Available online at www.sciencedirect.com
SCIENCE DIRECT
Bioorganic & Medicinal Chemistry 14 (2006) 4118–4129
Bioorganic & Medicinal Chemistry

In silico ADME modelling 2: Computational models to predict human serum albumin binding affinity using ant colony systems

Sitarama B. Gunturi,[a] Ramamurthi Narayanan[a,*] and Akash Khandelwal[b]

[a]*Life Sciences R&D Division, Advanced Technology Centre, Tata Consultancy Services Limited, # 1, Software Units Layout, Madhapur, Hyderabad 500 081, India*

[b]*College of Pharmacy, Department of Pharmaceutical Sciences, North Dakota State University, Sudro 8B2, Fargo, ND 58105, USA*

Received 25 November 2005; revised 31 January 2006; accepted 1 February 2006
Available online 28 February 2006

Abstract—Modelling of in vitro human serum albumin (HSA) binding data of 94 diverse drugs and drug-like compounds is performed to develop global predictive models that are applicable to the whole medicinal chemistry space. For this aim, ant colony systems, a stochastic method along with multiple linear regression (MLR), is employed to exhaustively search and select multivariate linear equations, from a pool of 327 molecular descriptors. This methodology helped us to derive optimal quantitative structure–property relationship (QSPR) models based on five and six descriptors with excellent predictive power. The best five-descriptor model is based on Kier and Hall valence connectivity index—Order 5 (path), Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 4, Auto-correlation descriptor (Broto-Moreau) weighted by atomic polarizabilities—Order 5, AlogP98, SklogS (calculated buffer water solubility) [$R = 0.8942$, $Q = 0.86790$, $F = 62.24$ and $SE = 0.2626$]; the best six-variable model is based on Kier and Hall valence connectivity index of Order 3 (cluster), Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 4, Auto-correlation descriptor (Broto-Moreau) weighted by atomic polarizabilities—Order 5, Atomic-Level-Based AI topological descriptors—AIdsCH, AlogP98, SklogS (calculated buffer water solubility) [$R = 0.9128$, $Q = 0.89220$, $F = 64.09$ and $SE = 0.2411$]. From the analysis of the physical meaning of the selected descriptors, it is inferred that the binding affinity of small organic compounds to human serum albumin is principally dependent on the following fundamental properties: (1) hydrophobic interactions, (2) solubility, (3) size and (4) shape. Finally, as the models reported herein are based on computed properties, they appear to be a valuable tool in virtual screening, where selection and prioritisation of candidates is required.

1. Introduction

Investigation of the causes of late stage failures in drug development revealed that inappropriate ADMET[1–9] properties [Absorption, Distribution, Metabolism, Excretion and Toxicity] were responsible for these failures. In an effort to reduce time and expence of the drug discovery and development process, it becomes apparent that early consideration of ADMET related issues is essential.[10,11] The need to know early the ADMET properties of drug candidates has propelled the development of numerous high-throughput in vitro screening methods. As these in vitro screens provide data using synthesised molecules, there is an increasing need for reliable and easily applicable computational models[12–14] to aid the design of new compounds with desirable ADMET profiles, prior even to chemical synthesis. Computational models are useful to rationalise a large number of experimental observations, offer potential for virtual screening applications and consequently can help in reducing time and cost of the drug discovery and development process. These potentials of in silico ADMET models created enormous interest among researchers from pharmaceutical industry[15–17] and academia [18–21] and thus it stands as an area of intense research.

The binding affinity of new chemical entities (NCEs) to HSA is one of the important ADMET properties considered in drug discovery and development. Probably, it is the most extensively studied protein because of its abundance, low cost, ease of purification and

Keywords: Human serum albumin; Protein binding; ADMET; QSPR; Ant colony systems.
* Corresponding author. Tel.: +91 40 55673581; fax: +91 40 55672222; e-mail: narayananr@atc.tcs.co.in

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.02.008

stability.[22] It plays a central role in drug pharmaco-kinetics,[23,24] in particular, in the distribution of drugs. Most drugs are transported in bound form to HSA and reach the target tissues. HSA allows solubilisation of hydrophobic compounds, thus contributing to a homogeneous distribution of drugs in the body and increases their biological lifetime.[25] Binding of a drug to serum albumin is a reversible process and is therefore in an equilibrium state. Only the unbound drug molecules contribute to the pharmacological efficacy; however, they are equally susceptible to metabolic reactions. Given the high concentration of albumin, the binding strength of any drug to serum albumin is the main factor that determines the availability of that drug and consequently, the diffusion of the drug from the circulatory system to target tissues.[22] All these factors cause the pharmaco-kinetics of almost any drug to be dramatically influenced and controlled by its binding affinity to serum albumin.

Extensive biochemical studies of Sudlow et al.[26,27] resulted in the proposition of two main drug binding sites in HSA denoted as I or warfarin site and II or indole-benzdiazepine site. Site I was shown to prefer large heterocyclic and negatively charged compounds and site II was the preferred site for small aromatic carboxylic acids. The ligand selectivity is comparatively broader for these two sites, allowing a range of drug molecules to bind at these sites. The broad selectivity is considered to be a result of the significant allosteric effects in HSA;[28] however, drug molecules can also interact non-specifically with HSA. Given the importance of drug binding to human serum albumin, it should be extremely useful to develop global quantitative structure–property relationship (QSPR) models to predict the binding affinity to human serum albumin, that are applicable to the whole medicinal chemical space. Global QSPR models can help to speed up the design of new compounds with appropriate HSA binding properties, thus leading to the optimization of the pharmaco-kinetics.

QSPRs have been successfully established to study and predict different important biopharmaceutical properties such as intestinal absorption,[29–31] oral availability,[32–35] blood–brain barrier transport,[21,36,37] metabolism,[38,39] toxicity[40,41] and skin[42] and corneal permeability.[43] There are several attempts reported[44–52] in the literature to generate predictive models for HSA binding affinity from molecular structures of drugs. Most of these models are based on compounds belonging to specific families,[44–49] thus are not global in nature and hence not applicable to virtual screening of diverse compounds. There are only three reports[50–52] as of date that can be described as global models for HSA binding affinity, to the best of our knowledge and these models are discussed briefly below: (a) Gonzalo Colmenarejo et al.[50] have applied genetic function approximation (GFA) to derive predictive models for human serum albumin binding. Their models are based on ClogP, topological descriptors and Jurs descriptors and the best being with six-descriptor combinations, (b) Hall et al.[51] have used multiple linear regression (MLR) for the generation of QSPR models for serum albumin binding using topological descriptors and their best models are based on nine descriptors and (c) Hou et al.[52] have reported predictive models for HSA using heuristic regression procedures based on seven descriptors.

Recently, we reported[53] predictive models for BBB permeation using a systematic variable selection method, namely, 'Variable Selection and Model Building Using Prediction' (VSMP). Systematic variable selection methods like VSMP provide the best solution from a given set of input variables by exploring all possible variable combinations; however, it suffers from the limitation of being computationally intensive, particularly when the number of variables to be selected is high (typically greater than 3). We encountered these bottlenecks while applying VSMP to the development of predictive models for HSA binding affinity and consequently, optimal solution could not be obtained using VSMP. Our continuing interest in the development of predictive ADME models prompted us to study the application of other variable selection procedures, to address these issues. We believe, the stochastic optimisation method based on the ant colony systems (ACS),[54–57] namely, ANTSELECT,[57] is expected to be computationally non-intensive and further will be able to provide the near optimal solution. The principal objectives of the present study are: (1) to develop accurate quantitative models to establish significant relationships between the relevant physicochemical descriptors and the HSA affinity and (2) to identify the important structural features that contribute to the binding affinity of drugs and drug-like compounds to HSA. In this paper, we describe the derivation of novel, global QSPR models for human serum albumin binding affinity using ant colony systems, originally developed by Dorigo et al.[54] along with multiple linear regression (MLR) for the first time, to the best of our knowledge. Further, we report: (a) the results of the internal and external validations to assess the predictive power of these QSPR models and (b) results of the comparison of the performances of the models reported herein with other published computational approaches, known as of date in the literature for a similar dataset. It is of high significance to note that the models reported herein are based on computed properties and hence, they are valuable tools for virtual screening of large sets of compounds, where selection and prioritisation of candidates is required.

