In silico ADME modeling: QSPR models for the binding of β-lactams to human serum proteins using genetic algorithms

Narayanan Ramamurthi* and Sitarama B Gunturi

Bioinformatics Division, Advanced Technology Center, Tata Consultancy Services
1, Software Units Layout, Madhapur, HYDERABAD – 500 081, INDIA
E-mail: narayananr@atc.tcs.co.in

Dedicated to Professor S. Swaminathan on the occasion of his 80th birthday
(received 30 Aug 04; accepted 23 Nov 04; published on the web 03 Dec 04)

Abstract
Quantitative structure-property relationship (QSPR) models based on *in vitro* serum proteins binding data of 113 diverse drugs and drug-like compounds are reported. For this purpose, two Genetic Algorithm (GA) based approaches GA1 and GA2 along with multiple linear regression (MLR) are employed to exhaustively search and to select multivariate linear equations, starting from a large pool of molecular descriptors (molecular properties or variables). The reported QSPR models are based on combinations of 5 and 6 molecular properties calculated from the 2D chemical structures. Internal (leave-one-out) and external validation tests have demonstrated that these models have excellent predictive power and can be applied to the design new β-lactams class of antibiotics. As the models reported herein, are based on computed properties, they appear as valuable *virtual* screening tools, where selection and prioritisation of candidates is required.

Keywords: β-Lactams, ADMET, QSPR, variable and training set selection, genetic algorithms

Introduction

Binding affinity of new chemical entities (NCE's) to serum proteins is one of the important ADME[1-7] (Absorption, Distribution, Metabolism and Excretion) properties considered in drug discovery and development. Serum proteins are grossly separated into albumin and globulins. Albumin is the protein of highest concentration in the serum (plasma is serum plus clotting proteins). It is a carrier of many small molecules,[8] and is very important in maintaining the oncotic pressure of the blood (that is keeping the fluid from leaking out into the tissues). Binding of a drug to serum proteins in human plasma is a major determinant of its pharmacodynamic behavior (the action of a drug to the body) and the pharmacokinetics of the

 ©ARKAT USA, Inc

drug (the action of the body to the drug) and consequently, can affect the systemic distribution of the drug in several ways. Binding of a drug to plasma protein is a reversible process and is therefore in an equilibrium state. The unbound drug molecules contribute to the pharmacological efficacy and are also susceptible to metabolic reactions.

Acidic drugs are known to bind tightly to human serum albumin (HSA), the major constituent of plasma proteins.[9] HSA has two ligand specific binding sites[10,11] namely, site-I and site-II. The ligand selectivity is comparatively broader for these two sites, allowing a range of drug molecules to bind at these sites. This broad selectivity is considered to be a result of the significant allosteric effects in HSA[12] and drug molecules can also interact nonspecifically with HSA. In addition, alpha 1-acid glycoproteins (AGP) and lipoproteins, constituents of plasma proteins, can also interact with drugs. Although the amount of AGP in the plasma is far smaller than that of HSA, it interacts strongly with basic and neutral drugs[9] in addition to some acidic drugs.[13] Binding to lipoproteins is considered non-specific due to hydrophobic interactions.[9]

Prediction of the serum protein binding percentage is more difficult than that of other ADME factors because, this is a composite parameter made up of the sum of interactions with multiple proteins, each with a different affinity. The prediction is further complicated by having to include both specific and nonspecific binding as well as significant allosteric interactions. Given the importance of drug binding to serum proteins[13], it should be extremely useful to develop quantitative structure-property relationships to predict the binding affinity to serum proteins, applicable to the whole medicinal chemical space. Computational models of this type are useful because they rationalize a large number of experimental observations and therefore allow us to save time and money in the drug design process. However, there are very few published reports [14–17] on the prediction of binding affinity of drugs and drug-like compounds to serum proteins and further, these predictive models are based on molecular properties chosen out of experience; consequently, the whole space of molecular properties is not explored. This prompted us to perform a study aimed at (i) gaining further insights into the molecular properties that influence serum protein binding (ii) develop improved mathematical models for serum protein binding and (iii) demonstrate their utility in drug discovery and development.

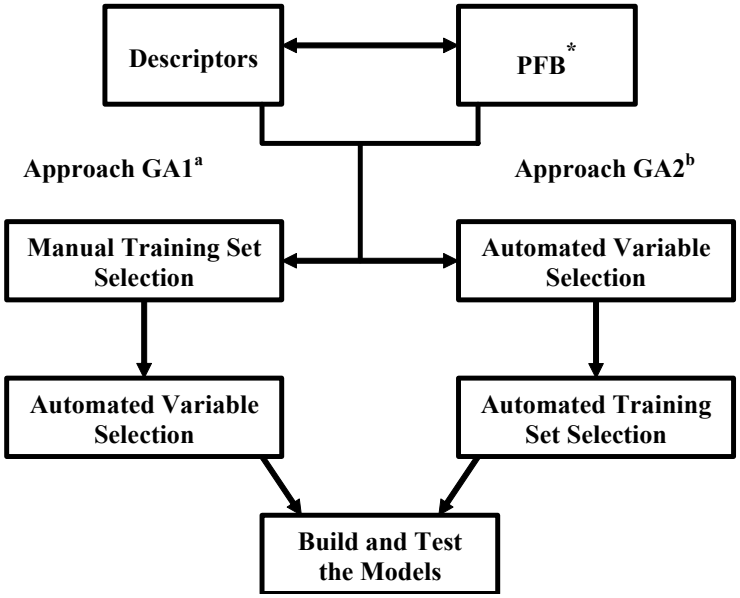
In this paper, we describe the application of two (GA) based approaches GA1 and GA2 to derive novel QSPR models to predict the binding affinity of a specific class of drugs, β-lactams, to human serum protein, for the first time using a large set of molecular properties, to the best of our knowledge. The predictive performance of the QSPR models on external data sets taken from literature is illustrated.

Results and Discussions

Methodology

The methodology adopted in the data analysis is depicted in Fig. 1. The QSPR models reported in the literature on protein binding are principally based on manually selected descriptors, where

experience and intuition are the key factors for success. However, in this approach, the whole molecular properties space is not explored and hence, it is highly desirable to explore all possible descriptor combinations in model generation. Our methodology provides many advantages and some of them are as follows: 1) it is based on a larger training set 2) it is also based on the largest number of molecular properties reported as of date and hence can provide an optimal solution 3) we have set the inter correlation coefficient between descriptors as 0.75 and thereby the selected descriptors are expected to be independent.



[a] Variable selection based on 100 observations

[b] Variable selection based on 113 observations

[*] PFB – Percentage Fraction bound

Figure 1. Data Analysis methodology.

QSPR models using GA1

In approach GA1, 100 compounds of training set their 322 descriptors and I are used to select the best three of the 5 and 6 variable combinations, keeping the experimental percentage fraction bound (PFB) values, as the dependent variable. As many of the 322 descriptors may be correlated, it is desirable to find variables that are not correlated and construct models using only these variables. It is well known that QSPR models based on un-correlated variables will improve the predictive performance and hence, we performed variable selection, setting the inter-variable correlation below 0.75, as this is expected to discard descriptors with a inter-correlation coefficient above 0.75. The selected combinations are shown in Table 1 and the inter-descriptor correlations are shown in Table 2.

