Cheminformatic Models To Predict Binding Affinities to Human Serum Albumin

Gonzalo Colmenarejo,*,† Ana Alvarez-Pedraglio, and José-Luis Lavandera*

Structural Chemistry Department, GlaxoSmithKline, Parque Tecnológico de Madrid, E-28760 Tres Cantos, Madrid, Spain

Received June 18, 2001

Models to predict binding affinities to human serum albumin (HSA) should be very useful in the pharmaceutical industry to speed up the design of new compounds, especially as far as pharmacokinetics is concerned. We have experimentally determined through high-performace affinity chromatography the binding affinities to HSA of 95 diverse drugs and druglike compounds. These data have allowed us the derivation of quantitative structure−activity relationship models to predict binding affinities to HSA of new compounds on the basis of their structure. Simple linear, one-variable models have been derived for specific families of compounds ($r^2 \geq 0.80$; $q^2 \geq 0.62$): β-adrenergic antagonists, steroids, COX inhibitors, and tricyclic antidepressants. Also, global models have been derived to be applicable to the whole medicinal chemical space by using the full database of HSA binding constants described above. For this aim, a genetic algorithm has been used to exhaustively search and select for multivariate and nonlinear equations, starting from a large pool of molecular descriptors. The resulting models display good fits to the experimental data ($r^2 \geq 0.78$; LOF ≤ 0.12). In addition, both internal (cross validation and randomization) and external validation tests have demonstrated that these models have good predictive power ($q^2 \geq 0.73$; PRESS/SSY ≤ 0.23; $r^2 \geq 0.82$ for the external set). Statistical analysis of the equation populations indicates that hydrophobicity (as measured by the ClogP) is the most important variable determining the binding extent to HSA. In addition, structural factors (especially the topological $^6\chi_{ring}$ index and some Jurs descriptors) also frequently appear as descriptors in the best equations. Therefore, binding to HSA turns out to be determined by a combination of hydrophobic forces together with some modulating shape factors. This agrees with X-ray structures of HSA alone or bound to ligands, where the binding pockets of both sites I and II are composed mainly of hydrophobic residues.

Introduction

Serum albumin is probably the protein most extensively studied because of its abundance, low cost, ease of purification, and stability.[1] It is the most abundant protein in plasma, where it reaches a concentration of about 40 mg mL^{-1} (0.6 mM).[2] This protein is extremely important from a biopharmacological point of view because it is the major transporter of non-esterified fatty acids, as well as of different drugs and metabolites, to different tissues. Serum albumin allows solubilization of hydrophobic compounds, contributes to a more homogeneous distribution of drugs in the body, and increases their biological lifetime.[2] Given the high concentration of this protein, the binding strength of any drug to serum albumin is the main factor for availability of that drug to diffuse from the circulatory system to target tissues.[3] All these factors cause the pharmacokinetics of almost any drug to be dramatically influenced and controlled by its binding to serum albumin.[2,3]

Binding of drugs and metabolites to human serum albumin (HSA) has been studied for many years (reviewed in ref 1). HSA is able to bind an enormous variety of ligands and displays two main binding sites, the so-called sites I and II. These sites have been structurally identified from the crystal structure of the protein bound to different ligands.[1,4−6] Binding strengths vary by several orders of magnitude, with association constants (K_A) ranging from 10^3 to 10^7 M^{-1}. These constants, however, have been determined through different techniques and/or under different conditions, making it impossible to confidently compare the results from different laboratories.

Quantitative structure−activity relationships (QSARs) have been successfully established to predict different important biopharmaceutical properties, such as metabolism,[7] toxicity,[8] oral bioavailability,[9−11] intestinal absorption,[12] blood−brain transport,[13,14] and skin[15,16] and corneal[17] permeability, from molecular structure. In some cases, these models have been developed for specific families of compounds (see, for instance, ref 18 and citations therein), but in others an attempt has been made to model (at least in principle) the whole medicinal chemical space.[9−15]

Given the importance of drug binding to HSA, it should be extremely useful to develop QSARs to predict the binding affinity to HSA. This would allow speeding up of the design of new compounds with appropriate HSA binding properties and therefore the optimization of the pharmacokinetics.

* To whom correspondence should be addressed. For G.C.: e-mail, gonzalo_colmenarejo@sbphrd.com; phone, +3491 8074048; fax, +3491 8074062. For J.-L.L.: e-mail, jll27677@gsk.com; phone, +34 91 8070551; fax, +34 91 8070550.

† Present address: Cheminformatics Department, GlaxoSmithKline, Centro de Investigación Básica, Parque Tecnológico de Madrid, E-28760 Tres Cantos, Madrid, Spain.

10.1021/jm010960b CCC: $20.00
Published on Web 10/27/2001

In this work we have determined in a systematic way the binding constants to HSA of a large set (95 molecules) of widely different compounds of biopharmaceutical interest. This has allowed the direct and unambiguous comparison of binding constants for those compounds previously studied with different techniques and/or at different conditions. Then, we have developed QSAR models for specific well-known families of drugs. These models will provide accurate predictions of HSA binding constants for new compounds of these families. Finally, the whole database has been used to create QSAR models to predict binding constants of any new compound on the basis of its structure.

Experimental Section

A total of 95 drugs and druglike compounds were selected from the literature, trying to maximize the diversity in both structure and physicochemical properties. These compounds are gathered in Table 1, where their clinical use and target (whenever it is known) are also displayed. It can be seen there that the set comprises many families of well-known compounds from many different therapeutic areas.

These compounds were assayed for HSA binding through high-performance affinity chromatography by using an immobilized HSA column (ThermoHypersil, 150 mm × 4.6 mm size). This technique is well established as a fast and reliable method to obtain HSA binding constants.[19,20] The mobile phase used was 25 mM Na_2HPO_4, 25 mM KH_2PO_4 (pH 7.0)/acetonitrile [85:15; v/v]. A flow rate of 0.8 mL min^{-1} was used throughout. Experiments were conducted at 25 ± 0.1 °C. A minimum of four different chromatograms were obtained for each compound to ensure the reproducibility of the measurements and to estimate their errors.

As is customary in protein binding studies by high-performance affinity chromatography, the binding constants were calculated in the logarithmic scale as $\log K'hsa = \log((t - t_0)/t_0)$, where t and t_0 are the retention times of the drug and $NaNO_3$ (dead time of the column), respectively.[19,20] This resulted in an appropriate wide, centered, and Gaussian-like distribution of binding constants, displayed in Figure 1. All the retention times, together with their corresponding errors and log K'hsa values, are listed in Table 1 too; 84 out of the 95 compounds showed errors in retention times below 1%, and in the remaining cases the errors were always below 5%. This indicates the high reproducibility and accuracy of the measurements. On the other hand, these compounds span a wide range of K'hsa binding constants (3 orders of magnitude) corresponding to retention times between 2 and 56 min. It is therefore expected to be appropriate to model a wide range of druglike molecules.

