74–84 (2011).doi:10.1002/jbm.a.33030

2. Engberg, A. E. *et al.* Evaluation of the hemocompatibility of novel polymeric materials.

ADDITIONAL ACTIVITIES

11th-15th June 2012 3rd ECO Summer School, Verona, Italy (coordinated by: Università degli Studi di Milano-Bicocca, Milan, Italy)