2. Results and discussions

2.1. QSPR models for human serum albumin binding using ant colony systems

QSPR models are typically generated based on manually selected compounds and their physico-chemical properties, wherein intuition and experience play a key role. The recent advances in the field of computational chemistry have resulted in the easy

calculation of many molecular descriptors[58–61] which have potential applications in QSPR studies. Consequently, the process of selecting the best combination of descriptors, having significant relationship to a biological property, becomes extremely difficult without automation, particularly, from a large pool of descriptors. Variable selection methods[62–65] can explore various descriptor combinations based on a pre-defined fitness criteria and provide efficient models; however, they are scarcely employed to build predictive HSA models, except for the report of Colmenarejo et al.[50] In their study, they applied Genetic Function Approximation[66] (GFA), which takes inspiration from Genetic algorithms (GAs),[67] and Multivariate Adaptive Regression Splines (MARS).[68] In GFA, the variable space is replaced by a space of basis functions and the regression equation is derived using the selected basis functions. But, due to the complexity involved in constructing the basis functions and searching the space of basis functions, the potential applications of GFA in QSAR studies are limited. On the other hand, Genetic Algorithms found several applications in QSAR studies due to their efficiency, simplicity and flexibility. Even though GAs are powerful search algorithms, the selection process is random in nature and thus they do not keep track of the individual significance of the variables, which is critical for the selection of the most significant descriptors/combinations. In the present study, we applied ant colony algorithm, first proposed by Dorigo and colleagues[54] to solve difficult optimisation problems, like, the travelling salesman problem (TSP) and the quadratic assignment problem (QAP).[55] Recently, Izailev and Agrafiotis[56] successfully applied ANTSELECT algorithm to construct optimal regression tree models for QSARs. In their later work,[57] they have employed ANTSELECT along with artificial neural networks (ANN) for variable selection and tested its performance on well-known datasets. In their approach, the selection of each descriptor depends on the weight assigned to it. Initially, all the variables are assigned equal weights and these weights are incremented by a quantity that is proportional to the fitness value, in this case R, the coefficient of determination of ANN. Of significance, this approach offers the advantage of increasing the probability of selection of descriptors having high significance as compared to GAs, as the selection process in ANTSELECT algorithm is based on the individual weights (significance) of the descriptors. The variable with more weight has more probability of being selected than the ones with lesser weight. In the present study, we have retained the original ANTSELECT algorithm of Izrailev et al.[57] and used MLR instead of ANN, to build models for predicting the binding affinity of drugs and drug-like compounds to HSA. Further, we have introduced a key improvement in the original algorithm by which only descriptors with inter-correlation coefficient below 0.75 are selected in a given model. The number of descriptors to be selected is fixed at the beginning of the selection process and in the present study, it is fixed as 1 to 8. The subsequent steps involved in the selection of the descriptor combinations are executed as given in the original report of Izrailev et al.[57] [Supplementary material]

The global QSPR models for HSA binding affinity, reported herein, are generated using a dataset of 94 drugs and drug-like compounds, originally reported by Gonzalo Colmenarejo et al.[50] keeping $\log K'$hsa, as the dependent variable. Out of these 94 compounds, we manually selected 84 compounds as training set and set aside the remaining ten as the test set compounds. In the present study, a total of 396 physico-chemical descriptors are used in the QSPR model generation. Of these, 392 molecular descriptors are calculated using the in-house software, 'Bio-suite,'[70] and they fall into the following classes: (1) structural descriptors, (2) physico-chemical descriptors, (3) geometrical descriptors and (4) topological descriptors. In addition to these 392 descriptors calculated using 'Bio-Suite', four physico-chemical descriptors, like, SklogP, SklogS in Pure water, SklogS in buffer system, are calculated using the web-based program 'PreADME'[71,72] and are considered for the analysis. Out of the 396 descriptors, sixty-nine molecular descriptors that satisfy one or all of the following criteria are dropped: (a) descriptors having zero or constant value for each compound, (b) descriptors having a small variation in magnitude for all compounds and (c) descriptors having zero or constant value in more than 95% of the compounds. Remaining 327 descriptors are considered for model generation.

The predictive models for HSA binding affinity reported as of date in the literature are based on six,[50] seven[52] and nine[51] descriptors. These reports do not describe multiple models with different combinations of descriptors. In this report, we describe multiple models with different combinations of descriptors and this approach is expected to identify the most significant descriptors among the input descriptors that determine the binding affinity to HSA. Further, multiple models can facilitate in decision making based on consensus prediction.[15] We started the model building exercise with one descriptor and then systematically included additional descriptors, one at a time, up to 8 descriptors, as the principal objective is to build a model with optimum number of significant molecular properties. We have considered the addition of a descriptor as significant, only when the R value of the resulting model showed an improvement of 0.015 or more. We believe this criterion will enable us not to overparameterize the final model. With a view to select the optimum number of descriptors, among the 327 descriptors, the calculated descriptors of the 84 training set compounds are given as an input to ANTSELECT algorithm to build models based on two to eight descriptors. In order to select independent descriptors, the threshold of inter-correlation coefficient of the descriptors is set to 0.75. The models based on descriptors of highest significance, from 2 to 8, selected by ANTSELECT program and their corresponding statistics are given in Table 1.

Table 1. Models based on descriptors from 1 to 8

Sample size	Model	Number of variables	Descriptors selected	*R*	Δ*R*	*Q*	*SE*	*F*
84	M1	1	307	0.7867	0.0000	0.7722	0.353	133.18
	M2	2	307, 311	0.8225	0.0358	0.8096	0.325	86.66
	M3	3	159, 307, 311	0.8508	0.0283	0.8197	0.304	69.92
	M4	4	108, 159, 307, 311	0.8691	0.0183	0.8427	0.288	60.95
	M5	5	92, 159, 166, 307, 311	0.8942	0.0251	0.8679	0.2626	62.24
	M6	6	88, 159, 166, 283, 307, 311	0.9128	0.0186	0.8922	0.2411	64.09
	M7	7	92, 159, 166, 262, 272, 307, 311	0.9184	0.0056	0.8950	0.2334	59.47
	M8	8	108, 155, 175, 209, 262, 307, 309, 311	0.9187	0.0003	0.8787	0.232	50.72

Note. Δ*R* represents the improvement in the *R* value with the addition of a descriptor. Where, 92, Kier and Hall valence connectivity index—Order 5 (path); 108, Delta connectivity index—Order 4 (cluster); 155, Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 0; 159, Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 4; 166, Auto-correlation descriptor (Broto-Moreau) weighted by atomic polarizabilities—Order 5; 175, Auto-correlation descriptor (Broto-Moreau) weighted by van der Waals radius—Order 2; 209, Auto-correlation descriptor (Geary) weighted by atomic polarizabilities—Order 2; 262, Atom type E-State descriptor—SsCl; 272, hydrogen type E-State descriptor—SHdsCH AlogP98; 311, SKlogS.

The influence of the addition of a descriptor on the *R* value of a model is presented in Table 1 and in Figure 1. From the analysis of Figure 1 and Table 1, we observed the following: (a) there is significant net improvement (Δ*R*) in the *R* value of the models M2 to M6 by the inclusion of an additional descriptor to models M1 to M5, respectively, (b) the *R* value of the models M7 and M8 did not show significant improvement by the inclusion of an additional descriptor to the models M6 and M7, thus suggesting that the six variable model, M6, provides the optimal solution based on the input of descriptors used in the present study and (c) models with five- and six-descriptor combinations provide the optimal solution based on the dataset used in the present study.

