Application of BCUT Metrics and Genetic Algorithm in Binary QSAR Analysis

Hua Gao†

Computer-Aided Drug Discovery, Pharmacia, 301 Henrietta Street, Kalamazoo, Michigan 49007

Received September 22, 2000

The application of three-dimensional H-suppressed BCUT metrics (BCUTs) in binary QSAR analysis was investigated using carbonic anhydrase II inhibitors and estrogen receptor ligands as test cases. Variable selection was accomplished with a genetic algorithm (GA). Highly predictive binary QSAR models were obtained for both sets of compounds within 200 GA generations. The derived binary QSAR models were validated with two sets of compounds not included in the training sets. The results indicate that BCUTs are very useful molecular descriptors, and the genetic algorithm is a very efficient variable selection tool in binary QSAR analysis.

INTRODUCTION

Combinatorial chemistry and high throughput screening are now firmly established as powerful techniques in modern drug discovery.[1] Database mining and virtual screening techniques, including different QSAR-type approaches and fast docking methods, play an important role in compound selection for high throughput screening and in focused or targeted combinatorial library design.[2−6] Traditional (2D/3D) QSAR methods are used to guide lead optimization and investigate action mechanisms of chemical−biological interactions (mechanistic QSAR).[7] In conventional QSAR analysis, biological activities are quantitatively expressed as a function of physicochemical properties of molecules. The application of QSAR is often difficult due to uncertainties involving molecular descriptor selection, in the case of the three-dimensional QSAR method, conformer generation, and alignment. It becomes even more difficult, if not impossible, when the QSAR method is used to analyze high throughput screening (HTS) data.[2] A binary QSAR methodology has been introduced recently,[8] in which biological activity expressed in a "binary" format (1 = active and 0 = inactive) is correlated with molecular descriptors of compounds, and a probability distribution for active and inactive compounds in a training set is estimated. The derived binary QSAR model can subsequently be used to predict the probability of new compound(s) to be active against a given biological target.

In previous binary QSAR analyses of carbonic anhydrase II (CA II) inhibitors[2] and estrogen receptor (ER) ligands,[9] three indictor variables, f_n (number of SO_2NH_2 groups) for binary QSAR model of CA II inhibitors and I−OH (equals to 1 for compounds containing phenolic OH) and I,es (equals to 1 for hexestrol derivatives) for ER ligands, had to be identified and used to describe structural features not captured by other molecular descriptors. Identifying indicator variables in QSAR analysis is difficult, if not impossible, especially in the analysis of HTS data that in most cases covers diverse chemical structures. Therefore, successfully deriving a meaningful binary QSAR model is highly dependent on the selection of molecular descriptors that can capture the crucial underlying structural features of chemical−biological interactions. Given the importance of molecular descriptors in binary QSAR analysis, searching for new molecular descriptors that can capture important structural features/properties is a ongoing effort. Since binary QSAR is intended for the analysis of HTS data, molecular descriptors that describe molecular diversity properties should be meaningful variables to be evaluated for application in binary QSAR analysis.

BCUT metrics are an extension of Burden's parameters which are based on a combination of the atomic number for each atom and a description of the nominal bond-type for adjacent and nonadjacent atoms and incorporate both connectivity information and atomic properties (e.g. atomic charge, polarizability, hydrogen bond abilities) relevant to intermolecular interactions.[12−15] Dependent on the choices of connectivity information, atomic information, and scaling factors controlling the relative balance of these two kinds of information, many BCUT metrics can be generated. The BCUT metrics have been used successfully in molecular diversity and related computational analysis. Since the BCUT metrics can capture sufficient structural features of molecules to yield useful measurement of molecular diversity, it is the intention of this study to explore the applicability of BCUT metrics in binary QSAR analysis.

METHODS

1. Biological Data. A set of 337 CA II inhibitors and a set of 463 ER ligands were collected from the literature.[2,9−11] The biological activity in CA II inhibitors was expressed as $\log 1/IC_{50}$ (or log1/Ki for some compounds), and for estrogen receptor ligands the biological activity was expressed as logRBA. RBA, "relative binding affinity", was calculated as a percentage of the ratio of IC_{50} values of the test compound to displace 50% of [^{3}H]estradiol from estrogen receptor binding. On the RBA scale, estradiol has a value of 100. The data profiles of the two data sets were summarized in Tables 1 and 2, respectively. For binary QSAR analysis, the continuous biological activity was transformed into a binary format (1 = active, 0 = inactive)

† Corresponding author phone: (616)833-4556; fax: (616)833-9183; e-mail: hua.gao@pharmacia.com.

