ELSEVIER

Journal of Molecular Graphics and Modelling 20 (2002) 259–268

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Enhancement of binary QSAR analysis by a GA-based variable selection method

Hua Gao*, Michael S. Lajiness, John Van Drie

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

Received 12 March 2001; accepted 12 March 2001

Abstract

Binary quantitative structure–activity relationship (QSAR) is an approach for the analysis of high throughput screening (HTS) data by correlating structural properties of compounds with a "binary" expression of biological activity (1 = active and 0 = inactive) and calculating a probability distribution for active and inactive compounds in a training set. Successfully deriving a predictive binary or any QSAR model largely depends on the selection of a preferred set of molecular descriptors that can capture the chemico–biological interaction for a particular biological target. In this study, a genetic algorithm (GA) was applied as a variable selection method in binary QSAR analysis. This GA-based variable selection method was applied to the analysis of three diverse sets of compounds, estrogen receptor (ER) ligands, carbonic anhydrase II inhibitors, and monoamine oxidase (MAO) inhibitors. Out of a variable pool of 150 molecular descriptors, predictive binary QSAR models were obtained for all three sets of compounds within a reasonable number of GA generations. The results indicate that the GA is a very effective variable selection approach for binary QSAR analysis.

Keywords: Binary QSAR; Genetic algorithm; CA II inhibitors; Estrogen receptor; MAO inhibitors; Variable selection

1. Introduction

Since Hansch's seminal work on quantitative structure–activity relationship (QSAR) analysis [1,2] many different QSAR methods including 2D and 3D QSAR approaches have been developed. These QSAR methods have been used to guide lead optimization and study action mechanisms of chemical–biological interactions (mechanistic QSAR) in modern drug discovery [3]. In most conventional QSAR methods, a linear relationship between a biological activity and molecular properties is assumed, which in most situations is a valid assumption for relatively small and congeneric sets of compounds. The linearity assumption may not hold true for large and diverse sets of data. The advent of combinatorial chemistry and high throughput screening (HTS) has greatly challenged conventional QSAR approaches. HTS produces a large amount of screening data, which in most cases identifies compounds as either active or inactive. In addition, compounds in HTS assay have diverse structures which makes it difficult, if not impossible, to analyze HTS data using conventional QSAR methods and make reliable predictions. A binary QSAR methodology has been introduced by Labute [4], 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. This binary QSAR method has been successfully used in several investigations [5,6].

Since hundreds of molecular descriptors are available for QSAR analysis and only a subset of them is statistically significant in terms of correlation with biological activity for a particular QSAR analysis, deriving an optimal binary QSAR model through variable selection needs to be addressed. Several variable selection methods including generalized simulated annealing [7,8], genetic algorithms (GAs) [9–13], and evolutionary algorithms [14–16] have been investigated for use with conventional QSAR methods. The results suggest that QSAR models derived with variable selection can have higher quality compared to those without variable selection. These considerations gave us an impetus for investigating the applicability of GAs as variable selection method in binary QSAR analysis.

* Corresponding author. Tel.: +1-616-833-4556; fax: +1-616-833-9183.
E-mail address: hua.gao@pharmacia.com (H. Gao).

1093-3263/02/$ – see front matter
PII: S1093-3263(01)00122-X

2. Methods

2.1. Biological data

2.1.1. Estrogen receptor (ER) ligands

A set of 463 ER ligands was collected from the literature [6,17]. The biological activity was expressed as log RBA. Relative binding affinity (RBA), was calculated as a percent from the ratio of IC_{50} values of test compounds to that of estradiol to displace 50% of [^{3}H]-estradiol from ER binding. On the RBA scale, estradiol has a value of 100. Fig. 1 shows some of the representative compounds for this data set. For binary QSAR analysis, the continuous biological activity was transformed into a binary format (1 = active, 0 = inactive) using a threshold value of 1.7 for log RBA. 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 the 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 binary QSAR model has been investigated in previous studies [4,6]. It has been shown that different binary threshold values had little impact on overall predictive accuracy of a binary QSAR model [8,9]. The set of compounds was randomly divided into two subsets, a training set of 400 compounds to derive the binary QSAR model and a test set of 63 compounds to test the derived binary QSAR model.

2.1.2. Carbonic anhydrase II (CA II) inhibitors

A set of 337 CA II inhibitors was collected from the literature [5]. This set of compounds covers a very diverse range of structural features (Fig. 2). The biological activity of the CA II inhibitors was expressed as log $1/IC_{50}$. A binary threshold value of 6 was used to classify compounds as active and inactive in the analysis. The set of compounds was also randomly divided into a training set of 280 compounds and a test set of 57 compounds.