Table 1. Variables selected using GA1

No of Selected Desc.	Model No	Selected Descriptors	R	F	Q	SE
5	G1	158, 235, 268, 284, 304	0.8935	74.45	0.8723	12.724
	G2	158, 235, 247, 284, 304	0.8933	74.28	0.8720	12.736
	G3	199, 231, 268, 284, 304	0.8926	74.68	0.8728	12.776
6	G4	158, 193, 231, 247, 284, 304	0.9074	72.25	0.8888	11.973
	G5	158, 193, 235, 268, 284, 304	0.9072	72.03	0.8884	11.988
	G6	158, 193, 235, 247, 284, 304	0.9068	71.73	0.8881	12.009

where
158 – Autocorrelation descriptor (Broto-Moreau) weighted by atomic masses – Order 2
193 – Autocorrelation descriptor (Moran) weighted by Pauling electronegativity – Order1
199 – Autocorrelation descriptor (Moran) weighted by atomic van der Waals radius – Order 1
231 – Mean information content on the distance equality
235 – Mean information content on the edge distance equality
247 – Atomic Type Electrotopological state index (E-state) - SsNH2
268 – Hydrogen Electrotopological state index (E-state) – SHsNH2
284 – Atomic-Level-Based AI topological descriptors – AIsssCH
304 – AlogP98

The best 5 descriptors model G1 is based on variables 158, 235, 268, 284 and 304 with a correlation coefficient, R of 0.8935 and cross-validated correlation coefficient, Q of 0.8723. The correlation coefficients of the other two models G2 and G3 are 0.8933 and 0.8926 respectively.

Significantly, the descriptors 284 (Atomic-Level-Based AI topological descriptor – AIsssCH) and 304 (AlogP98) are selected in all the best five variable models G1, G2 and G3. Descriptors 231 (Mean information content on the distance equality) in model G3 and 235 (Mean information content on the edge distance equality) in models G1 and G2 have high correlation (R = 0.9945), indicating that they provide the same information and significance to protein binding property of chemical compounds. Similarly, descriptor 247 (Atomic Type Electrotopological state index – SsNH2) of model G2 and 268 (Hydrogen Electrotopological state index – SHsNH2) of models G1 and G3 have inter-correlation coefficient R = 0.9992. It is interesting to note that descriptor 199 (Autocorrelation descriptor (Moran) weighted by atomic van der Waals radius – Order 1) of model G3 has low correlation with the rest of the descriptors in the five descriptors combinations.

The best six descriptors combination G4 is based on the variables 158, 193, 231, 247, 284 and 304 with a correlation coefficient of 0.9074, and cross-validated correlation coefficient, Q of 0.8888. Combinations of descriptors in models G5 and G6 are the same as those in models G1 and G2 with an additional new descriptor 193. Similarly, model G4, is the same as model G3,

with an additional descriptor 158. As there is no significant improvement in R and Q values between the five and six descriptor combinations, we did not proceed further to select models with higher number of descriptors and this prompted us to believe that the six descriptors model is an optimal solution to the prediction of protein binding based on the 332 descriptors.

Table 2. Correlation matrix of the selected variables

	158	193	199	231	235	247	268	284	304
158	1	0.1064	0.0358	0.0239	0.0244	0.1187	0.1148	0.0148	0.0646
193		1	0.9653	0.0742	0.1145	0.0808	0.0833	0.1417	0.0696
199			1	0.1679	0.2076	0.0556	0.0584	0.1572	0.0678
231				1	0.9945	0.2107	0.2150	0.1149	0.1799
235					1	0.1971	0.2013	0.1069	0.1625
247						1	0.9992	0.0935	0.5238
268							1	0.0930	0.5216
284								1	0.3260
304									1

The regression equations derived by performing MLR on the above variable combinations are given below:

PFB = -173.467 + 0.295338*Desc158 + 58.488900*Desc235 – 9.692590*Desc268 – 1.311910*Desc284 + 16.173280*Desc304 **(G1)**

PFB = -172.390 + 0.293301*Desc158 + 58.283730*Desc235 – 3.171120*Desc247 – 1.310580*Desc284 + 16.154600*Desc304 **(G2)**

PFB = -118.739 + 101.1195*Desc199 + 60.530060*Desc231 – 10.98010*Desc268 – 1.148630*Desc284 + 15.351260*Desc304 **(G3)**

PFB = -188.255 + 0.326134*Desc158 + 115.3533*Desc193 + 61.72461*Desc231 – 3.460850*Desc247 – 1.167410*Desc284 + 15.285760*Desc304 **(G4)**

PFB = -160.184 + 0.326561*Desc158 + 105.4771*Desc193 + 55.176880*Desc235 – 10.44420*Desc268 – 1.153940*Desc284 + 15.180520*Desc304 **(G5)**

PFB = -159.096 + 0.324263*Desc158 + 104.9271*Desc193 + 54.967900*Desc235 – 3.412100*Desc247 – 1.153440*Desc284 + 15.170140*Desc304 **(G6)**

Model validation

The above models, G1 to G6 are validated using Leave-One-Out approach. The results of LOO cross-validations are given in Table 3. A plot of cross-validated PFB values versus the experimental PFB values of compounds in training set using models G1–G6 are shown in Fig. 2 – Fig. 7 respectively. Based on the cross-validated results, we believe that the models G1 and G4 are the best five and six descriptors models, based on their overall predictive power and can be used for *virtual* screening of β-lactam analogs.

Table 3. Results of LOO cross validation

Serial No	Compound Name	Expt PFB	Predicted PFB G1	G2	G3	G4	G5	G6
1	Penicillin_31	12.00	-5.77	-6.04	-2.99	-4.46	-3.34	-3.60
2	Penicillin_2	15.00	8.52	8.72	8.63	12.93	9.12	9.32
3	Penicillin_32	16.80	23.07	23.46	27.93	27.07	26.96	27.37
4	Penicillin_9	20.00	32.33	32.42	28.15	28.19	26.87	27.00
5	Penicillin_11	25.00	46.54	46.59	40.92	41.98	40.12	40.20
6	Penicillin_30	26.00	37.29	37.27	41.41	42.06	42.14	42.11
7	Penicillin_6	28.00	35.17	35.25	27.80	29.45	27.50	27.62
8	Penicillin_72	32.00	35.93	35.05	31.85	35.40	39.12	38.18
9	Penicillin_12	33.00	10.68	10.54	11.89	13.08	11.22	11.08
10	Penicillin_34	38.00	32.05	32.17	31.91	29.47	29.26	29.42
11	Penicillin_35	42.00	51.69	53.61	57.47	60.28	56.80	58.86
12	Penicillin_8	47.00	54.31	54.33	52.18	51.67	50.08	50.12
13	Penicillin_37	53.20	50.20	50.24	55.88	54.20	55.01	55.02
14	Penicillin_28	55.00	43.25	43.19	43.46	45.66	45.75	45.68
15	Penicillin_76	57.00	68.17	68.19	68.66	64.18	64.99	65.03
16	Penicillin_71	58.00	74.93	74.91	69.46	74.71	77.50	77.46
17	Penicillin_7	58.80	63.31	63.32	58.26	58.70	56.71	56.74
18	Penicillin_73	59.00	65.42	65.40	59.04	61.60	64.18	64.15
19	Penicillin_27	60.00	35.75	35.77	40.43	38.18	38.21	38.23
20	Penicillin_46	60.00	58.84	58.85	80.35	76.81	74.69	74.61
21	Penicillin_77	61.70	71.60	71.62	73.04	68.10	68.65	68.69
22	Penicillin_38	62.00	56.18	56.21	56.63	53.54	54.79	54.82
23	Penicillin_36	63.00	67.78	67.77	71.67	67.60	68.62	68.60
24	Penicillin_45	65.00	60.86	60.88	62.15	61.68	61.88	61.88
25	Penicillin_13	66.20	51.62	51.26	53.55	53.08	51.33	50.96
26	Penicillin_19	68.00	67.72	67.74	73.01	70.32	71.02	71.02
27	Penicillin_10	74.00	63.90	63.92	61.96	60.89	59.40	59.44
28	Penicillin_79	74.50	77.66	77.67	79.81	74.79	74.78	74.81
29	Penicillin_18	77.00	68.21	68.22	65.05	60.83	60.70	60.75
30	Penicillin_22	78.00	73.33	73.32	74.06	75.95	76.67	76.64
31	Penicillin_62	80.00	78.35	78.32	78.27	78.68	78.53	78.49
32	Penicillin_67	80.00	79.18	79.20	83.53	78.20	78.62	78.64
33	Penicillin_54	81.50	80.08	80.04	78.50	79.69	79.24	79.20
34	Penicillin_48	81.50	71.79	71.78	70.70	68.50	69.63	69.62