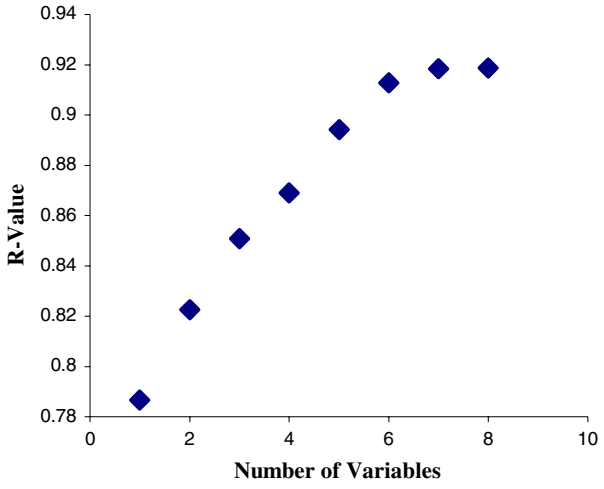

Figure 1. Plot of the *R* values of the models with variables from 1 to 8.

Based on the above observations, we decided to build multiple models based on five and six descriptors and to perform validation studies. Among the several five- and six-descriptor models generated by ANTSELECT, the best three models of each case are given in Table 2. The best five-descriptor model, A1, is based on descriptors 92, 159, 166, 307 and 311 with a coefficient of determination, *R* of 0.8942, and the cross-validated coefficient of determination, *Q* of 0.8679. The coefficients of determination of the other two five-descriptor models, A2 and A3, are 0.8923 and 0.8916, respectively, and their corresponding cross-validated coefficients of determination are 0.8742 and 0.8685, respectively. Significantly, the descriptors 159, 166, 307 and 311 are selected in all the best five-descriptor models of ANTSELECT, suggesting that these four descriptors have very high significance to the HSA binding behaviour of the compounds in the training set. The fifth descriptor in the models A1, A2 and A3 is 92, 88 and 55, respectively.

Table 2. Selected five and six-descriptor combinations and their statistical parameters

Model	Sample Size	No. of input descriptors	Descriptors selected	*R*	*Q*	*F*	*SE*
A1	84	327	92, 159, 166, 307, 311	0.8942	0.8679	62.2419	0.2626
A2			88, 159, 166, 307, 311	0.8923	0.8742	60.9417	0.2648
A3			55, 159, 166, 307, 311	0.8916	0.8685	60.4704	0.2656
A4	84	327	88, 159, 166, 283, 307, 311	0.9128	0.8922	64.0905	0.2411
A5			108, 155, 175, 262, 263, 311	0.9042	0.865	57.5442	0.2520
A6			79, 157, 163, 283, 307, 311	0.9017	0.8773	55.8025	0.2553

Where, 55, Subgraph count index—Order 3(cluster); 79, Kier and Hall molecular connectivity index of Order 3 (cluster); 88, Kier and Hall valence connectivity index of Order 3 (cluster); 92, Kier and Hall valence connectivity index—Order 5 (path); 108, Delta connectivity index of Order 4 (cluster); 155, Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 0; 157, Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 2; 159, Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 4; 163, Auto-correlation descriptor (Broto-Moreau) weighted by atomic polarizabilities—Order 2; 166, Auto-correlation descriptor (Broto-Moreau) weighted by atomic polarizabilities—Order 5; 175, Auto-correlation descriptor (Broto-Moreau) weighted by van der Waals radius—Order 2; 262, Atomic Type Electro-topological state index (E-state)—SsCl; 263, Atomic Type Electro-topological state index (E-state)—S-hydrophobic; 283, Atomic-Level-Based AI-topological descriptors—AidsCH; 307, AlogP98; 311, SklogS (calculated buffer water solubility).

The inter-correlation coefficients between the selected descriptors of the three models; A1, A2 and A3 range from 0.0057 to 0.3842 and consequently, the selected descriptors are independent of each other, thereby contributing significant structural information. The standard errors of the models are in the range of 0.2626 to 0.2656 and thus indicating that the models have good predictive power. The regression equations derived by performing MLR of the five-descriptor combinations are given below:

$$\begin{aligned}\log(\mathrm{K'hsa}) = {} & -0.730320 + 0.09505 * (\mathrm{Desc92}) \\ & - 0.008010 * (\mathrm{Desc159}) \\ & + 0.833715 * (\mathrm{Desc166}) \\ & + 0.142378 * (\mathrm{Desc307}) \\ & - 0.123680 * (\mathrm{Desc311}) \end{aligned} \quad (A1)$$

$$\begin{aligned}\log(\mathrm{K'hsa}) = {} & -0.655620 + 0.214465 * (\mathrm{Desc88}) \\ & - 0.008340 * (\mathrm{Desc159}) \\ & + 0.812193 * (\mathrm{Desc166}) \\ & + 0.150880 * (\mathrm{Desc307}) \\ & - 0.123420 * (\mathrm{Desc311}) \end{aligned} \quad (A2)$$

$$\begin{aligned}\log(\mathrm{K'hsa}) = {} & -0.774860 + 0.025875 * (\mathrm{Desc55}) \\ & - 0.007690 * (\mathrm{Desc159}) \\ & + 0.793349 * (\mathrm{Desc166}) \\ & + 0.160678 * (\mathrm{Desc307}) \\ & - 0.117140 * (\mathrm{Desc311}). \end{aligned} \quad (A3)$$

The best six-descriptor model, A4, is based on the descriptors 88, 159, 166, 283, 307 and 311 with a coefficient of determination, R of 0.9128, and the cross-validated coefficient of determination, Q of 0.8922. The coefficients of determination of the other two six-descriptor models, A5 and A6, are 0.9042 and 0.9017, respectively, and their corresponding cross-validated coefficients of determination are 0.8650 and 0.8773, respectively. The inter-correlation coefficients between the selected descriptors of the three models; A4, A5 and A6 range from 0.0057 to 0.4047 and consequently, the selected descriptors are independent of each other, thereby contributing significant structural information. The regression equations derived by performing MLR of the six-descriptor combinations are given below:

$$\begin{aligned}\log(\mathrm{K'hsa}) = {} & -0.509140 + 0.288682 * (\mathrm{Desc88}) \\ & - 0.008970 * (\mathrm{Desc159}) \\ & + 0.758452 * (\mathrm{Desc166}) \\ & - 0.028880 * (\mathrm{Desc283}) \\ & + 0.157323 * (\mathrm{Desc307}) \\ & - 0.133060 * (\mathrm{Desc311}) \end{aligned} \quad (A4)$$

$$\begin{aligned}\log(\mathrm{K'hsa}) = {} & -5.91993 + 1.826961 * (\mathrm{Desc108}) \\ & - 0.003230 * (\mathrm{Desc155}) \\ & + 1.864532 * (\mathrm{Desc175}) \\ & + 0.439154 * (\mathrm{Desc262}) \\ & + 0.019766 * (\mathrm{Desc263}) \\ & - 0.073610 * (\mathrm{Desc311}) \end{aligned} \quad (A5)$$

$$\begin{aligned}\log(\mathrm{K'hsa}) = {} & -0.542230 + 0.172591 * (\mathrm{Desc79}) \\ & - 0.008570 * (\mathrm{Desc157}) \\ & + 0.680459(\mathrm{Desc163}) - 0.027030 \\ & * (\mathrm{Desc283}) + 0.193151 \\ & * (\mathrm{Desc307}) - 0.112440 \\ & * (\mathrm{Desc311}). \end{aligned} \quad (A6)$$

Based on the R, Q, SE and F statistics, we have decided to use the best three of the five-descriptor models A1–A3 and the best three of the six-descriptor models, A4 to A6 for validation studies.

2.2. Validation

Two different tests of validation are performed on the models A1–A6. The first method of validation is using Leave-One-Out (LOO) method.[73] In this approach, the prediction of the property of a compound in a given set is based on the regression equation derived from the rest of the compounds of the set. The results of LOO cross-validations are given in Table 3 for models A1–A6. A plot of the predicted $\log K'$hsa values versus the observed $\log K'$hsa values for the training set based on models A1 and A4 is shown in Figure 2 and Figure 3, respectively, and the standard errors of the models A1–A6 range from 0.2411 to 0.2656, respectively, as given in Table 2.