Theoretical Calculations Section

For molecular structure calculations, descriptor calculations, and model generation and fitting, the programs Tsar[21] and QSAR+[22] (inside Cerius2[23]) were used, running on a four-processor Silicon Graphics Origin200 workstation under the IRIX 6.5 operating system.

Structure Calculations. Initial acceptable three-dimensional structures of the drugs at their neutral state were calculated with CORINA, inside Tsar; after that, accurate structures were obtained through energy minimization (EF algorithm) by using the semiempirical AM-1 Hamiltonian as the energy operator.[24] For this step, the program Vamp (inside Tsar) was employed. These quantum mechanical calculations were also used to determine some descriptors, like dipolar moments and mean atomic polarizabilities (see below).

Molecular Descriptor Calculation and Initial Selection. A wide range of molecular descriptors of different types were calculated for all 95 selected compounds. These included electronic, topological, information-content, spatial, structural, and thermodynamic descriptors. Electronic descriptors included AM-1 mean polarizability and dipole moment, CNDO/2 HOMO and LUMO energies, and superdelocalizability. Topological descriptors included Wiener,[25] Zagreb,[26] and Hosoya[27] indices, Kier and Hall molecular connectivity indices (χ's),[28] valence-modified connectivity indices (χ^v's),[28] subgraph count indices (SC's),[28] Kier's shape indices (κ's),[28] the molecular flexibility index,[28] and Balaban indices.[29] Information-content descriptors included indices of atomic composition, adjacency matrix, distance matrix, edge adjacency matrix, edge distance matrix, and multigraph information content.[30] Spatial descriptors included radius of gyration, Jurs descriptors,[31] shadow indices,[32] area, density, principal moment of inertia, and molecular volume. Structural descriptors included numbers of chiral centers, rotatable bonds, hydrogen-bond acceptors, hydrogen-bond donors, molecular weight, and aromatic density. Finally, the thermodynamic descriptors included AlogP,[33] Fh2o,[34] Foct,[34] MNDO heat of formation,[35] molar refractivity, ClogP,[36] and the Andrews binding free energy.[37]

This yielded a total of 107 initial descriptors. Several criteria were used to reduce this number while optimizing the information content of the descriptors set. First, descriptors for which no value was available for all the compounds were disregarded. Second, descriptors showing the same value for more than 70% of the molecules were removed. Third, if two descriptors showed a correlation coefficient greater than 0.9, one of them was left out. After these screening procedures, a set of 53 descriptors remained for model generation. These descriptors provided a set of uncorrelated variables with high information content.

QSAR's for Specific Families of Compounds. From inspection of Table 1 it can be seen that many of the 95 selected compounds can be grouped together in well-known "families", displaying structural similarity (measured through Tanimoto distances between Daylight[38] fingerprints) and also in most cases the same target and clinical use. Families were defined as displayed as additional entries in Table 1.

Similar compounds in principle should bind to the same binding site in HSA and establish mainly the same interactions with the protein. Therefore, it is expected that simple linear, one-variable models in this case can provide good predictions for new compounds inside each family because the structural variability is restricted and the binding site is the same. It is usually recommended to have at least five compounds per variable in linear regressions to produce reliable models.[39] So, only families with at least five member compounds were selected to derive specific univariate linear models. These are the so-called "COX inhibitors" (six compounds), comprising molecules inhibiting COX-1 and/or COX-2 (COX means cyclooxygenase); "penicillins" (five compounds); "β-adrenergic antagonists" (11 compounds); "steroids" (eight compounds), comprising steroids and structurally related molecules; and "tricyclic

Table 1. Database of Compounds Assayed for HSA Binding[a]