Table 3. Continued

35	Penicillin_66	82.10	87.82	87.82	96.01	90.95	91.23	91.22
36	Penicillin_29	82.20	55.81	55.58	53.27	58.43	58.07	57.83
37	Penicillin_21	82.50	83.79	83.72	78.25	87.15	87.91	87.81
38	Penicillin_44	83.00	77.16	77.14	75.59	78.09	77.62	77.59
39	Penicillin_53	83.50	77.80	77.78	76.20	77.36	77.09	77.06
40	Penicillin_75	83.60	82.15	82.05	69.35	80.07	81.06	80.96
41	Penicillin_41	84.00	83.47	83.43	77.80	80.67	81.28	81.25
42	Penicillin_61	84.00	76.22	76.20	76.10	76.47	76.50	76.47
43	Penicillin_64	86.00	80.88	80.89	88.55	84.11	84.50	84.49
44	Penicillin_50	86.10	79.68	79.67	79.16	76.49	77.53	77.52
45	Penicillin_14	87.00	62.16	62.20	65.11	63.82	64.59	64.62
46	Penicillin_39	88.00	79.73	79.67	72.87	78.02	78.44	78.38
47	Penicillin_68	89.30	85.43	85.43	90.42	85.07	84.84	84.85
48	Penicillin_74	89.70	100.98	100.71	73.76	99.91	101.25	100.96
49	Penicillin_16	91.00	94.83	94.82	102.85	97.82	98.05	98.02
50	Penicillin_52	91.50	76.76	76.74	74.68	75.84	76.08	76.05
51	Penicillin_42	92.00	88.76	88.71	85.05	86.01	86.06	86.01
52	Penicillin_3	92.40	98.59	98.54	100.93	100.72	97.75	97.71
53	Penicillin_49	92.50	83.33	83.32	82.12	79.15	80.36	80.35
54	Penicillin_4	93.30	104.11	104.11	109.50	105.94	103.81	103.82
55	Penicillin_23	94.00	88.26	88.21	86.31	91.57	91.34	91.27
56	Penicillin_24	94.00	90.69	90.63	88.75	94.08	93.66	93.58
57	Penicillin_69	94.70	92.78	92.73	96.91	92.00	92.53	92.47
58	Penicillin_63	95.20	107.51	107.42	100.78	104.15	104.53	104.46
59	Penicillin_65	95.60	86.28	86.27	93.59	89.21	89.59	89.56
60	Penicillin_57	96.00	99.16	99.08	92.07	95.83	96.20	96.13
61	Penicillin_43	96.50	101.09	101.00	93.13	98.90	98.26	98.18
62	Penicillin_25	97.00	105.24	105.14	100.20	109.09	108.09	107.97
63	Penicillin_59	97.00	100.84	100.75	94.36	98.18	97.81	97.73
64	Penicillin_51	97.20	88.99	88.97	93.31	88.16	89.05	89.02
65	Penicillin_58	97.40	102.84	102.74	96.33	100.22	99.70	99.62
66	Penicillin_70	97.40	88.17	88.14	92.68	87.66	88.13	88.09
67	Amoxicillin	18.00	25.48	25.77	30.34	29.44	29.24	29.55
68	Bacampicillin	20.00	60.19	60.20	53.54	50.92	50.73	50.81
69	Piperacillin	30.00	52.83	52.83	51.45	48.46	48.48	48.48
70	Methicillin	39.00	62.50	62.54	58.91	55.26	55.36	55.44
71	Carbenicillin	50.00	53.21	53.25	59.95	57.98	58.64	58.65
72	Penicillin_G	60.00	70.57	70.58	74.01	72.42	73.02	73.02
73	Ticarcillin	65.00	49.95	49.97	50.79	52.82	55.34	55.32

Table 3. Continued

74	Penicillin_V	80.00	73.57	73.55	70.85	69.33	70.22	70.21
75	Nafcillin	89.00	85.44	85.44	90.43	85.08	84.85	84.86
76	Oxacillin	92.00	79.25	79.24	92.14	87.53	87.70	87.65
77	Cloxacillin	95.00	92.71	92.67	103.21	100.92	100.71	100.63
78	Meropenem	2.00	-15.22	-15.07	-3.89	-10.67	-10.90	-10.78
79	Cephalexin	14.00	35.21	35.29	40.50	36.01	36.62	36.71
80	Cefadroxil	20.00	34.39	34.70	40.92	37.08	36.48	36.81
81	Cefepime	20.00	40.05	39.49	32.77	31.62	35.03	34.44
82	Ceftazidime	21.00	53.06	52.78	52.60	51.82	51.83	51.53
83	Cefaclor	25.00	37.83	37.87	38.57	38.86	39.52	39.56
84	Loracarbef	25.00	28.91	29.05	35.94	29.85	30.46	30.61
85	Cefpodoxime	27.00	48.72	48.16	34.20	31.43	32.13	31.52
86	Ceftizoxime	28.00	37.20	36.73	31.73	34.73	36.36	35.86
87	Ceftibutin	30.00	39.85	39.51	35.16	38.26	38.90	38.54
88	Cefotaxime	36.00	46.42	45.98	39.26	40.21	40.29	39.84
89	Cefprozil	40.00	45.91	46.21	53.71	49.15	47.73	48.05
90	Cephapirin	62.00	78.76	78.70	72.37	73.24	72.87	72.82
91	Cefixime	67.00	43.64	43.40	42.85	44.82	43.97	43.73
92	Cefmetazole	70.00	91.13	91.04	88.21	88.43	87.64	87.55
93	Cefoxitin	73.00	68.05	69.93	51.79	55.13	54.69	57.08
94	Cefmandole	74.00	79.07	79.00	93.10	88.15	88.69	88.55
95	Ceforanide	81.00	65.46	64.51	77.32	72.93	71.61	70.56
96	Cefotetan	83.00	72.61	74.54	66.97	83.57	79.90	81.97
97	Cefazolin	89.00	76.28	76.15	75.13	82.53	84.27	84.07
98	Cefoperazone	91.00	78.32	78.22	80.46	75.86	75.40	75.28
99	Ceftriaxone	93.00	46.15	45.84	45.55	44.77	44.86	44.54
100	Cefonicid	98.00	74.02	73.92	83.96	84.46	82.64	82.47

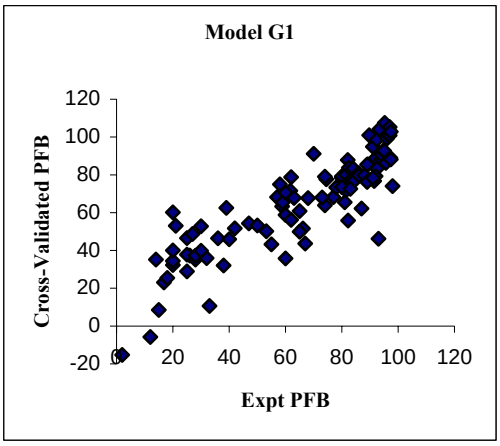


Figure 2. Cross-Validated PFB using G1.