The second method of validation involves the use of compounds that are not part of the training set, external validation. The quality of a predictive model is best tested using compounds that are not used in the training set. Based on the external and cross-validated results given in Table 3, we estimated the percentage of confidence of prediction of the training set compounds in various ranges of $\log K'$hsa and the results are given in Table 4. The confidence level of prediction of these models is 100% for compounds with experimental $\log K'$hsa values in the range of -1.33 to -0.50 and 77.8% for compounds in the range of -0.49 to -0.02.

Similarly, the percentage of confidence of prediction is 71.43% in the case of positive $\log K'$hsa compounds with experimental $\log K'$hsa in the range +0.0 to +0.50 and 100% for compounds in the range of +0.51 to +1.34. The predictive performance of the models is relatively less for compounds with experimental $\log K'$hsa values

Table 3. Results of the validation studies of the models A1–A6 and their comparison with the results of the reported models

Compound	Compound name	Expt	A1	A2	A3	A4	A5	A6	Hall et al.[51]	HM[52]
1	Acetyl salicylic acid[a,b]	***−1.39***	**−0.81**	**−0.80**	**−0.77**	**−0.80**	**−0.86**	**−0.77**	***−0.64***	***−0.48***
2	Cefuroxime	−1.33	−0.94	−1.08	−0.98	−1.07	−1.03	−0.93	−0.90	−1.26
3	Amoxicillin	−1.21	−0.97	−0.76	−0.99	−1.04	−1.03	−1.17	−1.08	−0.92
4	Cephalexin	−1.11	−0.88	−1.00	−0.94	−1.00	−0.69	−0.83	−0.40	−0.73
5	5-Fluoro Cytosin	−1.11	−1.20	−1.21	−1.23	−1.30	−0.92	−1.31	−0.79	−1.11
6	Cromolyn	−1.07	−0.41	−0.44	−0.29	−0.81	−0.67	−0.58	−0.45	−1.31
7	Ebselen	−1.04	−0.23	−0.40	−0.37	−0.45	−0.05	−0.29	−0.35	−1.07
8	Zidovudine	−1.02	−0.94	−0.94	−0.91	−1.02	−0.99	−0.96	−1.26	−1.26
9	Caffeine	−0.92	−1.00	−0.98	−0.97	−0.94	−0.74	−0.89	−0.88	−0.71
10	Acetaminophen[b]	−0.81	−0.49	−0.48	−0.50	−0.47	−0.52	−0.51	***−0.57***	***−0.92***
11	L-Tryptophan	−0.78	−0.19	−0.20	−0.20	−0.21	−0.32	−0.24	−0.56	−0.56
12	Methotrexate	−0.77	−0.69	−0.69	−0.60	−0.66	−0.79	−0.57	−0.35	−0.52
13	Propylthiouracil	−0.75	−0.29	−0.30	−0.30	−0.42	−0.81	−0.47	−0.83	−0.77
14	Antipyrine	−0.69	−0.21	−0.21	−0.20	−0.28	−0.20	−0.33	−0.37	−0.24
15	Penicillin V	−0.69	−0.91	−0.74	−0.92	−0.60	−1.05	−0.69	−0.71	−0.55
16	Salicylic Acid	−0.66	−0.81	−0.80	−0.78	−0.83	−0.71	−0.69	−0.69	−0.77
17	Cefuroxime Axetil	−0.56	−0.70	−0.81	−0.71	−0.76	−0.99	−0.67	−0.90	−0.61
18	Etoposide	−0.49	−0.62	−0.79	−0.62	−0.69	−0.81	−0.58	−0.64	−0.27
19	Atenolol	−0.48	−0.53	−0.48	−0.54	−0.42	−0.34	−0.51	−0.22	−0.32
20	Chloramphenicol[b]	−0.46	−0.66	−0.59	−0.61	−0.54	−0.53	−0.89	***−0.70***	***−0.81***
21	Cimetidine	−0.44	−0.60	−0.69	−0.69	−0.73	−0.66	−0.77	−0.59	−0.65
22	Chlorpropamide	−0.44	−0.23	−0.20	−0.21	−0.21	−0.43	−0.52	−0.42	−0.50
23	Sotalol	−0.44	−0.42	−0.25	−0.38	−0.16	−0.11	−0.35	−0.22	−0.13
24	Hydrochlorothiazide	−0.42	−0.67	−0.65	−0.68	−0.64	−0.53	−0.65	−0.76	−0.43
25	Tolazamide	−0.42	−0.24	−0.24	−0.25	−0.19	−0.32	−0.23	−0.14	−0.54
26	Hydrocortisone[a]	−0.40	**−0.08**	**−0.11**	**−0.15**	**−0.09**	**0.07**	**−0.23**	−0.23	−0.43
27	Nadolol	−0.40	−0.44	−0.21	−0.39	−0.05	−0.03	−0.19	0.08	−0.30
28	Prednisolone	−0.40	0.04	0.01	−0.02	−0.29	0.27	−0.42	−0.40	−0.29
29	Scopolamine	−0.34	−0.66	−0.75	−0.73	−0.68	−0.28	−0.62	−0.17	−0.26
30	Timolol[b]	−0.33	−0.59	−0.40	−0.57	−0.27	−0.40	−0.14	***−0.38***	***−0.51***
31	Metoprolol	−0.29	−0.39	−0.36	−0.42	−0.31	−0.14	−0.35	−0.10	−0.03
32	Trimethoprim	−0.26	−0.43	−0.44	−0.40	−0.41	−0.40	−0.31	−0.22	−0.35
33	Dansylglycine	−0.26	−0.03	0.02	−0.01	0.09	0.09	0.09	0.12	−0.30
34	Lidocaine	−0.23	−0.21	−0.20	−0.21	−0.16	−0.18	−0.10	0.15	−0.01
35	Methylprednisolone[a]	−0.22	**0.06**	**0.05**	**0.02**	**−0.21**	**0.22**	**−0.33**	−0.30	−0.25
36	Tolbutamide	−0.22	−0.30	−0.26	−0.27	−0.23	−0.41	−0.30	−0.27	−0.14
37	Sulfaphenazole	−0.21	0.30	0.30	0.29	0.36	0.07	0.22	−0.13	−0.12
38	Acebutolol	−0.21	−0.15	−0.09	−0.13	−0.01	−0.19	−0.04	−0.04	−0.05
39	Procaine	−0.19	−0.28	−0.26	−0.28	−0.23	−0.21	−0.27	−0.15	−0.19
40	Terazosin[b]	−0.16	−0.19	−0.27	−0.20	−0.21	−0.39	−0.20	***−0.26***	***−0.08***
41	Oxprenolol	−0.15	−0.10	−0.07	−0.12	−0.24	−0.13	−0.36	−0.20	−0.04
42	Lamotrigine	−0.13	−0.52	−0.50	−0.45	−0.53	−0.18	−0.41	−0.40	−0.26
43	Clonidine	−0.13	−0.35	−0.36	−0.33	−0.39	0.33	−0.11	−0.47	−0.18
44	Pindolol	−0.13	−0.01	0.02	−0.03	0.06	−0.12	0.03	−0.15	−0.25
45	Frusemide	−0.13	−0.60	−0.52	−0.51	−0.50	−0.39	−0.35	−0.64	−0.25
46	Carbamezapine	−0.10	0.52	0.46	0.48	0.30	0.34	0.26	−0.10	0.34
47	Ranitidine	−0.10	−0.15	−0.18	−0.23	−0.35	−0.34	−0.53	−0.30	−0.08
48	Camptothecin	−0.08	−0.04	−0.10	0.02	−0.12	−0.04	−0.05	−0.18	0.49
49	Tetracyclin	−0.08	−0.48	−0.49	−0.39	−0.33	−0.11	−0.26	−0.24	−0.35
50	Bupropion[b]	−0.05	−0.01	0.08	0.03	0.13	0.29	0.12	***−0.08***	***0.13***
51	Sumatriptan	−0.05	−0.23	−0.16	−0.25	−0.20	0.10	−0.44	0.19	−0.29
52	Warfarin	−0.04	0.11	0.07	0.11	0.14	0.10	0.22	0.05	0.30
53	Bumetamide	−0.03	0.11	0.13	0.16	0.21	0.06	0.31	−0.09	0.05
54	Oxyphenbutazone	−0.02	0.45	0.39	0.45	0.48	0.08	0.56	0.09	0.06
55	Acrivastine	−0.02	0.30	0.29	0.32	−0.18	0.30	−0.09	0.20	0.44
56	Phenytoin	0.00	0.07	0.04	0.09	0.09	0.13	0.11	−0.12	0.03
57	Doxycycline[a]	0.01	**−0.45**	**−0.51**	**−0.41**	**−0.39**	**−0.29**	**−0.31**	−0.38	−0.61
58	Ketoprofen	0.03	0.12	0.13	0.15	0.17	0.08	0.21	−0.01	−0.01
59	Alprenol	0.04	0.04	0.07	0.01	−0.06	0.10	−0.15	−0.10	−0.06
60	Prazosin[a,b]	0.06	**−0.06**	**−0.11**	**−0.03**	**−0.05**	**−0.31**	**0.00**	***−0.06***	***−0.21***
61	Digitoxin	0.13	0.55	0.36	0.38	0.25	−0.21	0.17	0.49	0.25
62	Levofloxacin	0.14	−0.05	−0.10	−0.05	−0.14	−0.05	−0.04		−0.03
63	Ciprofloxacin	0.14	−0.08	−0.15	−0.10	−0.07	0.02	−0.06	0.10	0.02
64	Labetalol	0.14	0.11	0.11	0.09	0.19	0.19	0.25	0.24	0.08
65	Norfloxacin	0.14	−0.09	−0.16	−0.11	−0.24	−0.03	−0.15	0.12	−0.16