10.1021/ci000306p CCC: $20.00
Published on Web 02/08/2001

Table 1. Data Profile of CA II Inhibitors

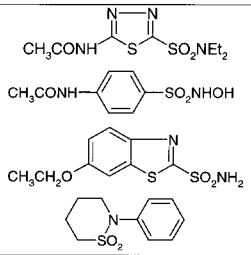
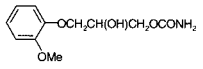
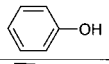
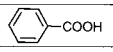
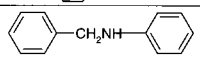
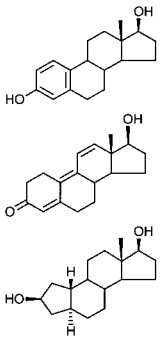
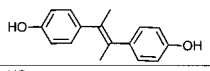
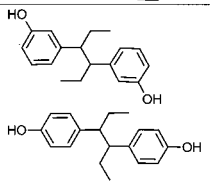
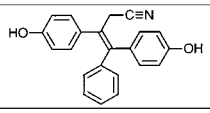
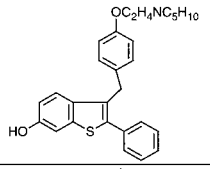
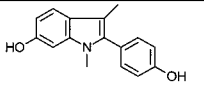
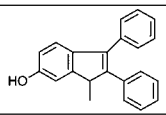
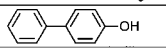
structural class	representative structures	number of compounds	range of log1/IC_{50}
sulfonamides		309	1.00 ~ 9.74
amides		19	0.00 ~ 2.80
alcohols and phenols		7	0.28 ~ 1.68
benzoic acid		1	0.25
amine		1	0.25

Table 2. Data Profile of ER Ligands

structural class	representative structures	number of compounds	range of log(RBA)
steroids		167	-1.00 ~ 2.60
diethylstilbestrol analogs		10	-0.10 ~ 2.46
hexstrol analogs		68	-2.00 ~ 2.48
tryphenylethylene analogs		40	-2.00 ~ 2.10
benzothiophene analogs		68	-0.70 ~ 1.61
indole analogs		61	-2.00 ~ 1.52
indene analogs		45	-2.02 ~ 2.47
phenols		4	-1.70 ~ 1.34

using a threshold criterion (6 of log1/IC_{50} or log1/Ki in the case of CA II inhibitors and 1.7 of logRBA in the case of estrogen receptor ligands).[2,8,9] Any compounds with biological activity higher than or equal to this criterion were classified as active, and any compounds with lower values were classified as inactive. The selection of a binary threshold value is arbitrary. On one hand, if the binary threshold value is too high, there will be too few "active" compounds to be selected; On the other hand, if the threshold value is too low, there will be too many "active" compounds to be selected. The effect of different binary threshold values on the binary QSAR model has been investigated in previous studies. It has been shown that different binary threshold values had little impact on overall predictive accuracy of a binary QSAR model.[8,9] Each set of compounds was divided into two subsets, a training set to derive a binary QSAR model and a test set to validate the derived binary QSAR model.

Table 3. Standard 3D H-Suppressed BCUT Metrics Calculated with DiverseSolutions

abbreviation	definition
BCUT_1	bcut_gastchrg_S_invdist2_0.05_R_H
BCUT_2	bcut_gastchrg_S_invdist2_0.08_R_H
BCUT_3	bcut_gastchrg_S_invdist2_1.25_R_L
BCUT_4	bcut_gastchrg_S_invdist2_2.75_R_L
BCUT_5	bcut_gastchrg_S_invdist6_0.60_R_H
BCUT_6	bcut_gastchrg_S_invdist6_2.25_R_L
BCUT_7	bcut_gastchrg_S_invdist_0.02_R_H
BCUT_8	bcut_gastchrg_S_invdist_1.50_R_L
BCUT_9	bcut_gastchrg_S_invdist_3.00_R_L
BCUT_10	bcut_haccept_S_invdist2_2.00_R_H
BCUT_11	bcut_haccept_S_invdist2_3.00_R_H
BCUT_12	bcut_haccept_S_invdist6_16.00_R_H
BCUT_13	bcut_haccept_S_invdist_0.60_R_H
BCUT_14	bcut_haccept_S_invdist_0.90_R_H
BCUT_15	bcut_hdonor_S_invdist2_1.20_R_H
BCUT_16	bcut_hdonor_S_invdist6_8.00_R_H
BCUT_17	bcut_hdonor_S_invdist_0.30_R_H
BCUT_18	bcut_hdonor_S_invdist_0.45_R_H
BCUT_19	bcut_tabpolar_S_invdist2_1.00_R_L
BCUT_20	bcut_tabpolar_S_invdist2_1.50_R_H
BCUT_21	bcut_tabpolar_S_invdist2_2.00_R_H
BCUT_22	bcut_tabpolar_S_invdist2_3.00_R_L
BCUT_23	bcut_tabpolar_S_invdist6_1.25_R_L
BCUT_24	bcut_tabpolar_S_invdist6_11.00_R_H
BCUT_25	bcut_tabpolar_S_invdist6_2.75_R_L
BCUT_26	bcut_tabpolar_S_invdist6_8.00_R_H
BCUT_27	bcut_tabpolar_S_invdist_0.50_R_H
BCUT_28	bcut_tabpolar_S_invdist_2.00_R_L
BCUT_29	bcut_tabpolar_S_invdist_4.00_R_L