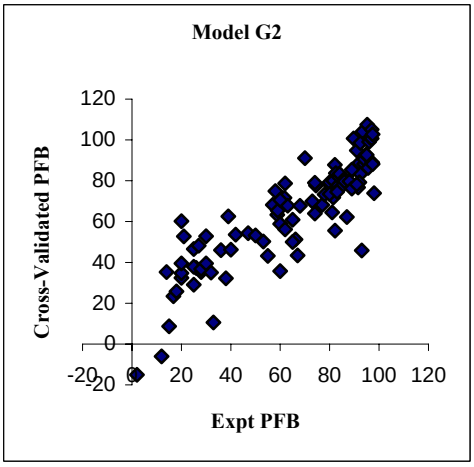


Figure 3. Cross-Validated PFB using G2.

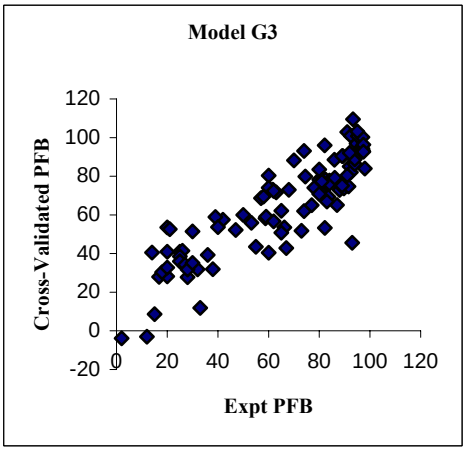


Figure 4. Cross-Validated PFB using G3.

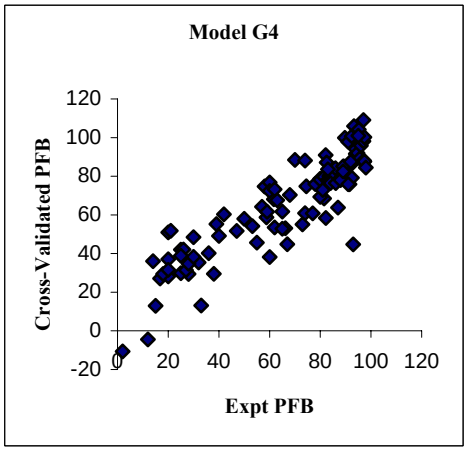


Figure 5. Cross-Validated PFB using G4.

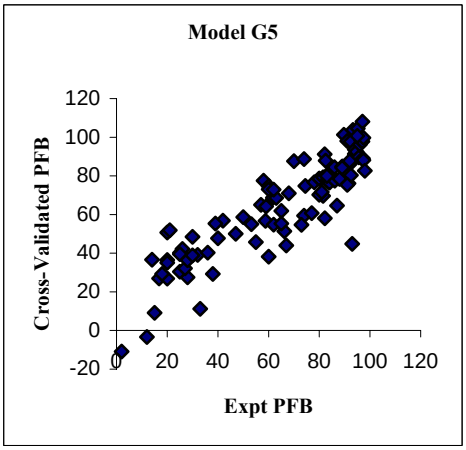


Figure 6. Cross-Validated PFB using G5.

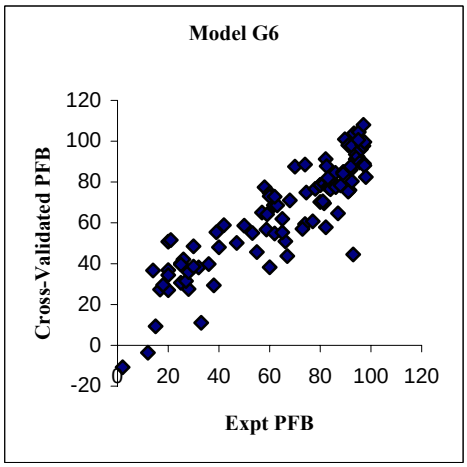


Figure 7. Cross-Validated PFB using G6.

External-validations of models G1 to G6

The manually selected members of test set I are used to study the predictive power of the models G1 – G6. The predicted PFB values of these test compounds are given in Table 4.

Table 4. Results of External Validation

Serial No	Compound Name	Expt PFB	Predicted PFB G1	G2	G3	G4	G5	G6
101	Penicillin_5	7.20	21.25	21.36	13.72	15.26	13.61	13.76
102	Penicillin_78	69.70	74.25	74.26	75.53	71.08	71.24	71.27
103	Penicillin_40	83.00	71.36	71.35	69.09	69.18	69.57	69.56
104	Penicillin_20	93.60	75.72	75.76	79.71	77.70	75.86	75.91
105	Penicillin_55	96.00	85.82	85.77	81.79	82.67	83.35	83.31
106	Penicillin_60	86.00	75.42	75.40	74.91	75.25	75.73	75.71
107	Penicillin_56	94.80	87.54	87.49	82.58	85.49	85.23	85.18
108	Ampicillin	18.00	26.36	26.42	29.40	28.06	29.17	29.23
109	Dicloxacillin	96.00	105.53	105.47	112.88	112.88	112.42	112.32
110	Cephradine	14.00	19.44	19.38	26.12	21.35	22.25	22.18
111	Cephalothin	71.00	81.77	81.73	76.74	77.72	78.63	78.59
112	Cefdinir	65.00	46.46	46.16	49.91	52.82	52.90	52.55
113	Cefuroxime	33.00	48.02	49.77	41.47	39.14	38.07	40.01

QSPR models using GA2

It is well known that manual selection of training and test sets followed by QSPR model generation is arduous and even impractical at times, particularly, when the data set is too large. Automated training set selection procedures are expected to be useful in such cases. Abraham et al[30] have employed automated training set selection successfully for the generating prediction models for absorption. They have employed Kennath and Stone algorithm[31], which selects compounds based on the maximin (maximizing the minimum distances) principle, thereby satisfies the diverse distribution of molecular descriptors. We have chosen to study the application of GA based automated training set selection for the first time to the best of our knowledge, in the development of prediction models for PFB. In this study, the selection of the compounds is based on the best correlation coefficient of MLR, as the models with high value of R are expected to possess good predictive power.

In this approach, same data set of 113 compounds and 322 molecular descriptors are used, as in the case of the approach GA1. However, the model generation involves the following two steps: 1) variable selection using the full data set of 113 compounds as compared to 100, used in the case of GA1 and 2) based on the best variable combinations of step 1, the best training set containing 100 members are selected by GA2.

The best three of the 5 and 6 variable combinations selected by GA2 are shown in Table 5 and the inter-descriptor correlations are shown in Table 6. The 5 variable combinations of V1, V2 and V3 are the same as that of G1, G2 and G3 respectively of GA1. The best six variable combination, V4, is based on the variables 133, 199, 235, 268, 284 and 304 with a correlation coefficient, R of 0.8996. It is a new descriptors combination and such a result is anticipated, as the training set size is different from that in approach GA1. Similarly, the second best six variable combination V5 also has a new descriptor 35 (mean square distance index) that contains the same information as descriptors 231 and 235. The third best six variable combination V6 is the same as G5 selected by GA1, but has different statistical numbers.