compound	t_r/min	SD_t	% error	log K'hsa	clinical use; target[b]	family
captopril	2.465	0.006	0.24	−2.69	antihypertensive; ACE inhibitor	
acetylsalicylic acid[c]	2.56	0	0	−1.39	antiinflamatory; COX-1 inhibitor	COX inhibitors
cefuroxime	2.575	0.006	0.23	−1.33	antibacterial; penicillin-binding protein inhibitor	penicillins
amoxicillin	2.61	0.01	0.38	−1.21	antibacterial; penicillin-binding protein inhibitor	penicillins
cephalexin	2.65	0.01	0.38	−1.11	antibacterial; penicillin-binding protein inhibitor	penicillins
5-fluorocytosine	2.65	0	0	−1.11	antifungal	
cromolyn	2.67	0.02	0.75	−1.07	antiasthmatic	
ebselen	2.685	0.006	0.22	−1.04	antiinflammatory	
zidovudine	2.695	0.006	0.22	−1.02	antiviral; reverse transcriptase inhibitor	
caffeine	2.755	0.006	0.22	−0.92	CNS cardiac stimulation; phosphodiesterase, A2 antagonist	
acetaminophen[c]	2.845	0.006	0.21	−0.81	antiinflammatory; COX-1 (COX-2) inhibitor	COX inhibitors
L-tryptophan	2.87	0.01	0.35	−0.78	amino acid	
methotrexate	2.88	0.01	0.35	−0.77	cancer chemotherapy; dihydropholate reductase	
propylthiouracil	2.892	0.006	0.21	−0.75	hyperthyroidism; thyroperoxidase inhibitor?	
antipyrine	2.96	0.01	0.34	−0.69	analgesic, antipyretic	phenazones
phenoxymethyl-penicillinic acid	2.96	0.02	0.67	−0.69	antibacterial; penicillin-binding protein inhibitor	penicillins
salicylic acid	3	0.01	0.33	−0.66	antiinflammatory; COX inhibitor	COX inhibitors
cefuroxime axetil	3.14	0	0	−0.56	antibacterial; penicillin-binding protein inhibitor	penicillins
etoposide	3.25	0.02	0.61	−0.49	cancer chemotherapy; topoisomerase II inhibitor	
atenolol	3.28	0.02	0.61	−0.48	antihypertensive, cardiac dysrithmias; β_1 antagonist	β-antagonists
chloramphenicol[c]	3.305	0.006	0.18	−0.46	antibacterial; ribosomal peptidyl transferase inhibitor	
cimetidine	3.355	0.006	0.18	−0.44	antiulcer; H_2 antagonist	H_2 antagonists
chlorpropamide	3.36	0.02	0.59	−0.44	hypoglycaemic; ATP-sensitive K^+ channel blocker at B-cells	sulfonylureas
sotalol	3.36	0.02	0.59	−0.44	antihypertensive, cardiac dysrithmias; β-antagonist	β-antagonists
hydrochlorothiazide	3.395	0.006	0.18	−0.42	diuretic; Na^+/Cl^- cotransporter inhibitor	
tolazamide	3.39	0.02	0.59	−0.42	hypoglycaemic; ATP-sensitive K^+ channel blocker at B-cells	sulfonylureas
hydrocortisone	3.43	0.02	0.58	−0.4	hormone	steroids
nadolol	3.44	0.01	0.29	−0.4	antihypertensive, cardiac dysrithmias; β-antagonist	β-antagonists
prednisolone	3.44	0.01	0.29	−0.4	antiinflammatory and immunosuppressive; glucocorticoid receptors	steroids
scopolamine	3.57	0.01	0.28	−0.34	antiemetic, antispasmodic; muscarinic antagonist	
timolol[c]	3.6	0.01	0.28	−0.33	glaucoma; β-antagonist	β-antagonists
metoprolol	3.72	0.04	1.07	−0.29	antihypertensive, antidysrhythmic; β_1-antagonist	β-antagonists
trimethoprim	3.8	0.01	0.26	−0.26	antibacterial; dihydrofolate reductase inhibitor	
dansylglycine	3.8	0.01	0.26	−0.26	nondrug	
lidocaine	3.92	0	0	−0.23	antidysrhythmic; Na^+ channel blocker	local anaesthetics
methylprednisolone	3.932	0.005	0.13	−0.22	antiinflammatory and immunosuppressive; glucocorticoid receptors	steroids
tolbutamide	3.95	0.02	0.51	−0.22	hypoglycaemic; ATP-sensitive K^+ channel blocker at B-cells	sulfonylureas
sulfaphenazole	3.97	0.03	0.75	−0.21	antimicrobial	
acebutolol	3.975	0.006	0.15	−0.21	antihypertensive, cardiac dysrithmias; β_1-antagonist	β-antagonists
procaine	4.05	0.03	0.74	−0.19	local anaesthetic; Na^+ channel blocker	local anaesthetics
terazosin[c]	4.17	0.02	0.48	−0.16	antihypertensive; α_1 antagonist	α_1 antagonist
oxprenolol	4.185	0.006	0.14	−0.15	antihypertensive, antidysrhythmic; β-antagonist	β-antagonists
lamotrigine	4.28	0.02	0.47	−0.13	antiepileptic; Na^+ channel blocker	
clonidine	4.29	0	0	−0.13	antihypertensive; α_2 agonist	
pindolol	4.295	0.006	0.14	−0.13	antihypertensive, cardiac dysrithmias; β-antagonist	β-antagonists
frusemide	4.3	0.02	0.46	−0.13	diuretic; $Na^+/K^+/2Cl^-$ cotransporter blocker	loop diuretics
carbamazepine	4.4	0.02	0.45	−0.1	antiepileptic; Na^+ channel blocker	tricyclic antidepressants
ranitidine	4.41	0.05	1.13	−0.1	antiulcer; H_2 antagonist	H_2 antagonists
camptothecin	4.492	0.005	0.11	−0.08	cancer chemotherapy; topoisomerase I inhibitor	
tetracycline	4.5	0.2	4.44	−0.08	antibacterial; A site at ribosomal 30S subunit	tetracyclines
bupropion[c]	4.64	0	0	−0.05	antidepressant	
sumatriptan	4.67	0.01	0.21	−0.05	antimigraine; $5HT_{1D}$ agonist	
warfarin	4.72	0.08	1.69	−0.04	anticoagulant; vitamin K reductase inhibitor	
bumetanide	4.76	0.03	0.63	−0.03	diuretic; $Na^+/K^+/2Cl^-$ cotransporter blocker	loop diuretics
oxyphenbutazone	4.8	0.02	0.42	−0.02	antiinflammatory	phenazones
acrivastine	4.83	0.03	0.62	−0.02	antiallergic; H1 antagonist	
phenytoin	4.88	0.01	0.2	0	antiepileptic; Na^+ channel blocker	
doxiciclyne	5	0.1	2	0.01	antibacterial; A site at ribosomal 30S subunit	tetracyclines
ketoprofen	5.12	0.02	0.39	0.03	antiinflammatory; COX inhibitor	COX inhibitors
alprenolol	5.145	0.006	0.12	0.04	antihypertensive, antidysrhythmic; β-antagonist	β-antagonists
prazosin[c]	5.28	0.02	0.38	0.06	antihypertensive; α_1 antagonist	α_1 antagonist
digitoxin	5.81	0.03	0.52	0.13	cardiotonic; Na^+/K^+ ATPase pump inhibitor	steroids
levofloxacin	5.86	0.008	0.14	0.14	antibacterial; topoisomerase II inhibitor	quinolones
ciprofloxacin	5.86	0.08	1.36	0.14	antibacterial; topoisomerase II inhibitor	quinolones
labetalol	5.897	0.005	0.08	0.14	antihypertensive in pregnancy; α/β-antagonist	β-antagonists
norfloxacin	5.9	0.1	1.69	0.14	antibacterial; topoisomerase II inhibitor	quinolones
phenylbutazone	6.31	0.04	0.63	0.19	antiinflammatory	phenazones
sancicline	6.42	0.07	1.09	0.21	antibacterial; A site at ribosomal 30S subunit	tetracyclines

Table 1 (Continued)