2. Database Construction and Calculation of Molecular Descriptors. All the structures were sketched using MOE software[16] and stored into databases. The structures were minimized with the inclusion of PEOE partial atomic charges. A SD file was generated for transfer to the DiverseSolutions software[13] for the calculation of BCUT metrics.

The set of 29 standard 3D H-suppressed BCUT metrics (Table 3) was computed for each of the compounds in both data sets with DiverseSolutions.[13] The calculated BCUT values were imported into MOE database and used in the binary QSAR analysis. All other molecular descriptors (Table 4) were calculated using 1999.05 version of MOE.[16]

3. Binary QSAR Analysis. In this study, binary QSAR analysis was carried out using the MOE binary QSAR function.[16] A detailed description of the binary QSAR methodology has been introduced in previous publications[2,8,9] and briefly described here.

Binary QSAR estimates, from a training set of compounds, the probability density $\Pr(Y = 1|X = x)$ according to the

Table 4. Molecular Descriptors Calculated with MOE

symbol	description
b_ar	number of aromatic bonds
b_1rotR	fraction of rotatable single bonds
$^0\chi$	zero-order atomic connectivity index
$^1\chi$	first-order atomic connectivity index
$^2\chi$	second-order atomic connectivity index
$^0\chi^v$	zero-order atomic valence connectivity index
$^1\chi^v$	first-order atomic valence connectivity index
$^2\chi^v$	second-order atomic valence connectivity index
$^1\kappa$	Kier first shape index
$^2\kappa$	Kier second shape index
$^3\kappa$	Kier third shape index
$^1\kappa_\alpha$	Kier first alpha modified shape index
$^2\kappa\_\alpha$	Kier second alpha modified shape index
$^3\kappa\_\alpha$	Kier third alpha modified shape index
Φ	Kier molecular flexibility index
ASA_H	total accessible hydrophobic surface area
logP(o/w)	partition coefficient

following formula

$$\Pr(Y = 1|X = x) \approx \left[1 + \frac{\Pr(Y = 0)}{\Pr(Y = 1)} \prod_{i=1}^{p} \frac{\Pr(Z_i = z_i|Y = 0)}{\Pr(Z_i = z_i|Y = 1)}\right]^{-1},$$
$$Z = Q(\mathrm{X} - u) = (Z_1,...,Z_p)$$

where Y is a Bernoulli random variable (i.e. Y takes on value of 0 or 1) representing biological activity in a given biological assay, and X is a random n-vector of real numbers (a random collection of physicochemical properties or molecular descriptors of the molecules).

The original molecular descriptors of the training set are transformed using principal component analysis (PCA) (n by p linear transform) by Q and u to obtain a decorrelated and normalized set of descriptors. Each probability density $\Pr(Z_i = z_i)$ is estimated by constructing a histogram. To minimize the sensitivity of a conventional procedure for histogram construction to bin-boundaries, each observation is replaced with a Gaussian density with variance σ^2. This variance can be interpreted as an observation error or as a *smoothing factor*. Once all of the $2p + 2$ probability densities have been estimated from the training set, the desired density $\Pr(Y = 1|X = x)$ is constructed using the above formula. The derived binary QSAR model can subsequently be used to predict the probability of a new compound to be active.