Table 5. Variables selected using GA2

Sample Size	No of Input Desc.	No of Selected Desc.	Descriptor Combination	Selected Descriptors	R
113	322	5	V1	158, 235, 268, 284, 304	0.8996
			V2	158, 235, 247, 284, 304	0.8991
			V3	199, 231, 268, 284, 304	0.8987
113	322	6	V4	133, 199, 235, 268, 284, 304	0.9128
			V5	35, 158, 193, 268, 284, 304	0.9126
			V6	158, 193, 235, 268, 284, 304	0.9124

where
35 – Mean square distance index
133 – Bound charge index (J) – Order 1
158 – Autocorrelation descriptor (Broto-Moreau) weighted by atomic masses of Order 2
193 – Autocorrelation descriptor (Moran) weighted by Pauling electro-negativity – Order 1
199 – Autocorrelation descriptor (Moran) weighted by atomic van der Waals radius – Order 1
231 – Mean information content on the distance equality
235 – Mean information content on the edge distance equality
245 – Atomic Type Electrotopological state index (E-state) - SsNH2
268 – Hydrogen Electro-topological state index (E-state) - SHsNH2
284 – Atomic-Level-Based AI topological descriptors – AIsssCH
304 – AlogP98

Table 6. Correlation matrix of the selected variables

	35	133	158	193	199	231	235	247	268	284	304
35	1	0.4054	0.0179	0.0814	0.1711	0.9883	0.979	0.1829	0.188	0.1248	0.1614
133		1	0.329	0.1823	0.2366	0.4113	0.430	0.2267	0.2272	0.2936	0.349
158			1	0.0972	0.036	0.0259	0.025	0.1434	0.1407	0.0163	0.079
193				1	0.9678	0.0941	0.135	0.1055	0.1038	0.0991	0.1052
199					1	0.183	0.222	0.0786	0.078	0.115	0.1092
231						1	0.994	0.2004	0.2053	0.1251	0.1442
235							1	0.19	0.1948	0.1201	0.1261
247								1	0.9991	0.1024	0.5363
268									1	0.104	0.5352
284										1	0.3462
304											1

The best 5 and 6 variable combinations having high correlation to PFB, V1 and V4 are considered for the selection of best training sets using GA2. The data set consisting of 113 compounds and the best 5 and 6 variable combinations are used as the input data for GA2 to select the training set of 100 compounds based on the correlation coefficient R. The two best training sets, II and III are presented in Table 7. The compounds that are not selected as training set members in the case of models T1 and T4 constitute test sets II and III respectively.

Table 7. Training set selected using GA2

Descriptor Combination	Model No	Selected members of the training set	R	Q	F
V1	T1	**II:** 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 53, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 96, 97, 98, 99, 100, 101, 104, 105, 106, 107, 109, 110, 111,113	0.9438	0.9323	153. 260

Table 7. Continued

V4	T4	**III:** 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 52, 53, 54, 55, 56, 57, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 71, 72, 73, 74, 75, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 93, 94, 95, 96, 97, 98, 99, 100, 105, 106, 107, 108, 109, 110, 111, 113	0.9473	0.9358	135.572

The compounds selected as training sets II and III for the models T1 and T4 with their corresponding PFB values are used for model generation using MLR and the resulting regression equations are given below:

PFB = -196.76 + 0.294636*Desc158 + 60.49762*Desc235 – 9.45805*Desc268 – 0.614930*Desc284 + 16.68958*Desc304 **(T1)**

N = 100, R = 0.9438, SE = 9.747849, Q = 0.9323

PFB = -195.554 + 160.736*Desc133 + 80.60265*Desc199 + 72.65097*Desc235 – 12.0217*Desc268 – 1.296300*Desc284 + 15.07384*Desc304 **(T4)**

N = 100, R = 0.9473, SE = 9.401961, Q = 0.9358

Model validation

The members of the training sets II and III, selected by GA2 (Table 7), are cross-validated by the LOO method. The LOO cross-validated PFB values of compounds in training sets II and III are given in Table 8. A plot of cross-validated PFB values versus the observed PFB values using T1 and T4 is shown in Fig. 8 and Fig. 9 respectively.

Table 8. Results of LOO cross-validation of compounds in training sets II and III

Training set II – Model T1				Training set III – Model T4			
Comp No.	Comp Name	Expt PFB	Calc PFB	Comp No.	Comp Name	Expt PFB	Calc PFB
1	Penicillin_5	7.20	13.94	1	Penicillin_5	7.20	13.84
2	Penicillin_31	12.00	-5.03	2	Penicillin_31	12.00	-4.07
3	Penicillin_2	15.00	-4.57	3	Penicillin_2	15.00	5.86
4	Penicillin_32	16.80	23.85	4	Penicillin_32	16.80	29.47
5	Penicillin_9	20.00	30.85	5	Penicillin_9	20.00	26.74

Table 8. Continued

6	Penicillin_11	25.00	39.99	6	Penicillin_11	25.00	39.33
7	Penicillin_30	26.00	38.46	7	Penicillin_30	26.00	42.84
8	Penicillin_6	28.00	27.18	8	Penicillin_6	28.00	26.38
9	Penicillin_72	32.00	37.02	9	Penicillin_72	32.00	35.04
10	Penicillin_12	33.00	9.83	11	Penicillin_34	38.00	33.55
11	Penicillin_34	38.00	34.17	12	Penicillin_35	42.00	64.24
13	Penicillin_8	47.00	57.54	13	Penicillin_8	47.00	54.28
14	Penicillin_37	53.20	50.07	14	Penicillin_37	53.20	54.02
15	Penicillin_28	55.00	44.48	15	Penicillin_28	55.00	45.06
16	Penicillin_76	57.00	65.66	16	Penicillin_76	57.00	65.15
17	Penicillin_71	58.00	68.98	17	Penicillin_71	58.00	74.43
18	Penicillin_7	58.80	58.35	18	Penicillin_7	58.80	56.48
19	Penicillin_73	59.00	66.74	19	Penicillin_73	59.00	62.70
20	Penicillin_27	60.00	36.91	20	Penicillin_27	60.00	41.88
21	Penicillin_46	60.00	61.57	21	Penicillin_46	60.00	79.44
22	Penicillin_77	61.70	69.76	22	Penicillin_77	61.70	70.68
25	Penicillin_45	65.00	62.06	23	Penicillin_38	62.00	53.80
26	Penicillin_13	66.20	53.66	24	Penicillin_36	63.00	68.41
27	Penicillin_19	68.00	68.12	25	Penicillin_45	65.00	64.10
28	Penicillin_78	69.70	72.60	26	Penicillin_13	66.20	48.95
29	Penicillin_10	74.00	64.54	27	Penicillin_19	68.00	70.90
30	Penicillin_79	74.50	76.55	28	Penicillin_78	69.70	70.22
31	Penicillin_18	77.00	66.27	29	Penicillin_10	74.00	57.58
32	Penicillin_22	78.00	73.73	30	Penicillin_79	74.50	75.60
33	Penicillin_62	80.00	81.04	31	Penicillin_18	77.00	66.71
34	Penicillin_67	80.00	76.99	32	Penicillin_22	78.00	71.97
35	Penicillin_54	81.50	77.73	33	Penicillin_62	80.00	86.34
36	Penicillin_48	81.50	73.31	34	Penicillin_67	80.00	78.91
37	Penicillin_66	82.10	85.26	35	Penicillin_54	81.50	86.24
39	Penicillin_21	82.50	84.22	36	Penicillin_48	81.50	73.63
40	Penicillin_40	83.00	72.44	37	Penicillin_66	82.10	92.32
41	Penicillin_44	83.00	79.43	39	Penicillin_21	82.50	75.70
42	Penicillin_53	83.50	75.29	40	Penicillin_40	83.00	70.78
43	Penicillin_75	83.60	84.42	41	Penicillin_44	83.00	79.19
44	Penicillin_41	84.00	85.38	42	Penicillin_53	83.50	83.24
45	Penicillin_61	84.00	78.70	43	Penicillin_75	83.60	76.94
46	Penicillin_60	86.00	77.54	44	Penicillin_41	84.00	77.46
47	Penicillin_64	86.00	77.65	45	Penicillin_61	84.00	83.53
48	Penicillin_50	86.10	81.69	46	Penicillin_60	86.00	78.77