compound	t_r/min	SD_t	% error	log K'hsa	clinical use; target[b]	family
minocycline	6.46	0.07	1.08	0.21	antibacterial; A site at ribosomal 30S subunit	tetracyclines
naproxen	6.8	0.1	1.47	0.25	antiinflammatory; COX-1 and COX-2 inhibitor	COX inhibitors
clofibrate[c]	7.08	0.008	0.11	0.27	lipid-lowering drug	
propranolol	7.11	0.01	0.14	0.28	antihypertensive, antidysrhythmic; β-antagonist	β-antagonists
tetracaine	7.56	0.07	0.92	0.32	local anaesthetic; Na^+ channel blocker	local anaesthetics
fusidic acid	7.78	0.06	0.77	0.33	antibacterial; EF2−ribosome complex	steroids
novobiocin	8.01	0.04	0.5	0.35	antibacterial; topoisomerase II inhibitor	
ondansetron	8.21	0.03	0.36	0.37	antiemetic; 5-HT3 antagonist	
droperidol	9.09	0.03	0.33	0.43	antipsychotic, antiemetic; D2 antagonist	
quinidine	9.31	0.04	0.43	0.44	antidysrhythmic; Na^+ channel blocker	
indomethacin	9.77	0.02	0.2	0.47	antiinflammatory; COX-1 inhibitor	COX inhibitors
quinine	10.11	0.01	0.1	0.49	antimalarial; heme polymerase inhibitor	
verapamyl[c]	10.69	0.02	0.19	0.52	antianginal; L-type heart Ca^{2+} channel blocker	
sulfasalazine	11.41	0.4	3.5	0.56	antirheumatoid	
progesterone	11.97	0.08	0.67	0.59	hormone	steroids
desipramine	12.6	0.1	0.79	0.61	antidepressant; noradrenaline transporter blocker	tricyclic antidepressant
estradiol	14.24	0.07	0.49	0.68	hormone	steroids
glibenclamide	14.28	0.007	0.05	0.68	hypoglycaemic; ATP-sensitive K^+ channel blocker at B-cells	sulfonylureas
testosterone	15.94	0.05	0.31	0.74	hormone	steroids
imipramine	16.4	0.1	0.61	0.75	antidepressant; noradrenaline transporter blocker	tricyclic antidepressant
ketoconazole	19.6	0.2	1.02	0.84	antifungal; P450 (ergosterol synthesis involved) inhibitor	azoles
promazine	23.12	0.04	0.17	0.92	antipsychotic; D_2 antagonist	tricyclic antidepressant
itraconazole[c]	29.7	0.09	0.3	1.04	antifungal; P450 (ergosterol synthesis involved) inhibitor	azoles
triflupromazine	30.2	0.2	0.66	1.05	antipsychotic; D_2 antagonist	tricyclic antidepressant
chlorpromazine	33.8	0.2	0.59	1.1	antipsychotic; D_2 antagonist	tricyclic antidepressant
terbinafine	39.1	0.1	0.25	1.17	antifungal; squalene epoxidase inhibitor	
clotrimazole	55.82	0.03	0.05	1.34	antifungal; P450 (ergosterol synthesis involved) inhibitor	azoles

[a] Experimental data obtained as described in the text. t_r = average retention time, obtained from the average of the detected peak maxima displayed in at least four chromatograms for each compound; SD_t = standard deviation of the retention time; %error = error percentage of the retention time. The log K'hsa values were all taken with two decimal numbers, no matter what the error in the retention times. [b] Taken from ref 2. [c] Compounds comprising the external validation set (see text).

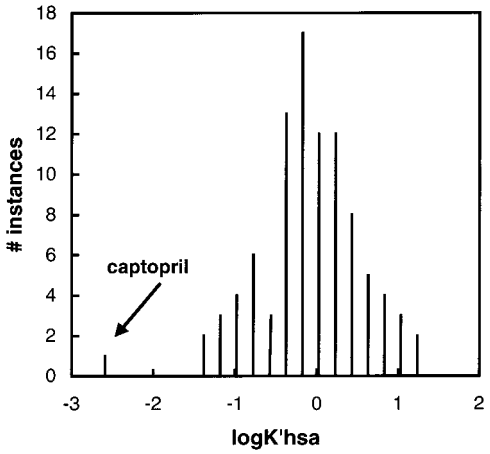

Figure 1. Histogram of instances distribution of log K'hsa values. The captopril value, calculated from the average value of t_r in Table 1, is pointed out; its retention time is the same as t_0 within experimental error, so it was not included in model generation (see text).

antidepressants" (six compounds), comprising tricyclic antidepressants and structurally similar compounds.

To derive the models, all possible univariate linear regressions were calculated for all 53 descriptors described above, and the best model for each family was selected. In principle, the goodness of the fit, as judged by the squared correlation coefficient, r^2, was used as selection criterion for the best model. The closer this value is to 1, the better the fit (and the model) is. However, special care was also taken of the predictive power of the models such that in some cases second-best models were finally the selected ones because they scored better in the validation tests (see below).

Because the families contained relatively small numbers of molecules, no external validations were possible. Therefore, the models were *internally* validated by two methods. First, leave-out-one (LOO) cross-validation tests were conducted and the PRESS (predictive residual sum of squares) statistic was computed. It is defined as

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

where y_i' are the predicted values of the dependent variable at each of the models generated after elimination of one molecule, for that molecule, and y_i are the actual (experimental) values. For reasonable regression models, the PRESS/SSY ratio,

$$\frac{\mathrm{PRESS}}{\mathrm{SSY}} = \frac{\sum_{i=1}^{n} (y_i - y_i')^2}{\sum_{i=1}^{n} (y_i - \hat{y})^2} \tag{2}$$

where $\hat{y}$ is the average value of the dependent variable and SSY is given by the denominator of eq 2, should be smaller than 0.4.[40] Also, the cross-validated squared correlation coefficient, q^2, was computed. It is obtained from the equation

$$q^2 = 1 - \frac{\mathrm{PRESS}}{\mathrm{SD}} \tag{3}$$

where SD is the sum of squared deviations of the dependent variable values from their mean. For reason-

Table 2. QSAR Models for Specific Families of Compounds in the Database[a]

family	*n*	QSAR equation	PRESS/SSY	r^2	q^2	r_{nr}	$\hat{r}_r$	SD-r_r
β-adrenergic antagonists	11	−0.52 + (0.23)(ClogP)	0.25	0.83	0.74	0.91	0.70	2.50
steroids	8	(1.63−5.44)(JursRPSA)	0.11	0.94	0.89	0.97	0.69	2.40
COX inhibitors	6	−4.82 + (12.61)(DensArom)	0.38	0.80	0.62	0.89	0.85	0.52
tricyclic antidepressants	6	−5.53 + (1.64)(RadOfGyr)	0.08	0.97	0.92	0.98	0.79	1.37
penicillins	5	6.37−11.25(Shadow-YZ_{fr})	0.71	0.67	0.28	0.82	0.92	−0.77

[a] *n* = number of compounds in the family; PRESS/SSY, r^2, q^2, as described in the text; r_{nr} = correlation coefficient for the nonrandom model; $\hat{r}_r$ = average correlation coefficient for the random models; SD-r_r = number of standard deviations of the mean value of *r* of all random trials to the nonrandom r_{nr} value.

able regression models, q^2 should be close to r^2, usually smaller.[40]

The second validation method was a randomization test, where log K'hsa data were scrambled several times and new regression models were derived each time. This is particularly important for this problem because we are selecting one variable from a pool of descriptors much larger than the number of molecules in the models, so chances are that good fits will be obtained, but just from chance correlations not representing true relationships between log K'hsa and the descriptor selected.[41−43] In our case, nine randomization trials were performed to achieve a 90% confidence level. Both the mean value of the correlation coefficient *r* for random trials, $\hat{r}_r$, and the number of standard deviations of the mean value of *r* of all random trials to the nonrandom *r* value, SD-r_r, were determined. For models not originating from chance correlations, the $\hat{r}_r$ value should be well below the nonrandom one, r_{nr}, and the SD-r_r should be large.[39] This should indicate that the nonrandom model represents a true relationship between the selected variables and the log K'hsa.