The performance of a binary QSAR model was measured by predictive accuracy on active and inactive compounds.[8,9] The derived binary QSAR model was cross-validated by the leave-one-out procedure.[17]

4. Variable Selection. In binary QSAR analysis, variable selection is another important factor. We either overwhelm the binary QSAR algorithm thus decreasing the predictive ability of the binary QSAR model by using too many molecular descriptors or cannot derive a meaningful model by not using enough descriptors thus not capturing sufficient structural properties of intermolecular interactions. Therefore, selecting a preferred set of molecular descriptors is crucial to obtain a highly predictive and robust binary QSAR model. Genetic algorithm (GA) is a novel algorithm rooted in Darwin's theory[18] and attracts much attention as a key solution to various optimization problems in many fields including QSAR analysis.[19−23] In this study, variable selection is achieved by using a genetic algorithm developed in our group. In this method, binary QSAR is used as a statistical method, and best variable combinations are selected by GA using a cross-validated predictive accuracy of binary QSAR model as a scoring function. In this procedure, the initial population of chromosomes is created by setting all bits in each chromosome to a random value (1 or 0). Bit "1" denotes a selection of a variable, and bit "0" denotes a nonselection. The fitness of each chromosome is evaluated by the cross-validated predictive accuracy of the binary QSAR model using the leave-one-out procedure. The chromosome with the least number of variables and the highest fitness is chosen as the best chromosome in a generation. The best chromosome is protected and is chosen to survive and will be replaced if a chromosome with a lower number of variables gives a better fitness in the next generation. The chromosomes with highest fitness are selected from the population in an arbitrary proportion as a good population. A new population is generated by uniform crossover and mutation of the good population (except the best chromosome). The cycle is repeated until the number of generations reaches the given maximum.

5. Molecular Diversity. In this study, molecular diversity of compounds was analyzed using Tanimoto coefficients (T_c). Average T_c was calculated using FP:MACCS fingerprint implemented in MOE[16,24] according to the following formula

$$T_c = B_c/(B_1 + B_2 - B_c)$$

where B_c is the number of common bits, and B_1 and B_2 are the bits in the fingerprints of molecules 1 and 2, respectively.

RESULTS AND DISCUSSION

1. Binary QSAR Analysis of Carbonic Anhydrase II Inhibitors. A set of 287 compounds was chosen as a training set to derive a binary QSAR model. The range of the biological activities for the training set is −0.20 to 9.30 log units. A value of 6.0 was chosen as a binary threshold value. Based on this threshold criterion, 50 compounds are inactive and 237 are active in the training set. The CA II inhibitors analyzed are a very diverse set of compounds with an average Tanimoto coefficient of 0.55 ± 0.17 (average ± SD). The average Tanimoto coefficient for sulfonamides, the major class of compounds, is 0.60 ± 0.11. The average Tanimoto coefficients indicate the high diversity of this set of compounds even within the sulfonamide group of compounds. There are several structural templates in the sulfonamide class (see Table 1).

With an initial population of 200 chromosomes, a good population of 20 chromosomes, a uniform crossover rate of 0.5, and a mutation rate of 0.05, a highly predictive binary QSAR model was obtained with a maximum GA generation of 200. The cross-validated predictive accuracy vs the number of GA generations is plotted in Figure 1. From the plot, it can be seen that the cross-validated predictive accuracy on actives increased from 0.82 to 0.91 in about 70 GA generations. The result indicates that the genetic algorithm is a very efficient method for variable selection or optimization in the binary QSAR analysis.

The best binary QSAR model was obtained with a combination of 23 molecular descriptors including 4 connectivity indexes ($^2\chi$, $^0\chi^v$, $^1\chi^v$, and $^2\chi^v$), 3 shape indexes ($^2\kappa$,

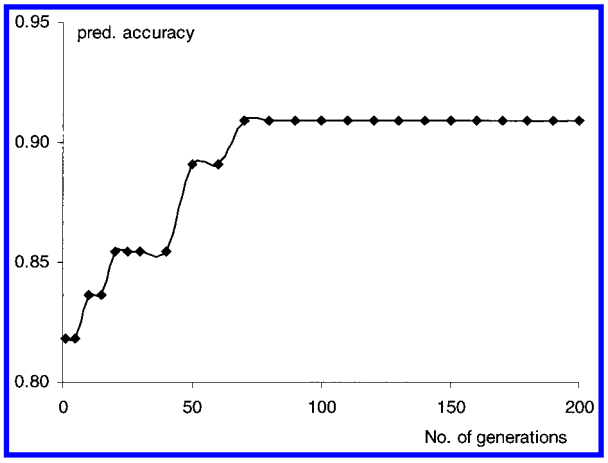


Figure 1. Plot of cross-validated predictive accuracy on actives vs number of GA generations in the analysis of CA II inhibitors..