Table 8. Continued

50	Penicillin_39	88.00	81.04	47	Penicillin_64	86.00	83.38
51	Penicillin_68	89.30	83.99	48	Penicillin_50	86.10	79.93
52	Penicillin_74	89.70	103.91	50	Penicillin_39	88.00	74.23
53	Penicillin_16	91.00	93.45	52	Penicillin_74	89.70	80.74
55	Penicillin_42	92.00	91.52	53	Penicillin_16	91.00	96.20
56	Penicillin_3	92.40	95.57	54	Penicillin_52	91.50	78.24
57	Penicillin_49	92.50	80.82	55	Penicillin_42	92.00	93.07
58	Penicillin_4	93.30	101.63	56	Penicillin_3	92.40	97.13
60	Penicillin_23	94.00	89.58	57	Penicillin_49	92.50	94.47
61	Penicillin_24	94.00	92.21	59	Penicillin_20	93.60	84.51
62	Penicillin_69	94.70	98.05	60	Penicillin_23	94.00	89.06
63	Penicillin_56	94.80	89.72	61	Penicillin_24	94.00	92.12
64	Penicillin_63	95.20	111.40	62	Penicillin_69	94.70	102.06
65	Penicillin_65	95.60	83.66	63	Penicillin_56	94.80	85.87
66	Penicillin_55	96.00	88.03	64	Penicillin_63	95.20	104.00
67	Penicillin_57	96.00	102.55	65	Penicillin_65	95.60	93.90
68	Penicillin_43	96.50	104.47	66	Penicillin_55	96.00	85.52
69	Penicillin_25	97.00	107.73	67	Penicillin_57	96.00	97.25
70	Penicillin_59	97.00	104.52	68	Penicillin_43	96.50	97.82
71	Penicillin_51	97.20	92.47	69	Penicillin_25	97.00	104.79
72	Penicillin_58	97.40	106.71	71	Penicillin_51	97.20	87.15
73	Penicillin_70	97.40	92.76	72	Penicillin_58	97.40	105.54
74	Amoxicillin	18.00	26.49	73	Penicillin_70	97.40	92.99
75	Ampicillin	18.00	26.96	74	Amoxicillin	18.00	32.59
77	Piperacillin	30.00	60.85	75	Ampicillin	18.00	26.20
78	Methicillin	39.00	59.01	77	Piperacillin	30.00	53.65
79	Carbenicillin	50.00	53.83	78	Methicillin	39.00	56.70
80	Penicillin_G	60.00	65.84	79	Carbenicillin	50.00	57.66
81	Ticarcillin	65.00	49.34	80	Penicillin_G	60.00	75.28
82	Penicillin_V	80.00	70.02	81	Ticarcillin	65.00	51.64
83	Nafcillin	89.00	84.01	82	Penicillin_V	80.00	73.19
84	Oxacillin	92.00	77.38	83	Nafcillin	89.00	83.86
85	Cloxacillin	95.00	91.79	84	Oxacillin	92.00	87.81
86	Dicloxacillin	96.00	105.97	85	Cloxacillin	95.00	100.02
87	Meropenem	2.00	3.72	86	Dicloxacillin	96.00	112.20
88	Cephalexin	14.00	30.38	87	Meropenem	2.00	-17.28
89	Cephradine	14.00	20.46	88	Cephalexin	14.00	26.69
90	Cefadroxil	20.00	30.03	89	Cephradine	14.00	10.95
91	Cefepime	20.00	31.02	90	Cefadroxil	20.00	33.08

Table 8. Continued

92	Ceftazidime	21.00	46.03	91	Cefepime	20.00	38.77
93	Cefaclor	25.00	32.99	93	Cefaclor	25.00	24.18
94	Loracarbef	25.00	24.13	94	Loracarbef	25.00	21.42
96	Ceftizoxime	28.00	27.42	95	Cephpodoxime	27.00	34.77
97	Ceftibutin	30.00	30.97	96	Ceftizoxime	28.00	21.29
98	Cefuroxime	33.00	40.50	97	Ceftibutin	30.00	29.27
99	Cefotaxime	36.00	37.76	98	Cefuroxime	33.00	34.66
100	Cefprozil	40.00	42.83	99	Cefotaxime	36.00	35.98
101	Cephapirin	62.00	70.04	100	Cefprozil	40.00	43.68
104	Cefmetazole	70.00	77.37	105	Cephalothin	71.00	75.02
105	Cephalothin	71.00	73.21	106	Cefoxitin	73.00	56.35
106	Cefoxitin	73.00	54.05	107	Cefmandole	74.00	84.37
107	Cefmandole	74.00	78.12	108	Ceforanide	81.00	72.15
109	Cefotetan	83.00	71.68	109	Cefotetan	83.00	65.45
110	Cefazolin	89.00	66.55	110	Cefazolin	89.00	73.93
111	Cefoperazone	91.00	77.62	111	Cefoperazone	91.00	83.68
113	Cefonicid	98.00	74.12	113	Cefonicid	98.00	85.70

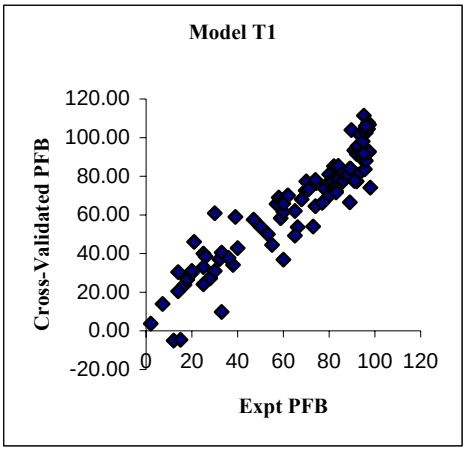


Figure 8. Cross-Validated PFB using T1.

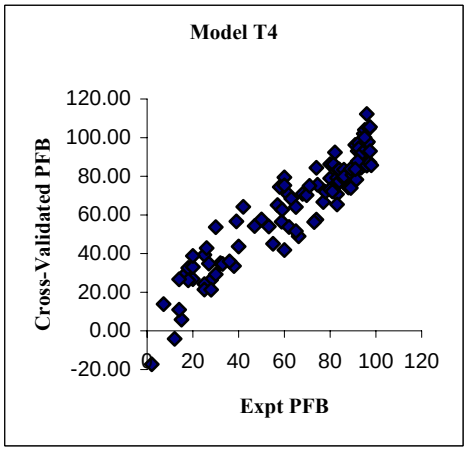


Figure 9. Cross-Validated PFB using T4.

External validation

The compounds of the test sets II and III are used for external validation of the models T1 and T4 to assess the actual predictive power of the models. The results of the external validation of models T1 and T4 using the test sets I and II are given in Table 9.

Table 9. Results of external validation of compounds in test sets II and III

Test set II – Model T1				Test set III – Model T4			
Comp No.	Comp Name	Expt PFB	Calc PFB	Comp No.	Comp. Name	Expt PFB	Calc PFB
12	Penicillin_35	42.00	57.06	10	Penicillin_12	33.00	9.93
23	Penicillin_38	62.00	56.71	38	Penicillin_29	82.20	57.67
24	Penicillin_36	63.00	71.62	49	Penicillin_14	87.00	63.00
38	Penicillin_29	82.20	59.79	51	Penicillin_68	89.30	84.12
49	Penicillin_14	87.00	57.32	58	Penicillin_4	93.30	108.84
54	Penicillin_52	91.50	74.42	70	Penicillin_59	97.00	105.86
59	Penicillin_20	93.60	70.95	76	Bacampicillin	20.00	49.92
76	Bacampicillin	20.00	67.44	92	Ceftazidime	21.00	55.11
95	Cephpodoxime	27.00	60.80	101	Cephapirin	62.00	68.69
102	Cefdinir	65.00	37.50	102	Cefdinir	65.00	41.18
103	Cefixime	67.00	37.09	103	Cefixime	67.00	38.42
108	Ceforanide	81.00	61.15	104	Cefmetazole	70.00	85.45

Analysis of the models G1-G6, T1 & T4 and their implications in drug discovery

As mentioned earlier, the unbound drugs are susceptible to metabolic clearance and at the same time, the unbound drugs are also responsible for pharmacological efficacy. Thus, the higher the protein binding, the lesser is the metabolic clearance of the drug, especially, when the binding is restrictive[27] Thus, an increase in the lifetime of the drug is observed for strongly binding drugs. Hence, the greater accuracy of prediction of high binding percent of drugs is desirable, as even small errors in this range will have significant impact on the pharmacodynamic effect of the drug. For example, the difference in the predicted values of percent fraction bound of drugs from 96 to 92 corresponds to a doubling of unbound fraction.