The resulting models are gathered in Table 2, together with their PRESS/SSY, r^2, q^2, r_{nr}, $\hat{r}_r$, and SD-r_r statistics. As can be seen in Table 2, different descriptors provide for each family the best model: ClogP, JursRPSA, aromatic density, radius of gyration, and shadow-YZ_{fr}, for β-adrenergic antagonists, steroids, COX inhibitors, tricyclic antidepressants, and penicillins, respectively. In the case of COX inhibitors, the equation shown is not the best one as far as r^2 is concerned; the best fit corresponded to one equation having an AM-1 dipole moment as descriptor. However, this model displayed low predictive power, so it was replaced with the one shown in Table 2.

In all the cases but penicillins, the fit is good, with r^2 greater than 0.80 and PRESS/SSY below 0.4. Also, the internal validation tests indicate good predictive power and no chance correlations for all the models but penicillins; q^2 is close to r^2, $\hat{r}$ is well below r_{nr}, and SD-r_r is large. This is especially fair in the case of β-adrenergic antagonists, steroids, and tricyclic antidepressants. In the case of penicillins, no model could be derived from any of the 53 descriptors showing acceptable fit and reasonable cross validation and randomization test results.

These equations should be useful for predicting HSA binding affinities for new molecules of the corresponding families: β-adrenergic antagonists, steroids, COX inhibitors, and tricyclic antidepressants.

Global QSAR. All the compounds but one in the database of log K'hsa values were used to generate models to be employed for predicting binding affinities to HSA of *any* new druglike compound. The only exception, captopril, was not used because its retention time is the same as that of $NaNO_3$ (t_0 = 2.46′) within experimental error. The genetic funcion approximation (GFA) facility in QSAR+ was used for this aim to exhaustively search for models and to select the best ones. This facility consists of a genetic algorithm, which at the outset generates randomly an ensemble of model equations, each of them being codified by a "gene". The genes fight for reproduction with success proportional to their "fitness" (in this case the minimization of the so-called lack of fit (LOF) statistic of the model[44]). Both crossover and mutation are allowed to increase the search in the model space. Here, mutations represent modifications in the model equation, while crossovers correspond to transfers of some equation terms between models. In this way, the search is made very quickly without the need of trying all the possible models, which would result in an unaffordable computation. On the other hand, the fight imposed inside the population of genes (models) produces a fast selection of the best equations. After a relatively small number of generations the system converges to a final population comprising mainly the best model, together with a set of models similar to that, both in functional form and LOF. Genetic algorithms have proved to be very powerful optimization algorithms for complex systems in very different contexts.[45] Here, we have a large set of descriptors and possible functional forms of the QSAR equations, so GFA appears to be the appropriate choice for this problem. In addition, GFA ends up with an equation, therefore allowing the interpretation of the model in physical terms. GFA is also able to discriminate between good and bad descriptors and to select the former. These two properties are not available for other optimization algorithms, like neural networks; when used for similar problems, they are normally supplemented with a genetic algorithm to search and select for descriptors, the neural network being only used for regression purposes.[12]

The dataset, comprising 94 compounds, was split into two subsets: a training one (84 molecules) used to derive the models and an external validation one (10 molecules, shown with a *c* in Table 1). The last set was selected to span uniformly the whole range of log K'hsa values. Multiple runs were initially conducted in order to set the parameters of the GFA for this problem. For instance, it was observed that runs with an excessive number of generations converged on slightly better models but had worse external and internal validation test results, indicating they had moved to the overfitted side. In contrast, too few generations failed to yield good models and convergence. For any number of initial random equations, an approximate optimum number of generations could be suggested. In addition,

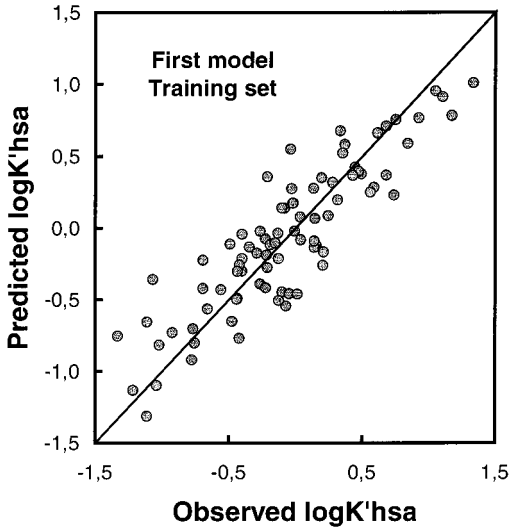

Figure 2. Observed vs predicted log K'hsa values plot for the first global model, obtained with the training set.

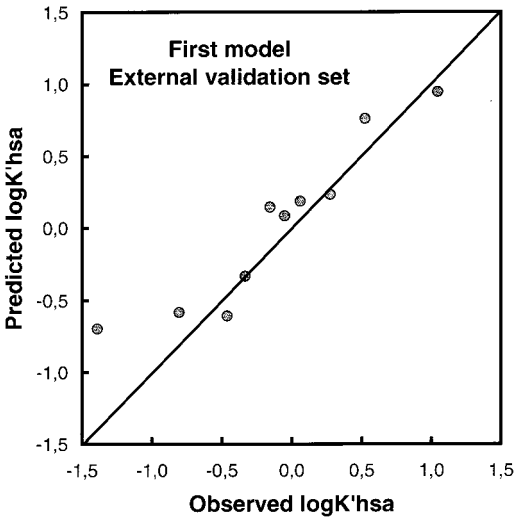

Figure 3. Observed vs predicted log K'hsa values plot for the first global model, obtained with the external validation set.

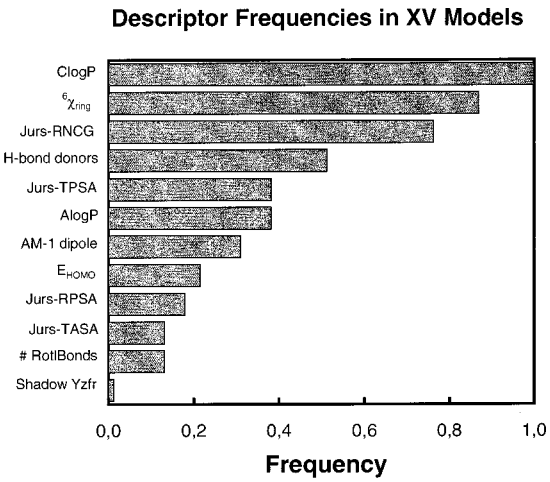

Figure 4. Frequency distribution of descriptors in the best model equations obtained from the global cross validation of the first global model.

not biasing the evolution toward equations with small numbers of terms also resulted in overfitted models with a lot of descriptors. Similar considerations were applied for other parameters: type of equation terms, mutation probabilities, and so on. The final parameters and running conditions are described below.