Table 5. Descriptor Importance of Molecular Descriptors in the Binary QSAR Model of CA II Inhibitors

descriptor	importance	descriptor	importance
$^{1}\kappa_{\alpha}$	0.17	$^{1}\chi^{v}$	0.11
logP(o/w)	0.15	BCUT_24	0.11
$^{2}\chi$	0.13	$^{2}\chi^{v}$	0.11
$^{3}\kappa_{\alpha}$	0.13	BCUT_18	0.11
$^{2}\kappa$	0.13	BCUT_11	0.11
BCUT_26	0.12	BCUT_5	0.11
$^{0}\chi^{v}$	0.12	BCUT_14	0.11
BCUT_3	0.12	BCUT_7	0.11
BCUT_6	0.12	BCUT_1	0.11
BCUT_8	0.11	BCUT_2	0.10
BCUT_20	0.11	BCUT_29	0.09
BCUT_15	0.11		

$^{1}\kappa_{\alpha}$, $^{3}\kappa_{\alpha}$),[25,26] logP(o/w), and 15 BCUTs (BCUT_1, _2, _3, _5, _6, _7, _8, _11, _14, _15, _18, _20, _24, _26, and _29). The non-cross-validated predictive accuracy is 94% on active compounds, 94% on inactive compounds, and 94% for all the compounds. The cross-validated accuracy is 91% on active compounds, 92% on inactive compounds, and 91% for all the compounds. Thus, the predictive power of the binary QSAR model is quite high. Since binary QSAR is highly nonlinear, there is no simple way to describe the descriptor coefficient or descriptor contribution as in other QSAR methods. The descriptor importance of a molecular descriptor in a binary QSAR model gives the degree (between 0 and 1) to which each descriptor is useful in distinguishing actives from inactives. The descriptor importance of the molecular descriptors used in the binary QSAR model is summarized in Table 5. The result indicates that BCUTs are equally, if not more, important as other molecular descriptors. The summation of the importance of BCUTs in the binary QSAR model is significantly higher than other descriptors because of the number of BCUTs contained in the binary QSAR model. Since the high structural diversity of the compounds analyzed, none of the descriptors could capture significant structural properties alone to distinguish actives from inactives.

Further analyses were conducted to compare the usefulness of different molecular descriptors used. The binary QSAR models derived using different sets of molecular descriptors are summarized in Table 6. The binary QSAR model derived using BCUTs alone is nearly as good as the one derived using the combination of all the descriptors for this set of compounds. This result directly reveals the usefulness of BCUT metrics in binary QSAR analysis.

To evaluate the predictive value of the derived binary QSAR model, 50 known CA II inhibitors not included in the training set were analyzed. The range of the biological activity for the test set of compounds is −0.28 to 9.74. Based on the estimation of the binary QSAR model, 9 out of 10 (90%) inactive compounds were correctly predicted, and 39 out of 40 active compounds (98%) were correctly predicted (overall accuracy of 96%), consistent with the cross-validation result.

The overall prediction accuracy of the binary QSAR model is very high (94%). However, careful examination of the binary QSAR model revealed that prediction accuracy was significantly lower for compounds with biological activity near the binary threshold value. The accuracy for compounds with log1/IC_{50} or log1/Ki values below 5 is 100%, for compounds with activity value between 5.5 and 6.5 is only 74%, for compounds with activity value between 6.5 and 7.5 is 91%, and for compounds with activity value above 7.5 is 97%.

2. Binary QSAR Analysis of Estrogen Receptor Ligands. A set of 400 compounds was chosen as a training set to derive a binary QSAR model. The range of the biological activity (logRBA) for the training set is −2.02 to 2.48. A value of 1.7 of logRBA, which corresponds to 50% of RBA, was chosen as a binary threshold value. Based on this threshold criterion, 64 compounds are active and 336 are inactive in the training set.

The estrogen receptor ligands are composed of several classes of compounds (see Table 2). The diversity of this set of compounds is very high with an average Tanimoto coefficient of 0.49 ± 0.18 calculated using the MACCS fingerprint.

With the same combination of GA parameters, a reasonable binary QSAR model was obtained. The cross-validated predictive accuracy on active compounds vs the number of GA generations is plotted in Figure 2. One can see that the cross-validated predictive accuracy was improved dramatically within 70 GA generations, indicating the usefulness of the GA variable selection method in the binary QSAR analysis.