Given the significance of high binding behavior of drugs on the pharmacological efficacy, it is desirable for any model to perform well, especially, in the higher range of protein binding (75%-100%). Analyses of the predicted values of PFB of the β-lactam compounds in the range of 0-33%, 34-66% and 67-100% versus the observed PFB values were performed using G4. Among the 23 compounds in the range 0-33% of PFB, model, G4 predicted effectively in 63.63% (14/22) cases respectively keeping the acceptable standard error as 10%. Similarly its performance in the range of 34-66% is 77.27% (17/22). Significantly, the predictive performance of the model is excellent in the case of high binding drugs, i.e., of the 55 compounds in the range of 67-100%; the model G4 predicted PFB values successfully for 47/56 compounds. The performance of the models is at the upper limits of the expectations for a property that is measured by a variety of methods with results of modest accuracy.

Another significant aspect of the models is the finding that most of the variables are common among the datasets. This suggests that we have been able to identify a number of common

structural features contributing to the protein binding of these drugs. It is important to keep in mind that there is more than one binding site on albumin and that there are multiple proteins involved in the measured percent of binding. However, it is possible from these results to infer that certain structural features enhance binding, whereas others are detrimental.

As mentioned earlier, the GA based approaches used in the present study, GA1 and GA2, generated models with various combinations of molecular properties. This provides an advantage when one is interested not only in prediction, but also in understanding the mechanisms behind the modeled phenomenon. In this respect, it is clear that hydrophobicity increases drug binding to serum proteins because all the models contain a term for hydrophobicity factor, AlogP98. This has also been observed previously in other models of limited set of compounds[28, 29] and further supported by X-ray structure of HSA, both alone and bound to different ligands[27]. From a drug design point of view, an increase of hydrophobicity within a series of compounds is expected to result in an increased serum protein binding as long as the corresponding chemical modifications do not result in an opposing effect of other types of interactions that affect binding. In our models in addition to AlogP98, we uncovered new molecular properties that effect serum protein's binding and further studies are needed to understand the correlation between these molecular properties and serum protein binding.

Conclusions

In this paper, we have derived novel predictive models for serum protein binding affinity β-lactam class of antibiotics based on genetic algorithms and *in vitro* data of percent fraction of β-lactams bound to serum proteins. Further, the utility of automated variable and training set selection using genetic algorithms in the development of chemoinformatic models for protein binding affinity is demonstrated and this approach is expected to have significant impact in drug discovery and development of this class of compounds. The predictive performance of the models (both internal and external) is excellent and hence they are applicable to the design of new derivatives of β-lactams. Unlike other reported approaches, in this paper, models with various combinations of molecular properties are presented, which provides options to the end users.

Significantly, the models reported herein are based on molecular properties that are easy and fast to compute and hence can be applied for *virtual* screening of drug like compounds. Further, new molecular properties contributing to the binding of β-lactam analogs to serum proteins are uncovered.

Methods

Genetic algorithm

GAs[18–21] are computational algorithms constructed in analogy with the process of evolution and are widely used to find near-optimal solution where the variable space is exponentially proportional to the problem dimensions. Three fundamental mechanisms drive the evolutionary process: selection, crossover and mutation within chromosomes. Selection occurs on the current population by choosing the fittest individuals to reproduce. Reproduction, then, can result in the crossover and/or mutation of parent *genes* to form new solutions. In the present study, the crossover and mutation probabilities were set to 0.9 and 0.1 respectively. The population size has been fixed to 100 and the number of iterations to 5000.

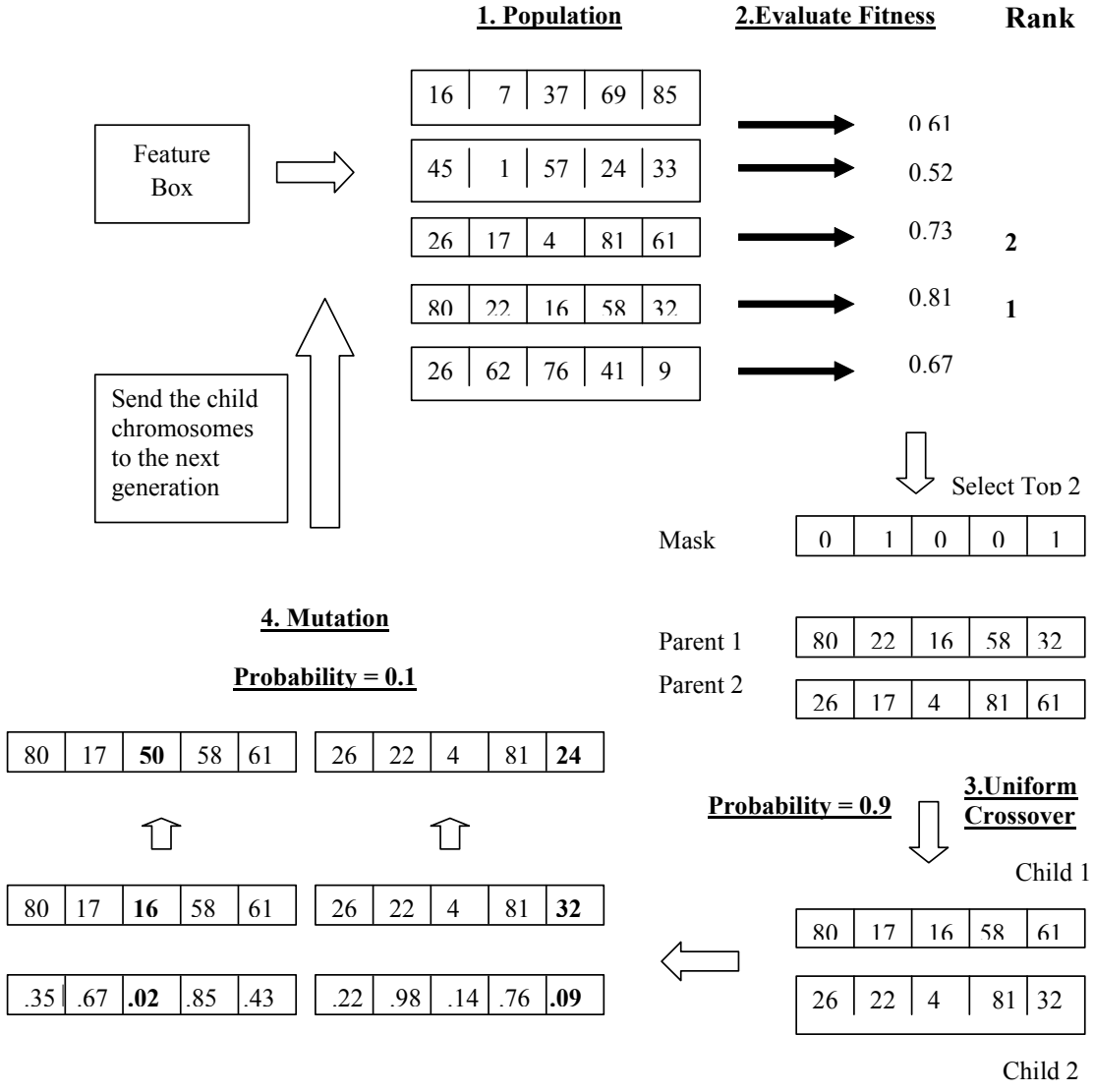


Figure 10. Genetic algorithm sequence.