The GFA search and selection was performed in two steps. An initial run of GFA was performed with the training set, using the 53 descriptors. Given the large number of descriptors, as well as the use of multiple types of equation terms (see below), a big initial population of random equations (1000 models) was generated to search as much as possible the equation space. All these initial models had five terms, one of them a constant. Terms were of five types: linear, quadratic, spline, offset quadratic, and quadratic spline. The mutation probabilities for evolution were set as follows: 0.5 for adding a new term, 0.5 for shifting the spline knot, 0.25 for reducing the equation, and 0 for extending the equation. A smoothing parameter of 2 was used to bias toward smaller models. Equation lengths were not fixed. Least squares were used as the regression method. The system was allowed to evolve for a total of 100 000 generations, after which the resulting best model (as judged by the LOF) was

$$\log \text{K'hsa} = 0.020141 + (0.055367)(\text{AM1dip}) - (1.22294)(\text{JursRPSA}) - 0.028267(E_{\text{HOMO}} + 7.4076)^2 + (0.14905)(\text{ClogP}) - 3.48408(0.18539 - {}^6\chi_{\text{ring}}) \quad (4)$$

The resulting statistics for this model equation were as follows: LOF = 0.12; $r^2 = 0.78$; $q^2 = 0.73$; PRESS/SSY = 0.27; five outliers. (Both q^2 and PRESS/SSY where computed using eqs 2 and 3 above; they are shown in the "equation viewer" of QSAR+ inside Cerius2.) Therefore, a good fit (as judged by the large r^2 and the small LOF), with good predictive power (as judged by a q^2 close to r^2 and a PRESS/SSY well below 0.4) was obtained. Figure 2 shows a plot of the theoretical log K'hsa values vs the experimental ones.

The whole process was validated first by means of a randomization test, where the whole evolution was repeated nine times (to yield a 90% confidence level) after scrambling the log K'hsa values every time. As a result, all the derived models were clearly worse than the one above, having $\hat{r}_r = 0.30$, well below $r_{nr} = 0.88$. Moreover, the SD-r_r value was large: 3.27. This indicates that the obtained regression does not result from chance correlations but is the result of real dependencies.

In addition, the whole process was also further validated with a LOO cross-validation test. As with the randomization test, the same evolution was repeated 84 times after leaving out all the molecules, one each time. The 84 best models were used to predict the corresponding log K'hsa of the absent compound. This yielded a $q^2 = 0.81$, very close to the value obtained for the validated model.

Finally, the model was externally validated with the external set not used in the derivation of the model at any moment. The resulting actual vs predicted log K'hsa plot is shown in Figure 3. The values are very well correlated ($r^2 = 0.88$). This, together with the internal validation tests, demonstrates the good predictive power of the model and the absence of both overfitting and chance correlations in the model.

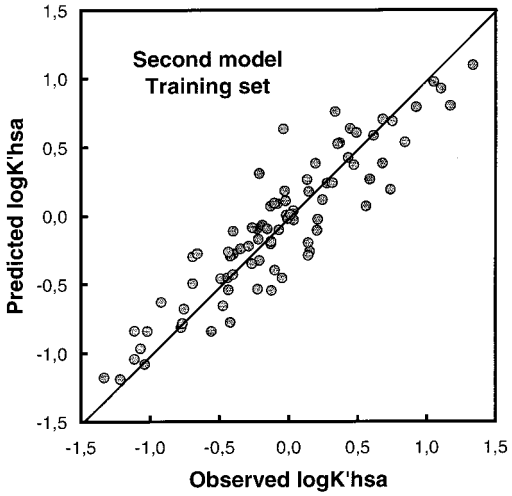

Figure 5. Observed vs predicted log K'hsa values plot for the second global model, obtained with the training set.

The cross-validation test described above was also useful for screening in a robust way the descriptors important for this system because they represent 84 independent evolutions with datasets lacking one different molecule each time. The descriptor frequencies were computed from the resulting 84 best models. Figure 4 shows these frequencies. It can be seen that the best models are made up of a subset of only 12 descriptors. ClogP is the most important parameter because it occurs in *all* the models, followed by ${}^6\chi_{ring}$. All the descriptors in the previous model belong to this subset of important descriptors.

Therefore, a second GFA step was taken, starting only with the subset of 12 important descriptors, to obtain an improved model. The conditions were the same as before except the system was allowed to evolve for 300 000 generations to achieve better convergence. The resulting best model was

$$\log \text{K'hsa} = -0.607873 + 0.06784(\text{HBondDon} - 3)^2 - (9 \times 10^{-6})(\text{JursTPSA}) - 0.028261(E_{\text{HOMO}} + 7.4076)^2 + (0.005697)(\text{AM1dip}^2) + (0.182595)(\text{ClogP}) + (2.33529)({}^6\chi_{ring}) \quad (5)$$

This model is similar to the previous one in that it keeps the most important parameters, ClogP and ${}^6\chi_{ring}$, together with others less important: AM-1 dipole moment, and HOMO energy. JursRPSA has been replaced with JursTPSA, and a new term, having a number of hydrogen bond donors as descriptor, appears.

The statistics for this model are as follows: LOF = 0.10; $r^2 = 0.83$; $q^2 = 0.79$; PRESS/SSY = 0.20. Therefore, both the goodness of the fit and the predictive power of the model from the cross validation have been improved compared with those from the previous model. Figure 5 displays a plot of the predicted log K'hsa vs the experimental one; Figure 6 displays the descriptor usage vs the number of generations in the evolution, where it can be seen that good convergence has actually been achieved.

A randomization test, with the same conditions as the previous one, was also conducted to validate the model generation with the subset of important descriptors. The

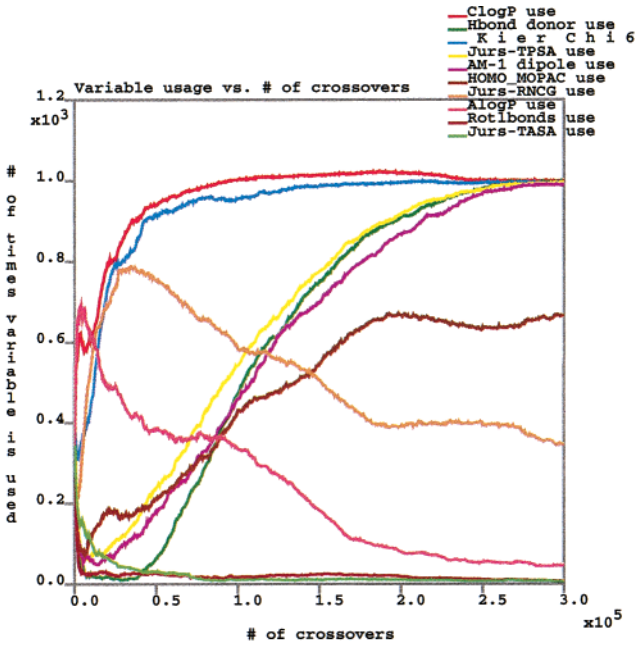

Figure 6. Variable usage vs genetic algorithm generations for the second global model. Selection of model descriptors during evolution can be seen, as well as the achieved convergence.