(*continued on next page*)

Table 3 (*continued*)

Compound	Compound name	Expt	A1	A2	A3	A4	A5	A6	Hall et al.[51]	HM[52]
66	Phenylbutazone	0.19	0.53	0.46	0.51	0.54	0.23	0.61	0.20	0.38
67	Sancycline[a]	0.21	**−0.31**	**−0.37**	**−0.28**	**−0.25**	**−0.19**	**−0.19**	−0.24	−0.02
68	Minocycline	0.21	−0.16	−0.19	−0.12	−0.03	−0.09	0.02	−0.01	0.11
69	Naproxen	0.25	−0.01	−0.01	0.01	0.00	−0.13	0.00	−0.01	0.03
70	Clofibrate[b]	0.27	−0.07	0.08	0.04	0.15	0.02	−0.05	***−0.12***	***−0.03***
71	Propranalol	0.28	0.22	0.24	0.19	0.28	0.13	0.20	0.26	0.05
72	Tetracaine	0.32	0.14	0.18	0.13	0.25	0.09	0.16	0.16	0.31
73	Fusidic Acid	0.33	0.65	0.71	0.58	0.78	0.12	0.66	0.72	0.60
74	Novobiocin	0.35	0.36	0.53	0.60	0.35	−0.06	0.54	0.13	0.30
75	Ondansetron	0.37	0.11	0.04	0.07	0.09	0.10	0.18	0.18	0.33
76	Droperidol	0.43	0.54	0.46	0.50	0.43	0.48	0.49	0.47	0.63
77	Quinidine	0.44	0.41	0.28	0.27	0.15	0.39	0.17	0.57	0.41
78	Indomethacin	0.47	0.21	0.20	0.28	0.25	0.08	0.39	0.16	0.31
79	Quinine	0.49	0.40	0.27	0.27	0.15	0.39	0.16	0.57	0.40
80	Verapamil[b]	0.52	0.63	0.71	0.73	0.88	0.67	0.97	***1.16***	***0.98***
81	Sulfasalazine	0.56	0.30	0.33	0.37	0.41	−0.12	0.46	−0.04	0.21
82	Progesterone[a]	0.59	**0.47**	**0.42**	**0.33**	**0.47**	**0.47**	**0.33**	0.30	0.49
83	Desipramine[a]	0.61	**0.55**	**0.45**	**0.46**	**0.51**	**0.60**	**0.47**	0.72	0.56
84	Estradiol	0.68	0.35	0.27	0.23	0.41	0.34	0.34	0.36	0.37
85	Glibenclamide	0.68	0.30	0.28	0.30	0.39	0.34	0.45	0.58	0.58
86	Testosterone	0.74	0.31	0.27	0.18	0.33	0.36	0.18	0.20	0.30
87	Imipramine	0.75	0.65	0.62	0.58	0.70	0.67	0.64	0.91	0.77
88	Ketoconazole[a]	0.84	**−0.04**	**−0.05**	**0.03**	**0.04**	**0.73**	**0.25**	0.76	0.86
89	Promazine[a]	0.92	**0.75**	**0.68**	**0.63**	**0.73**	**0.51**	**0.62**	0.77	0.81
90	Itraconazole[b]	1.04	1.19	1.17	1.31	0.96	1.52	1.18	***1.50***	***1.70***
91	Triflupromazine	1.05	0.81	0.77	0.81	0.86	1.25	0.76	1.42	1.02
92	Chlorpromazine	1.10	0.89	0.82	0.76	0.88	0.76	0.68	0.83	0.89
93	Terbinafine	1.17	1.01	1.30	1.08	1.17	1.20	1.03	0.71	0.82
94	Clotrimazole	1.34	0.92	0.84	0.91	0.91	1.12	0.86	1.05	1.20

[a] Test compounds in models A1–A6.
[b] Test compounds in Refs. 51 and 52.

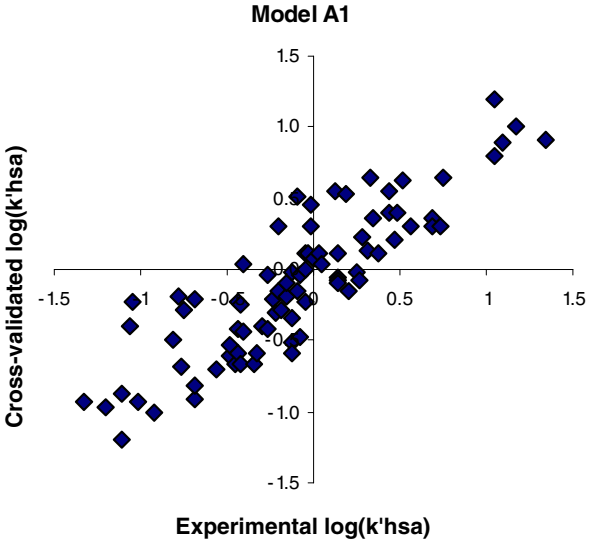

Figure 2. Plot of the cross-validated log(k'hsa) values using model A1 against the experimental log(k'hsa) values.

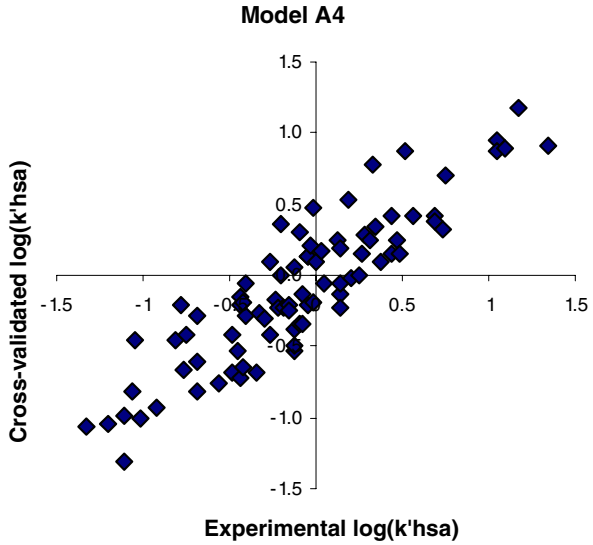

Figure 3. Plot of the cross-validated log(k'hsa) values using model A4 against the experimental log(k'hsa) values.

that are close to zero and otherwise the models performed excellently for other compounds. Thus, the models reported herein have excellent predictive power and offer excellent potential for applications in virtual screening studies.[74–76] Based on these analyses, we believe, the models A1 and A4 are the best five- and six-descriptor models, generated by ANTSELECT in the present study for the given dataset.