Table 6. Comparison of Binary QSAR Models Derived Using Different Sets of Descriptors of CA II Inhibitors

QSAR model		MOE molecular descriptors only	BCUT metrics only	combination of descriptors
predictive accuracy	actives	0.94 (0.92)[a]	0.93 (0.89)	0.94 (0.91)
	inactives	0.72 (0.58)	0.94 (0.92)	0.94 (0.92)
	overall	0.90 (0.86)	0.93 (0.89)	0.94 (0.91)
descriptors used		logP(o/w), b_ar, b_1rotR, $^{1}\chi$, $^{2}\chi$, $^{0}\chi^{v}$, $^{1}\chi^{v}$, $^{2}\chi^{v}$,$^{2}\kappa$, $^{3}\kappa$, $^{3}\kappa_{\alpha}$	BCUT-2, 3, 4, 5, 6, 7, 8,10, 11, 12, 13, 14, 17, 29	$^{2}\chi$, $^{0}\chi^{v}$, $^{1}\chi^{v}$, $^{2}\chi^{v}$, $^{2}\chi^{v}$, $^{1}\kappa_{\alpha}$, $^{3}\kappa_{\alpha}$, BCUT-1, 2, 3, 5, 6, 7, 8, 11, 14, 15, 18, 20, 24, 26, 29

[a] Value in parentheses is cross-validated accuracy.

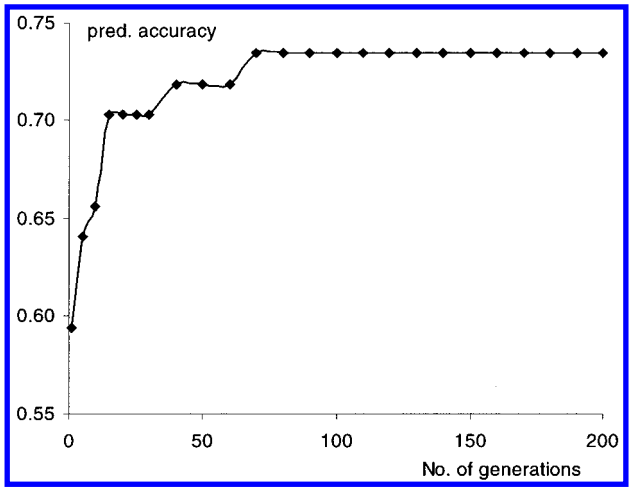


Figure 2. Plot of cross-validated predictive accuracy on actives vs number of GA generations in the analysis of ER ligands.

Table 7. Descriptor Importance of Molecular Descriptors in the Binary QSAR Model of ER Ligands

descriptor	importance	descriptor	importance
logP(o/w)	0.13	BCUT_18	0.08
$^{1}\kappa_{\alpha}$	0.11	$^{2}\chi$	0.08
$^{2}\kappa_{\alpha}$	0.10	BCUT_19	0.08
$^{1}\chi^{v}$	0.10	BCUT_10	0.08
BCUT_8	0.10	BCUT_17	0.08
BCUT_25	0.10	BCUT_4	0.08
$^{0}\chi$	0.09	BCUT_28	0.08
BCUT_2	0.09	BCUT_13	0.08
$^{1}\chi$	0.09	BCUT_16	0.07
BCUT_7	0.08	Φ	0.07
BCUT_9	0.08	BCUT_14	0.06
BCUT_15	0.08	BCUT_21	0.06

The best binary QSAR model was obtained with a combination of 24 molecular descriptors including 4 connectivity indexes ($^{0}\chi$, $^{1}\chi$, $^{2}\chi$, and $^{1}\chi^{v}$), 2 shape indexes ($^{1}\kappa_{\alpha}$, $^{2}\kappa_{\alpha}$), flexibility index (Φ), logP(o/w), and 16 BCUTs (BCUT_2, _4, _7, _8, _9, _10, _13, _14, _15, _16, _17, _18, _19, _21, _25, and _28). The non-cross-validated accuracy of the derived binary QSAR model is 80% on active compounds, 90% on inactive compounds, and 89% for all the compounds. The cross-validated accuracy is 73% on active compounds, 90% on inactive compounds, and 88% for all the compounds. The importance of molecular descriptors (see Table 7) revealed the similar fact that BCUTs are important in the binary QSAR model for this set of compounds.

The binary QSAR models derived using different sets of molecular descriptors are summarized in Table 8. A pretty good binary QSAR model could be obtained using BCUTs alone. The results also indicate the usefulness of BCUTs in the binary QSAR analysis for this set of compounds.

In the validation test, 63 known estrogen receptor ligands not included in the training set were analyzed. Based on the prediction of the binary QSAR model, five out of seven (71%) active compounds were correctly predicted, and 51 out of 56 inactive compounds (91%) were correctly predicted, consistent with the cross-validation result.