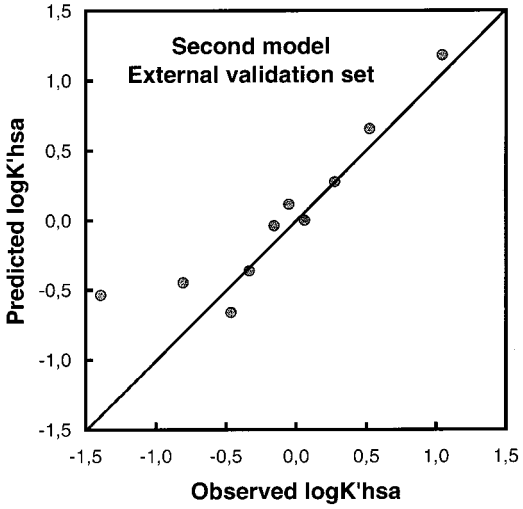

Figure 7. Observed vs predicted log K'hsa values plot for the second global model, obtained with the external validation set.

results showed again that the random models were clearly worse, with $\hat{r} = 0.25$ ($r_{nr} = 0.92$), and a large SD-$r_r = 4.57$. Thus, it is even more unlikely that this model results from chance correlations.

Finally, the model was used to predict the log K'hsa of the external validation set. Again, the predicted and observed log K'hsa values are very well correlated ($r^2 = 0.82$, slightly worse than those from the previous model). Figure 7 displays a plot of the actual vs predicted log K'hsa values for this second model.

In summary, two models with good predictive power have been worked out that should be useful to successfully predict binding affinities to HSA for new drugs of any family from their structure. The models cover a wide range of retention times (between 2 and 56 min) most optimally. Table 3 lists the range of values displayed by the descriptors comprising the two models in the dataset used to derive the models; inside these

Table 3. Ranges of Values of Model Descriptors in the Training Set

descriptor	minimum value	maximum value
ClogP	−1.87	7.48
$^6\chi_{ring}$	0.00	0.33
JursRPSA	0.00	0.67
JursTPSA	0.0	428.3
E_{HOMO}	−10.26	−7.41
Am-1 dipole moment	0.66	8.89
hydrogen bond donors	0	7

ranges the reliability of the predictions are greater because they should come from interpolations of the model, not from extrapolations.

Implications for Drug−HSA Binding Studies and Drug Design. This work presents a new set of 95 binding constants of drugs and druglike compounds to HSA, systematically obtained with the same technique and under the same conditions. Therefore, it represents a rich source of information about drug−HSA binding that can be useful for extracting general conclusions about the forces that stabilize these interactions. This information, in turn, should allow the rationalization of the design of new drugs, as far as their HSA binding is concerned.

As stated above, GFA generates equation models for regression. This is an advantage when one is interested not only in predicting but also in *understanding* the mechanism behind the modeled phenomenon. The descriptor frequencies analysis described above is therefore very informative because it provides the variables important to describe the behavior of this system.

In this respect, it is clear that hydrophobicity increases drug binding to HSA because *all* the models contain a term proportional to ClogP and some models a term proportional to AlogP too (not shown). This has also been observed previously in other models of limited sets of compounds, like 1,4-benzodiazepines,[19] 2,3-substituted 3-hydroxypropionic acids,[20] and also in a heterogeneous set,[46,47] where also hydrophobic terms display a prevailing role. This is supported by the X-ray structures of HSA, both alone and bound to different ligands.[1,4−6] These structures show both sites I and II made up mainly of hydrophobic residues and also that drug binding is stabilized to a large (if not primary) extent by hydrophobic interactions. This allows us to suggest this type of interaction to be probably the most important for drug−HSA binding. From a drug-design point of view, 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 also result in an opposing effect of other types of interactions that affect binding (see below).

Both the model equations and the descriptor frequencies analysis indicate that there are other factors modulating the binding strength to HSA, which are mainly of geometric or shape nature. $^6\chi_{ring}$, JursRPSA, and JursTPSA are probably the most important ones in the models. $^6\chi_{ring}$ is a sixth-order, ring type Kier and Hall topological index. Its presence in the models is probably reflecting the effect that the nature of six-membered rings in the drug can have in the interaction: heteroatoms in the ring, substituents present in the ring, etc. This topological index especially describes the degree of substitution or branching in six-membered rings such that high values of $^6\chi_{ring}$ correspond to molecules with many nonsubstituted atoms in six-membered rings. The equations show a direct proportionality of log K'hsa with $^6\chi_{ring}$ and therefore indicate that molecules with nonsubstituted six-membered rings are expected to bind more tightly to HSA than others with no six-membered rings or highly branched rings.

On the other hand, Jurs descriptors are obtained by mapping partial charges on solvent-accessible surface areas of particular atoms. In particular, JursTPSA is the sum of solvent-accessible surface areas of atoms with the absolute value of partial charges greater than or equal to 0.2. JursRPSA is JursTPSA divided by the total molecular solvent-accessible surface area. The corresponding terms in the model equations indicate that log K'hsa is inversely proportional to these descriptors; therefore, binding is favored for molecules with large nonpolar surfaces. This echoes the importance of hydrophobicity in binding.

Finally, other additional factors, like hydrogen bonds, number of rotatable bonds, and HOMO energy, can be important in determining HSA binding extent.

In summary, drug binding to HSA seems to be driven by hydrophobic interactions that should be modulated by structural factors, which are modeled by different terms in the two global models presented here.