In order to identify the most significant descriptors with regard to HSA binding affinity, we computed the frequencies of the selected descriptors among all the five- and

Table 4. Confidence score of model A4 and its comparison with reported models

Range	No. of Compound	A4		Hall et al.[51]		HM[52]	
		No. of correct predictions	% of success	No. of correct predictions	% of success	No. of correct predictions	% of success
[−1.39, −1.01]	8	8	100	8	100	8	100
[−1.00, −0.51]	9	9	100	9	100	9	100
[−0.50, −0.01]	38	30	78.95	31	81.58	31	81.58
[0.01, 0.50]	23	15	65.22	15	65.22	15	65.22
[0.51, 1.00]	10	10	100	9	90	10	100
[1.01, 1.34]	5	5	100	5	100	5	100

six-descriptor models, in particular, models with the coefficient of determination, the R value, above 0.88 and the results are given in Table 5. The frequency of a descriptor is derived from the number of occurrences of a descriptor in all the five- and six-descriptor combinations (Table 5).

Analysis of Table 5 revealed the following observations:

(1) There are 106 five-descriptor models with 'R' value greater than 0.88.
(2) Of these,
   (a) 104 models are based on solubility (311) as one of the descriptors and the probability of occurrence of solubility is the highest among the descriptors constituting the five-descriptor combinations. (0.9811). Further, solubility has correlation of 0.5471 with $\log K'$hsa.
   (b) 98 models contain AlogP98, 307 as the descriptor with second highest probability (0.9245). AlogP98 has the highest correlation coefficient of 0.7867 with $\log K'$hsa among all of the selected descriptors, reported in the present study.
   (c) 53 models contain the descriptor 159, with third highest probability of 0.5000.
   (d) 35 models contain the descriptor 166, with a probability of 0.3302.

Similar analysis of Table 5 reveals the following significant observations about the six-descriptor models:

Table 5. Frequencies of the descriptor combinations with $R > 0.88$

Descriptor No.	Six-variable combinations (2024)		Five-variable combinations (106)	
	# of occurrences	Frequency	# of occurrences	Frequency
311	1848	0.9130	104	0.9811
307	1432	0.7075	98	0.9245
283	336	0.1660	9	0.0849
272	278	0.1374	—	—
263	542	0.2678	—	—
262	269	0.1329	—	—
178	311	0.1537	—	—
175	319	0.1576	9	0.0849
166	257	0.1270	35	0.3302
159	579	0.2861	53	0.5000
158	297	0.1467	10	0.0943
157	316	0.1561	15	0.1415
155	264	0.1304	—	—
108	279	0.1378	—	—
97	—	—	9	0.0849

(1) There are 2024 six-descriptor models with 'R' value greater than 0.88.
(2) Out of these
   (a) 1848 models contain solubility (311) as the descriptor with the highest probability (0.9130).
   (b) 1432 models contain 307, AlogP98, as the descriptor with second highest probability (0.7075).
   (c) 579 models contain the descriptor 159, with third highest probability of 0.2861.
   (d) 542 models contain the descriptor atomic type electro-topological state index (E-sate)-S-hydrophobic, 263, with a probability of 0.2648. Of interest is the high correlation of 263 with AlogP98, 0.8018, thus it provides similar physico-chemical information as that of AlogP98 with regard to human serum albumin binding affinity. Interestingly, if one were to add the frequencies of these two descriptors, 263 and 307, then, it turns out to be 0.9753, thus suggesting that hydrophobic interactions as the single most significant descriptor that contributes to HSA binding affinity of drugs and drug-like compounds. This inference is further substantiated by the fact that AlogP98, 307 provides the best R value for the single-descriptor model. (Table 1).

2.3. Description of the selected descriptors in models A1–A6 and their significance to HSA binding affinity

From the interpretation of the descriptors in the regression models, A1 to A6, it is possible to gain insights into the physico-chemical forces that govern the binding affinities of drugs and drug-like compounds to HSA. Generally, small molecules are bound to macromolecules through several types of interactions such as hydrogen bonding, van der Waals surface area, electrostatic and hydrophobic interactions. The descriptors used in the regression models A1–A6 and their descriptions are given in Table 6.

In order to identify the most significant descriptors that determine the HSA binding affinity of small organic molecules, we carried out the analyses of the physical meaning of the selected descriptors in the models A1–A6 and the results are given below:

(1) Hydrophobic interactions are the single most significant feature that determines the HSA binding affinity of drugs and drug-like compounds. This is because all the models reported herein contain

Table 6. Description of the descriptors selected in models A1–A6

Descriptor No.	Calculated property	Description
55	Subgraph count index—Order 3 (cluster)	Number of atoms with 3 heavy substituents
79	Kier and Hall molecular connectivity index of Order 3 (cluster)	Accessibility of bond to interaction with another molecule
88	Kier and Hall valence connectivity index of Order 3 (cluster)	Accessibility of bond to interaction with another molecule
92	Kier and Hall valence connectivity index—Order 5 (path)	Accessibility of bond to interaction with another molecule
108	Delta connectivity index of Order 4 (cluster)	Accessibility of bond to interaction with another molecule
155	Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 0	Sum of square root of atomic weight of each heavy atom in molecule
157	Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 2	Sum of square root of product of atomic weights of all 1–3 atoms
159	Auto-correlation descriptor (Broto-Moreau) weighted by atomic masses—Order 4	Sum of square root of product of atomic weights of all 1–4 atoms
163	Auto-correlation descriptor (Broto-Moreau) weighted by atomic polarizabilities—Order 2	Sum of square root of product of polarizability of all 1–2 atoms
166	Auto-correlation descriptor (Broto-Moreau) weighted by atomic polarizabilities—Order 5	Sum of square root of product of polarizability of all 1–5 atoms
175	Auto-correlation descriptor (Broto-Moreau) weighted by van der Waals radius—Order 2	Sum of square root of product of radius of all 1–3 atoms
262	Atomic Type Electro-topological state index (E-state)—SsCl	Electronic accessibility of atom
263	Atomic Type Electro-topological state index (E-state)—S-hydrophobic	Hydrophobicity
283	Atomic-Level-Based AI-topological descriptors—AidsCH	Sum of AI-topological descriptor of CH atom type attached to a double bond and a single bond
307	AlogP98	Partition coefficient
311	SklogS (calculated buffer water solubility)	Solubility

a term proportional to either, AlogP98, 307 or E-State index S-hydrophobic, 263. This has also been observed previously in published models developed using limited sets of compounds of the same chemical class such as 1,4-benzodiazepines,[69] 2,3-substituted 3-hydroxypropionic acids[77] and heterogeneous sets.[50,78,79] This observation is further supported by the X-ray structures of HSA, both in bound and unbound forms.[24,80–82] These structures show both sites I and II are made up mainly of hydrophobic residues and also that drug binding is stabilised to a large (if not primary) extent by hydrophobic interactions. From a drug design perspective, an increase of hydrophobicity within a series of compounds is expected to result in an increased HSA binding, as long as the corresponding chemical modifications do not result in opposing effects of other physico-chemical properties mentioned below.

(2) Solubility of the drugs and drug-like compounds as the second most significant property that determines the HSA binding affinity.

(3) Descriptors, 55, 79, 88, 92 and 108, describe different aspects of atom connectivity within a molecule, such as branching, flexibility. Further, they can be considered to describe the steric crowding around an atom and or bond, which in turn impacts the accessibility of an atom /bond to interact with HSA. In a simple sense, these descriptors describe the 'shape' of drugs and drug-like compounds and their role on HSA binding of small organic molecules.

(4) Descriptors, 155, 157, 159 and 175, are based on atomic weights of atoms and thus may have some relation to the effect of the size of a compound on HSA binding affinity; it is worth mentioning that, the actual physical meaning of these descriptors to HSA binding affinity is unclear and to the best of our knowledge, there is no report in the literature, describing the physical meaning of these descriptors.

(5) Descriptors, 163 and 166, describe the polarizability of all 1–3 and 1–5 atoms in a compound and they describe the accessibility of the corresponding atom types to interact with HSA.

(6) Descriptor 262 encodes the electron accessibility of chlorine atom(s) in a molecule and it is a favourable feature for binding to HSA.