Datasets

The quality of a QSPR model depends on: 1) size and quality of the data set used 2) the molecular propertied considered and 3) the mathematical methods employed. The serum protein binding data of 113 drugs and drug like compounds used in the present study is collected from literature [15–17, 24–26] and is expressed as a percentage of drug bound to total serum proteins (percent fraction bound, PFB)". These values are *in vitro* measurements carried out in several concentrations and the mean value is referred as "PFB. The PFB values range from 2.00 (Morepenem) to 98.0 (Cefonicid) and the molecular weights range from 258 (Penicillin_2) to 646 (Cefoperazone). The names, compound numbers, structures and their corresponding PFB values of the dataset are given in the supplemental material.

Training and test sets

In approach GA1, 100 drugs and drug like compounds are selected manually, and they constitute "Training set I" (Table 3). It consists of compounds with a wide range of molecular size. The selection is based on the substituents on the β-lactam ring system and the diversity of the experimental PFB data. The rest of the 13 compounds constitute the Test set I (Table 4).

In approach GA2, the same numbers of compounds are selected as training sets II and III (Table 7), in an automated fashion using GA from the initial pool of 113 compounds. The remaining 13 compounds are used as Test sets II and III (Table 9).

Software

All the programs used in the present study are developed in-house, are part of the in-house product TATA-Biosuite [22] and are used to perform the following: 1) to draw the 2D structures of the compounds 2) to calculate the molecular descriptors 3) to perform variable selection and the training set selection and 4) multiple linear regression and validation.

Descriptor generation

For all the compounds used in the study, 322 descriptors[23] are calculated using the QSAR module of "Tata-Biosuite" and are stored as a text file. Among the 322 calculated descriptors, 7 are physicochemical descriptors (AlogP98, SklogP, calculated vapour pressure etc.,), 7 are geometrical descriptors (topological surface area, 2D-van der Waals surface area, 2D-van der Waals volume etc.), 11 are structural descriptors (molecular weight, number of rotatable bonds, number of aromatic rings, number of hydrogen bond donors, number of hydrogen bond acceptors etc.,) and the remaining 297 are topological descriptors. The topological descriptors include Weiner index, Balaban index, Kier and Hall molecular connectivity indices, Kappa shape indices, autocorrelation indices, information content descriptors, Electrotopological indices, atomic-level-based AI topological descriptors etc.,

Cross-validation

We have used 'Leave-One-Out (LOO)" method, the simplest and commonly used cross validation approach, in our studies. In this approach, the property value for a given compound in the training set is predicted using the regression equation derived from the data of the remaining compounds. The PRESS (predictive residual sum of squares) statistic is computed using the formula

$$PRESS = \sum_{i=1}^{N} (y_i - y_i')^2$$

where y_i' *is the predicted LogBB value calculated after eliminating the* i^{th} *compound and* y_i *is the experimental LogBB value. The* Q *value given by*

$$Q = \sqrt{1 - \frac{PRESS}{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

Supplementary information available

Structures and their corresponding PFB data of the compounds used in this study are provided as the supplementary information.

References

1. van de Waterbeemd, H. *Curr. Opin. Drug Discov. Devel.* **2002**, *5*, 33.
2. Viswanathan, V. N.; Balan, C.; Hulme, C.; Cheetham, J. C.; Yaxiong Sun, Y. *Curr. Opin. Drug Discov. Devel.* **2002**, *5*, 400.
3. Chaturvedi, P. R.; Decker, C. J.; Odinecs, A. *Curr. Opin. Chem. Biol.* **2001**, *5*, 452.
4. Banik, G. M. *Current Drug Discovery* **2004**, 31.
5. Segall, M. D. *Future Drug Discovery* **2004**, 81.
6. van de Waterbeemd, H.; Gifford, E. *Nat. Rev. Drug Discov.* **2003**, *2*, 192.
7. Clark, D. E.; Grootenhuis, P. D. J. *Curr. Opin. Drug Discov. Devel.* **2002**, *5*, 382.
8. Herve, F.; Urien, S.; Albengres, E.; Duche, J. C.; Tillement, J. *Clin. Pharmacokinet.* **1994**, *26*, 44.
9. Naranjo, C. A.; Sellers, E. M. *Drug–protein binding*; Praeger: New York, 1986, 233.
10. Carter, D. C.; He, X. M.; Munson, S. H.; Twigg, P. D.; Gernert, K. M.; Broom, M. B.; Miller, T. Y. *Science* **1994**, *244*, 1195.
11. Curry, S.; Mandelkow, H.; Brick, P.; Franks, N. *Nat. Struct. Biol.* **1998**, *5*, 827.
12. Diaz, N.; Suarez, D.; Sordo, T. L.; Merz, K. M. Jr. *J. Med. Chem.* **2001**, *44*, 250.
13. Kremer, J. M.; Wilting, J.; Janssen, L. H. *Pharmacol. Rev* **1988**, *40*, 1.

14. Hall, M. L.; Hall, L. H.; Kier, L. B. *J. Comput. Aided Mol. Des.* **2003**, *17*, 103.
15. Colmenarajo, G.; Alvarez-Pedraglio, A.; Lavandera, J.-L. *J. Med. Chem.* **2001**, *44*, 4370.
16. Hall, M. L.; Hall, L. H.; Kier, L. B. *J. Chem. Inform. Comp. Scien.* **2003**, *43*, 2120.
17. Saiakov, R. D.; Stefan, L. R.; Klopman, G. *Perspect. Drug Discovery Des.* **2000**, *19*, 133.
18. Antonisse, J. *Proceedings of the Third International Conference on Genetic Algorithms*; 1989, 86.
19. Davis, L., Ed; *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York 1991.
20. Oliver, M.; Smith D. J. and Holland, J. R. C. *Proceedings of the Second International Conference on Genetic Algorithms*; 1987, 224.
21. Mitchel, T. *Machine Learning*; McGraw Hill Press: 1997.
22. TCS Bio-suite unveiled, *The Hindu Business Line;* 15th July 2004.
23. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors;* Wiley-VCH: 2000.
24. Hardman, J. G.; Limbird, L. E. Goodman & Gillman *The Pharmacological Basis of Therapeutics*, 10th ed; McGraw-Hill publishers: 2001.
25. Hardman, J. G.; Limbird, L. E.; Bird, A. E.; Marshall, A. C. *Biochem. Pharmacol.* **1967**, *16*, 2275.
26. Rolinson, G. N.; Sutherland, R. *Br. J. Pharmac. Chemother.* **1965**, *25*, 638.
27. Herve, F.; Urien, S.; Albengres, E.; Duche, J.C.; Tillement, J. *Clin Pharmacokinet* **1994**, *26*, 44.
28. Kaliszan, R.; Noctor, T. A. G.; Wainer, I. W. *Chromatogr.* **1992**, *33*, 546.
29. Andrisano, V.; Bertucci, C.; Cavrini, V.; Recanatini, M.; Cavalli, A.; Varoli, L.; Felix, G.; Wainer, I. W. *J. Chromatogr. A* **2000**, *876*, 75.
30. Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. *J. Pharm. Sci.* **2001**, *90*, 749.
31. Kennard, R.W.; Stone, L.A. *Technometrics* **1969**, *11*, 328.