References

(1) Carter, D. C.; Ho, J. X. Structure of Serum Albumin. *Adv. Protein Chem.* **1994**, *45*, 152−203.
(2) Rang, H. P.; Dale, M. M.; Ritter, J. M. *Pharmacology*; Churchill Livingstone: Edinburgh, 1999.
(3) Hervé, F.; Urien, S.; Albengres, E.; Duché, J.-C.; Tillement, J. Drug Binding in Plasma. A Summary of Recent Trends in the Study of Drug and Hormone Binding. *Clin. Pharmacokinet.* **1994**, *26*, 44−58.
(4) Carter, D. C.; He, X.-M.; Munson, S. H.; Twigg, P. D.; Gernert, K. M.; Broom, M. B.; Miller, T. Y. Three-dimensional Structure of Human Serum Albumin. *Science* **1994**, *244*, 1195−1198.
(5) Carter, D. C.; He, X.-M. Structure of Human Serum Albumin. *Science* **1990**, *249*, 302−303.
(6) He, X. M.; Carter, D. C. Atomic structure and chemistry of human serum albumin. *Nature* **1992**, *358*, 209−215.
(7) Ekins, S.; Obach, R. S. Three-dimensional quantitative structure activity relationship computational approaches for prediction of human in vitro intrinsic clearance. *J. Pharmacol. Exp. Ther.* **2000**, *295*, 463−473.
(8) Cronin, M. T. D. Computational methods for the prediction of drug toxicity. *Curr. Opin. Drug Discovery Dev.* **2000**, *3*, 292−297.
(9) Yoshida, F.; Topliss, J. G. QSAR Model for Drug Human Oral Bioavailability. *J. Med. Chem.* **2000**, *43*, 2575−2585.
(10) Dressman, J. B.; Amidon, G. L.; Fleisher, D. Absorption Potential: Estimating the Fraction Absorbed for Orally Administered Compounds. *J. Pharm. Sci.* **1985**, *74*, 588−589.
(11) Oprea, T. I.; Gottfries, J. Toward minimalistic modeling of oral drug absorption. *J. Mol. Graphics Modell.* **1999**, *17*, 261−274.
(12) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(13) Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. Predicting blood−brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43*, 2204−2216.
(14) Basak, S. C.; Gute, B. D.; Drewes, E. R. Predicting Blood−Brain Transport of Drugs: A Computational Approach. *Pharm. Res.* **1996**, *13*, 775−778.
(15) Cronin, M. T. D.; Dearden, J. C.; Moss, G. P.; Murray-Dickson, G. Investigation of the mechanism of flux across human skin in vitro by quantitative structure−permeability relationships. *Eur. J. Pharm. Sci.* **1999**, *7*, 325−330.
(16) Gute, B. D.; Grunwald, G. D.; Basak, S. C. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PHAs): a hierarchical QSAR approach. *SAR QSAR Environ. Res.* **1999**, *10*, 1−15.

(17) Yoshida, F.; Topliss, J. G. Unified Model for the Corneal Permeability of Related and Diverse Compounds with Respect to their Physicochemical Properties. *J. Pharm. Sci.* **1996**, *85*, 819−823.
(18) Hansch, C.; Leo, A. Exploring QSAR. *Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(19) Kaliszan, R.; Noctor, T. A. G.; Wainer, I. W. Quantitative Structure−Enantioselective Retention Relationships for the Chromatography of 1,4-Benzodiazepines on a Human Serum Albumin Based HPLC Chiral Stationary Phase: An Approach to the Computational Prediction of Retention and Enantioselectivity. *Chromatographia* **1992**, *33*, 546−550.
(20) Andrisano, V.; Bertucci, C.; Cavrini, V.; Recanatini, M.; Cavalli, A.; Varoli, L.; Felix, G.; Wainer, I. W. Stereoselective binding of 2,3-substituted 3-hydroxypropionic acids on an immobilised human serum albumin chiral stationary phase: stereochemical characterisation and quantitative structure−retention relationship study. *J. Chromatogr. A* **2000**, *876*, 75−86.
(21) *Tsar*, version 3.3; Oxford Molecular, Ltd.: Oxford, England, 2000.
(22) *QSAR+*, version 4.5; MSI: San Diego, CA, 2000.
(23) *Cerius2*; MSI: San Diego, CA, 2000.
(24) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM-1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(25) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(26) Gutman, I.; Ruscic, B.; Trinajstic, N.; Wilcox, C. F., Jr. Graph theory and molecular orbitals. XII. Acyclic polyenes. *J. Chem. Phys.* **1975**, *62*, 3399−3405.
(27) Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(28) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. *Rev. Comput. Chem.* **1991**, *2*, 367−422.
(29) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(30) Bonchev, D.; Trinajstic, N. Chemical information theory: structural aspects. *Int. J. Quantum Chem., Symp.* **1982**, *16*, 463−80.
(31) Stanton, D. T.; Jurs, P. C. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure−property relationship studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(32) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Anal. Chim. Acta* **1987**, *199*, 99−109.
(33) Ghose, A. K.; Pritchett, A.; Crippen, G. M. Atomic physicochemical parameters for three dimensional structure directed quantitative structure−activity relationships III: modeling hydrophobic interactions. *J. Comput. Chem.* **1988**, *9*, 80−90.
(34) Pearlman, R. S. *Physical Chemistry Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, Y. C., Eds.; Dekker: New York, 1980.
(35) Dewar, M. J.; Theil, W. Ground States of Molecules. 38. The MNDO Method. *J. Am. Chem. Soc.* **1997**, *99*, 44899−44907.
(36) Leo, A. J. *CLOGP*, version 3.63; Daylight Chemical Information Systems: Santa Fe, NM, 1997.
(37) Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional group contributions to drug−receptor interactions. *J. Med. Chem.* **1984**, *27*, 1648−1657.
(38) *Daylight*, version 4.51; Daylight Chemical Information System, Inc.: Santa Fe, NM, 1997.
(39) *QSAR+* Manual, version 4.5; MSI: San Diego, CA, 2000.
(40) Wold, S. Validation of QSAR's. *Quant. Struct.−Act. Relat.* **1991**, *10*, 191−193.
(41) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(42) Ivanciuc, O. Artificial neural networks applications. Part 4. Quantitative structure−activity relationships for the estimation of the relative toxicity of phenols for *Tetrahymena*. *Rev. Roum. Chim.* **1998**, *43*, 255−260.
(43) Ivanciuc, O. Artificial neural networks applications. Part 7. Estimation of bioconcentration factors in fish using solvatochromic parameters. *Rev. Roum. Chim.* **1998**, *43*, 347−354.
(44) Friedman, J. H. Multivariate Adaptive Regression Splines. *Ann. Statist.* **1991**, *19*, 1−141.
(45) Banzhaf, W.; Nordin, P.; Keller, R. E.; Francone, F. D. *Genetic Programming. An Introduction*; Morgan Kaufmann Publishers, Inc.: San Francisco, CA, 1998.
(46) Koizumi, K.; Ikeda, C.; Ito, M.; Suzuki, J.; Kinoshita, T.; Yasukawa, K.; Hanai, T. Influence of glycosylation on the drug binding of human serum albumin. *Biomed. Chromatogr.* **1998**, *12*, 203−210.
(47) Hanai, T.; Miyazaki, R.; Kinoshita, T. Quantitative analysis of human serum albumin−drug interactions using reversed-phase and ion-exchange liquid chromatography. *Anal. Chim. Acta* **1999**, *378*, 77−82.

JM010960B