From the above analysis, we infer that the binding affinity of a compound to HSA is, principally, dependent on: (1) hydrophobic interactions, (2) solubility, (3) shape and (4) size. The models A1–A6 are based on descriptors that describe the above fundamental properties and further, they (descriptor nos: 55, 155, 157, 159, 163, 166, 175 and 311) are used in HSA models for the first time. It is worth mentioning that there are many QSAR models reported in the literature based on calculated descriptors and often they are accompanied with an important drawback: a lack of interpretation in terms of simple structural and physico-chemical concepts.[83,84] We believe that with better understanding of the calculated physico-chemical properties having high significance to an ADME property, the performance of predictive models may be further improved. Finally, as the models reported herein are based on computed properties with high significance to HSA binding affinity of small organic molecules, they appear as valuable tools for virtual screening, where selection and prioritisation of candidates is required.

2.4. Conclusion

In this paper, we have described predictive models based on 5 and 6 descriptors, for human serum albu-

min binding affinity using a dataset of 94 diverse drugs and drug-like compounds and 327 molecular descriptors. For this aim, ANTSELECT algorithm along with MLR is employed for the first time in the literature, to the best of our knowledge. Unlike the reported approaches, we have reported multiple models with various combinations of descriptors, thus providing an end user with option to apply consensus prediction approach. The reported models contain contributions from new descriptors, whose significance to HSA binding is not known as of date. The models described herein possess excellent predictive power (both internal and external) and have potential for applications in virtual screening studies. From the analyses of the descriptors selected in the various models reported herein and their significance with regard to HSA binding affinity, we infer that HSA binding affinity of drugs and drug-like compounds is dependent on: (a) hydrophobic interactions, (b) solubility, (c) size and (d) shape. Finally, as the models reported herein are based on computed properties, they appear as valuable tools for virtual screening, where selection and prioritisation of candidates is required.

3. Methods

3.1. Datasets and human serum albumin binding

Gonzalo Colmenarejo et al.[50] reported the binding affinities of 94 diverse drugs and drug-like compounds to HSA, determined through high-performance affinity chromatography. In this assay, the compounds were assayed for HSA binding through high-performance affinity chromatography by using immobilized HSA column, a technique well established to obtain HSA binding constants.[69,48] As is customary in protein binding studies by high-performance chromatography, the binding constants were calculated in the logarithmic scale as $\log K'\text{hsa} = \log((t - t_0)/t_0)$, where t and t_0 are the retention times of the drug and $NaNO_3$ (dead time of the column), respectively.[69,48] We created a database of these 94 compounds and they constitute the training set and test set used in the present study. Out of these 94 compounds, we selected 84 compounds as training set and set aside the remaining ten as the test set compounds. The dataset includes heterogeneous groups of commercially available drugs such as COX-inhibitors, penicillins, β-adrenergic antagonists, quinalones, etc.

3.2. Training and test sets

84 drugs and drug-like molecules selected from the above database constitute the training set. The training set consists of compounds with a wide range of molecular size and complexity. The $\log K'\text{hsa}$ values range from −2.69 (Captopril) to 1.34 (Clotrimazole) and the molecular weights range from 129.0935 (5-Fluoro Cytosine) to 764.9488 (Digitoxin). A total of 10 compounds were set aside as test set for validation studies. The names and their corresponding $\log K'\text{hsa}$ values are shown in Table 4.

3.3. Ant colony systems

Ant Colony Optimisation (ACO) algorithms are stochastic search algorithms that are inspired by the behaviour of real ants.[54,55] Izrailev and Agrafiotis[57] developed a variable selection procedure, namely, ANTSELECT based on the ant colony systems and applied it to QSAR studies. In their algorithm, each ant is represented by a descriptor combination (or iteration) and the pheromone deposited by the ant is represented by the weight assigned to the descriptors in that combination. We have introduced a correlation check in their algorithm that checks all the pairwise correlations of the descriptors in a selected descriptor combination. If any pair has a correlation greater than 0.75, that descriptor combination is not considered for further study. The rest of the algorithm is executed as explained in the original paper[57] and presented in detail in the supplemental material. In our computation, we performed 30,000 iterations and the best combination is selected based on the R value of MLR. This process is repeated 1000 times and the selected 1000 combinations are cross validated and the combination with the maximum Q value is selected as the best model.

3.4. Software

All the programs used in the present study are developed in-house, are a part of the in-house product, 'Biosuite,'[70] and are used to perform the following: (1) draw the 2D structures of the compounds, (2) calculate the molecular descriptors, (3) descriptor selection and (4) multiple linear regression and validation. A web-based software, PreADME,[71,72] is used for the calculation of four physico-chemical descriptors.

3.5. Molecular descriptor calculation

A total of 392 descriptors[58–61] are calculated using the QSAR module of the in-house software, 'Biosuite.'[70] In addition, four physico-chemical descriptors are calculated using the web-based program 'PreADME' and all the calculated descriptors are stored as a text file, formatted into a table using an in-house program. In this paper, the terms 'variable' and 'descriptor' are used interchangeably as appropriate to the context of discussion.

3.6. Cross-validations

We have used 'Leave-One-Out (LOO)' method,[73] 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:

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

where y'_i is the predicted $\log K'\text{hsa}$ value calculated after eliminating the ith compound and y_i is the experimental $\log K'\text{hsa}$ value. The Q value is given by

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

Acknowledgments

AK would like to thank the providers of PreADME, a web based software product. RN and SBG would like to thank Dr. Rajgopal Srinivasan, Life Sciences R&D Division, TCS for constructive discussions during the preparation of the manuscript.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2006.02.008. The following informations are provided in the Supplementary material: (1) Ant Colony Algorithm; (2) Flow chart of the Ant colony algorithm; (3) Correlation matrix of the selected variables; (4) Histograms of the frequencies of the selected variables in the five and six variable combinations with $R > 0.88$.