A similar boundary effect was also observed with the binary QSAR model of estrogen receptor ligands. The overall prediction accuracy of the binary QSAR model is very high (90%). The accuracy for compounds with logRBA values below 1 is 95%, for compounds with logRBA values between 1 and 1.6 is 85%, for compounds with logRBA values between 1.6 and 1.8 is 63%, and for compounds with logRBA values above 1.8 is 82%.

The two data sets analyzed have different data profiles. In the data set of carbonic anhydrase II inhibitors, 83% of the compounds in the training set are active, while in the data set of estrogen receptor ligands, 16% of the compounds in the training set are active. In both cases, highly predictive and stable binary QSARs were obtained. The results indicate that the binary QSAR method is a reliable statistical tool irrespective of the data profile to be analyzed and that the BCUTs are very useful molecular descriptors in the binary QSAR analysis. The binary QSAR method has been successfully used in several HTS data analyses in Pharmacia (data not shown). Our experience has shown that the quality of HTS data had the most impact on the quality of a binary QSAR model, and the unbalanced nature of HTS data in terms of active vs inactive compounds had much less impact on a binary QSAR model. This has also been observed in other investigations.[8,9] Like other QSAR methods, the binary QSAR methodology cannot be expected to be successful for all the cases of HTS data analysis.

3. Comparison with Previous Studies. In the previous studies, a binary QSAR model with a predictive accuracy of 96% on actives, 82% on inactives, and an overall accuracy of 94% was derived for CA II inhibitors. The cross-validated accuracy for this model was 96% on actives, 82% on inactives, and 93% for all the compounds. To successfully derive the binary QSAR model, an indicator variable (f_n, number of SO_2NH_2 groups) had to be identified and used in the binary QSAR model.[2] In the case of ER ligands, a derived binary QSAR model had a predictive accuracy of 85% on actives, 93% on inactives, and 92% for all the compounds. The cross-validated accuracy was 76% on actives, 93% on inactives, and 90% for all the compounds. Two indicator variables, I,OH (for phenolic group) and I,es (for hexstrol analogues) were used.[9] The quality of the derived binary QSAR models in this study is as good as the ones in the previous investigations. However, comparing with the previous binary QSAR analyses of the two data sets, the inclusion

Table 8. Comparison of Binary QSAR Models Derived Using Different Sets of Descriptors of ER Ligands

QSAR model		MOE molecular descriptors only	BCUT metrics only	combination of descriptors
predictive accuracy	actives	0.58 (0.48)[a]	0.78 (0.67)	0.80 (0.73)
	inactives	0.91 (0.89)	0.85 (0.84)	0.90 (0.90)
	overall	0.86 (0.83)	0.84 (0.81)	0.89 (0.88)
descriptors used		logP(o/w), $^{2}\chi$, $^{0}\chi^{v}$, $^{2}\chi^{v}$, $^{1}\kappa$, $^{2}\kappa$, $^{3}\kappa$, $^{3}\kappa_{\alpha}$ Φ, b_1rotR, ASA_H	BCUT-3, 9, 10, 11, 13, 15, 16, 17, 19, 20, 22, 23, 25, 28, 29	$^{0}\chi$, $^{1}\chi$, $^{2}\chi$, $^{1}\chi^{v}$, $^{1}\kappa_{\alpha}$, $^{2}\kappa_{\alpha}$, BCUT-2, 4, 7, 8, 9, 10, 13, 14, 15, 16, 17, 18, 19, 21, 25, 28

[a] Value in parentheses is the cross-validated accuracy.

of BCUT metrics in this study eliminated the necessity to identify indicator variables thus enhancing the flexibility of applying binary QSAR method in HTS data analysis in which identifying indictor variables is difficult, if not impossible.

In conclusion, BCUT metrics capture sufficient structural information of intermolecular interactions thus enabling the elimination of indicator variables in previously derived binary QSAR models. Combined with GA-based variable selection method, the function of binary QSAR in the analysis of high throughput screening data would be significantly enhanced.

ACKNOWLEDGMENT

The author likes to thank Prof. R. S. Pearlman and Dr. Veerabahu Shanmugasundaram for their help in the calculation of BCUT metrics using DiverseSolutions.