References and notes

1. Kennadt, T. *Drug Discov. Today* **1997**, *2*, 436–444.
2. Eddy, E. P.; Maleef, B. E.; Hart, T. K.; Smith, P. L. *Adv. Drug Delivery Rev.* **1997**, *23*, 185–198.
3. Reichel, A.; Begley, D. *J. Pharm. Res.* **1998**, *15*, 1270–1274.
4. Colmenarejo, G.; Alvarez-Prdraglio, A.; Lavandera, J.-L. *J. Med. Chem.* **2001**, *44*, 4370–4378.
5. van de waterbeemd, H. *Curr. Opin. Drug Discov. Dev.* **2002**, *5*, 33–43.
6. Viswanathan, V. N.; Balan, C.; Hulme, C.; Cheetham, J. C.; Sun, Y. *Curr. Opin. Drug Discov. Dev.* **2002**, *5*, 400–406.
7. Chaturvedi, P. R.; Decker, C. J.; Odinecs, A. *Curr. Opin. Chem. Biol.* **2001**, *5*, 452–463.
8. Banik, G. M. *Curr. Drug Discov.* **2004**, 31–34.
9. Segall, M. D. *Future Drug Discov.* **2004**, 81–84.
10. Lombardo, F.; Gifford, E.; Shalaeva, M. *Mini Rev. Med. Chem.* **2003**, *3*, 861–875.
11. van de waterbeemd, H.; Gifford, E. *Nat. Rev. Drug Discov.* **2003**, *2*, 192–204.
12. Clark, D. E.; Grootenhuis, P. D. J. *Curr. Opin. Drug Discov. Dev.* **2002**, *5*, 382–390.
13. Oprea, T. I. *Molecules* **2002**, *7*, 51–62.
14. Hou, T.; Xu, X. *Curr. Pharm. Res.* **2003**, *10*, 1011–1033.
15. Sean, E. O'Brean; Marcel, J. de Groot *J. Med. Chem.* **2005**, *48*, 1287–1291.
16. Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. Z. *J. Comput. Aided Mol. Des.* **2002**, *16*, 381–401.
17. Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X.-Q.; Doweyko, A.; Li, Y. *J. Comput. Aided Mol. Des.* **2003**, 1–10.
18. Aureli, L.; Cruciani, G.; Cesta, M. C.; Anacardio, R.; Simone, L. D.; Moriconi, A. *J. Med. Chem.* **2005**, *48*, 2469–2479.
19. Rose, K.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 651–666.
20. Hutter, M. C. *J. Comput. Aided Mol. Des.* **2003**, *17*, 415–433.
21. Abraham, M. H. *Eur. J. Med. Chem.* **2004**, *39*, 235–240.
22. Carter, D. C.; Ho, J. X. *Adv. Protein Chem.* **1994**, *45*, 152–203.
23. Herve, F.; Urien, S.; Albengres, E.; Duche, J. C.; Tillement, J. *Clin. Pharmacol.* **1994**, 44–58.
24. Henrik, V. *Danish Med. Bull.* **1999**, *46*, 379–399.
25. Rang, H. P.; Dale, M. M.; Ritter, J. M. *Pharmacology*; Churchill Livingstone: Edinburgh, 1999.
26. Sudlow, G.; Birkett, D. J.; Wade, D. N. *Mol. Pharmacol.* **1976**, *12*, 1052–1061.
27. Sudlow, G.; Birkett, D. J.; Wade, D. N. *Mol. Pharmacol.* **1976**, *11*, 824–832.
28. Diaz, N.; Suarez, D.; Sordo, T. L.; Merz, K. M., Jr. *J. Med. Chem.* **2001**, *44*, 250–260.
29. Wegner, J. K.; Frohlich, H.; Zell, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 931–939.
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. *J. Pharmacol. Sci.* **2001**, *90*, 749–784.
31. Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. *J. Med. Chem.* **2001**, *44*, 1927–1937.
32. Yoshida, F.; Topliss, J. G. *J. Med. Chem.* **2000**, *43*, 2575–2585.
33. Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. *J. Med. Chem.* **2002**, *45*, 2615–2623.
34. Mandagere, A. K.; Thompson, T. N.; Hwang, K.-K. *J. Med. Chem.* **2001**, *45*, 304–311.
35. Turner, J. V.; Maddalena, D. J.; Snezana, A.-K. *Pharmacol. Res.* **2004**, *21*, 68–81.
36. Hou, T. J.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137–2152.
37. Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. *J. Med. Chem.* **2000**, *43*, 2204–2216.
38. Borodina, Y.; Rudik, A.; Filimonov, D.; Kharchevnikova, N.; Dmitriev, A.; Blinova, V.; Poroikov, V. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1998–2009.
39. Zamora, I.; Afzelius, L.; Cruciani, G. *J. Med. Chem.* **2003**, *46*, 2313–2324.
40. Basak, S. C.; Balasubramanian, K.; Gute, B. D.; Mills, D.; Gorczynska, A.; Roszak, S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1103–1109.
41. Klopman, G.; Chakravarti, S. K.; Zhu, H.; Ivanov, J. M.; Saiakhov, R. D. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 704–715.
42. Filho, O. A.-S.; Hopfinger, A. J.; Zheng, T. *Mol. Pharm.* **2004**, *1*, 466–476.
43. Yoshida, F.; Topliss, J. G. *J. Pharm. Sci.* **1996**, *85*, 819–823.
44. Kaliszan, R.; Noctor, T. A. G.; Wainer, I. W. *Chromatographia* **1992**, *33*, 546–550.
45. Markuszewski, M.; Kaliszan, R. *J. Chromatogr. B* **2002**, *768*, 55–66.
46. Ashton, D. S.; Beddell, C. R.; Cockerill, G. S.; Gohil, K.; Gowrie, C.; Robinson, J. E.; Slater, M. J.; Valko, K. *J. Chromatogr. B* **1996**, *677*, 194–198.
47. Andrisano, V.; Booth, T. D.; Cavrini, V.; Wainer, I. W. *Chirality* **1997**, *9*, 178–183.
48. Andrisano, V.; Bertucci, C.; Cavrini, V.; Recanatini, M.; Cavalli, A.; Varoli, L.; Felix, G.; Wainer, I. W. *J. Chromatogr. A* **2000**, *876*, 75–86.
49. Hanai, T.; Koseki, A.; Yoshikawa, R.; Ueno, M.; Kinoshita, T.; Homma, H. *Anal. Chim. Acta* **2002**, *454*, 101–108.
50. Colmenarejo, G.; Pedraglio, A. A.; Lavandera, J.-L. *J. Med. Chem.* **2001**, *44*, 4370–4378.
51. Hall, L. M.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2120–2128.

52. Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1693–1700.
53. Narayanan, R.; Gunturi, S. B. *Bioorg. Med. Chem.* **2005**, *13*, 3017–3028.
54. Dorigo, M.; Gambardella, L. M. *BioSystems* **1997**, *43*, 73–81.
55. Dorigo, M.; Caro, G.; Gambardella, L. M. *Artificial Life* **1999**, *5*, 137–172.
56. Israilev, S.; Agrafiotis, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 176–180.
57. Israilev, S.; Agrafiotis, A. *J. Chem. Inf. Comput. Sci.* **2002**, *13*, 417–423.
58. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH, 2000.
59. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley Interscience: New York, 2000.
60. *From Chemical Topology to 3D Molecular Geometry*, Balaban, A. T., Ed.; Plenum: New York, 1997.
61. Kubinyi, H.; Folkers, G.; Martin, Y. T. 3D QSAR in Drug Design Kluwer ESCOM, 1996–1998; Vol. 1–3.
62. Kubinyi, H. *QSAR* **1994**, *13*, 285–294.
63. Kubinyi, H. *QSAR* **1994**, *13*, 393–401.
64. Liu, S.-S.; Liu, H.-L.; Yin, C.-S.; Wang, L.-S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 964–969.
65. Izrailev, S.; Agrafiotis, D. K. *SAR QSAR Environ. Res.* **2002**, *13*, 417–423.
66. Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
67. Holland, J. Adaptation in Artificial and natural systems, 1975.
68. Friedman, J. Technical Report No. 102, 1988, Stanford University, Stanford, CA.
69. Kaliszan, R.; Noctor, T. A. G.; Wainer, I. W. *Charomatographia* **1992**, *33*, 546–550.
70. Tata Consultancy Services Limited, 'TCS Launches TATA Bio-Suite, Bioinformatics software to facilitate life science research through information technology', Press Release, June 7th, 2004.
71. http://preadmet.bmdrc.org/preadmet/index.php.
72. Lee, S. K.; Chang, G. S.; Lee, I. H.; Chung, J. E.; Sung, K. Y. No, K. T. 15th *European Symposium on Quantitative Structure–Activity relationships & Molecular Modeling* 2004, 5th–11th September, Istanbul, Turkey.
73. Wold, S. *QSAR* **1991**, *10*, 191–193.
74. Walters, W. P.; Stahl, M. T.; Murcko, M. A. *Drug Discov. Today* **1998**, *3*, 160–178.
75. Fox, S.; Farr-Jones, S.; Yund, M. A. *J. Biomol. Screening* **1999**, *4*, 183–186.
76. Bajorath, J. *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894.
77. Andrisano, V.; Bertucci, C.; Cavrini, V.; Recanatini, M.; Cavalli, A. *J. Chromatogr. A* **2000**, *876*, 75–86.
78. Koizumi, K.; Ikeda, C.; Ito, M.; Suzuki, J.; Kinoshita, T.; Yasukawa, K.; Hanai, T. *Biomed. Chromatogr.* **1998**, *12*, 203–210.
79. Hanai, T.; Miyazki, R.; Kinoshita, T. *Anal. Chim. Acta* **1999**, *378*, 77–82.
80. Carter, D. C.; He, X. M.; Munson, S. H.; Twigg, P. D.; Gernert, K. M.; Broom, M. B.; Miller, T. Y. *Science* **1994**, *244*, 1195–1198.
81. Carter, D. C.; He, X. M. *Science* **1990**, *249*, 302–303.
82. He, X.; Carter, D. *Naure* **1992**, *358*, 209–215.
83. Randic, M.; Zupam, J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550–560.
84. Hoffmann, R. *J. Mol. Str. (Theochem)* **1998**, *424*, 1–6.