REFERENCES AND NOTES

(1) Dolle, R. E. Discovery Of Enzyme Inhibitors Through Combinatorial Chemistry. *Mol. Div.* **1996**, *2*, 223−236.
(2) Gao, H.; Bajorath, J. Comparison Of Binary And 2D QSAR Analyses Using Inhibitors Of Human Carbonic Anhydrase II As A Test Case. *Mol. Diver.* **1999**, *4*, 115−130.
(3) Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction Of A Virtual High Throughput Screen By 4D-QSAR Analysis: Application To A Combinatorial Library Of Glucose Inhibitors Of Glycogen Phosphorylase b**.** *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1151−1160.
(4) van Drie, J. H.; Rohrer, D. C.; Blinn, J. R.; Gao, H. *3*D Database Searching And Related Methods For The Structure-Based Design Of Combinatorial Libraries. In *Modern Methods of Drug Discovery*; Hilgenfeld, R., Hillisch, A., Eds.; Springer-Verlag: in press.
(5) Gussio, R.; Pattabriraman, N.; Kellogg, G. E.; Zaharevitz, D. W. Use Of 3D QSAR Methodology For Data Mining The National Cancer Institute Repository Of Small Molecules: Application To HIV-1 Reverse Transcriptase Inhibition. *Methods.* **1998**, *14*, 255−263.
(6) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paul, K. D.; Friend, S. H.; Weinstein, J. N. Mining The NCI Anticancer Drug Discovery Databases: Genetic Function Approximation For The QSAR Study Of Anticancer Ellipticine Analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.
(7) Hansch, C.; Hoekman, D.; Gao, H. Comparative QSAR: Toward A Deeper Understanding Of Chemicobiological Interactions. *Chem. Rev.* **1996**, *96*, 1045−1075.
(8) Labute, P. Binary QSAR: A New Method For The Determination Of Quantitative Structure−Activity Relationships. In *Pacific Symposium On Biocomputing '99*; Altman, R. B., Dunker, A. K., Hunter, L., Klein, T. E., Lauderdale, K., Eds.; World Scientific: NJ, 1999; p 444.
(9) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary-QSAR Analysis Of Estrogen Receptor Ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.
(10) Lien, E. J.; Hussain, M.; Tong, G. L. Role Of Hydrophobic Interactions In Enzyme Inhibition By Drugs. *J. Pharm. Sci.* **1970**, *59*, 865−868.
(11) Gao, H.; Katzenellenbogen, J. A.; Garg, R.; Hansch, C. Comparative QSAR Analysis Of Estrogen Receptor Ligands. *Chem. Rev.* **1999**, *99*, 723−744.
(12) Burden, F. R. Molecular Identification Number For Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(13) Pearlman, R. S.; Smith, K. M. In *3D-QSAR and Drug Design: Recent Advances*; Kubinyi, H., Martin, Y., Folkers, G., Eds.; Kluwer Academic: Dordrecht, Netherlands, 1997; p 339.
(14) Pearlman, R. S.; Smith, K. M. Metric Validation And The Receptor-Related Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.
(15) Stanton, D. T. Evaluation And Use Of BCUT Metrics In QSAR And QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(16) Chemical Computing Group Inc. MOE 1998.03; 1255 University Street, Montreal, Quebec, Canada.
(17) Cramer, R. D.; Bunce, J. D.; Patterson, D. E.; Frank, I. E. Crossvalidation, Bootstrapping, And Partial Least Squares Compared With Multiple Regression In Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18−25.
(18) *Practical Genetic Algorithms*; Haupt, R. L., Haupt, S. E., Eds.; John Wiley & Sons: New York, 1998.
(19) Syswerda, G. Genetic Algorithms And Their Applications. In *Handbook of Genetic Algorithms*; Davis, L., Ed.; New York: Van Nostrand Reinhold: 1991; p 332.
(20) Rogers, D.; Hopfinger, A. J. Application Of Genetic Function Approximation To Quantitative Structure−Activity Relationships And Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(21) Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. Application Of Genetic Algorithms On The Structure−Activity Relationship Analysis Of Some Cinnamamides. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 775−781.
(22) Hasegawa, K.; Kimura, T.; Funatsu, K. GA Strategy For Variable Selection In QSAR Studies: Enhancement Of Comparative Molecular Binding Energy Analysis By GA-Based PLS Method. *Quant. Struct.−Act. Relat.* **1999**, *18*, 262−272.
(23) Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA Strategy For Variable Selection In QSAR Studies: GA-Based PLS Analysis Of Calcium Channel Antagonists. *J. Chem. Inf. Comput. Sci.* **1997**, 3066−310.
(24) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification Of A Preferred Set Of Molecular Descriptors For Compound Classification Based On Principal Component Analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699−704.
(25) Kier, L. B.; Hall, L. H. The Nature Of Structure−Activity Relationships And Their Relative To Molecular Connectivity. *Eur. J. Med. Chem.* **1997**, *12*, 307−312.
(26) Kier, L. B. Indexes Of Molecular Shape From Chemical Graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.

CI000306P