2.1.3. Monoamine oxidase (MAO) inhibitors

A set of 1608 MAO inhibitors was used in this analysis and was kindly provided by Dr. Yvonne C. Martin of Abbott [18]. This set of compounds includes a number of classes

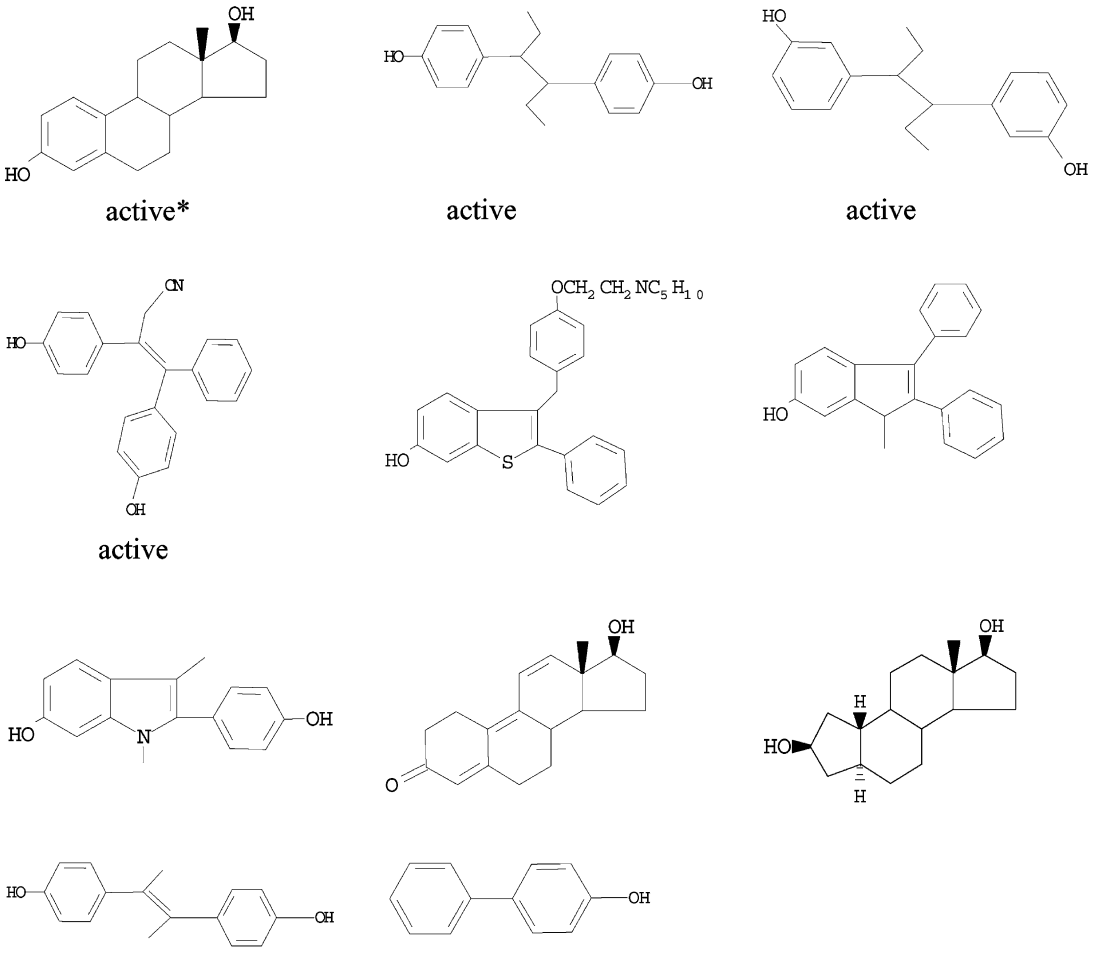

Fig. 1. Representative structures of ER ligands (compounds marked with (∗) were classified as active in the analysis).

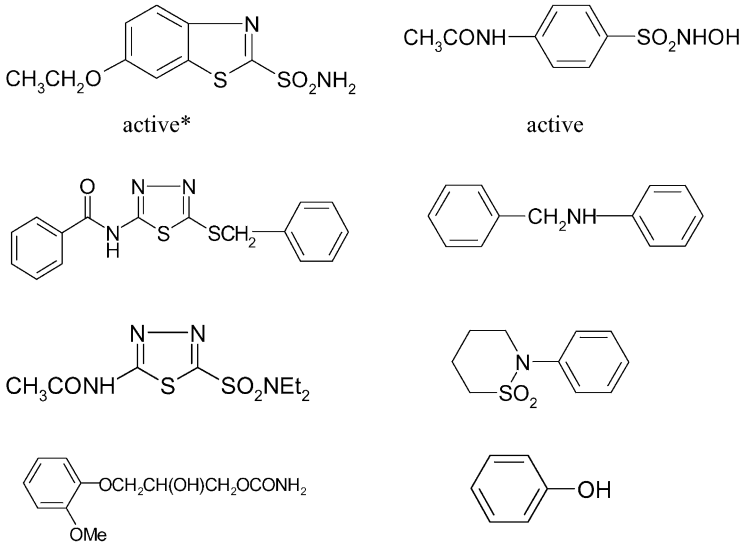

Fig. 2. Representative structures of CA II inhibitors (compounds marked with (∗) were classified as active in the analysis).

of structures; some representative structures are shown in Fig. 3. The MAO inhibitory activity was previously divided into four classes, 3, 2, 1, and 0, with class 3 being the most active and class 0 the least active. In our binary QSAR analysis, a binary threshold value of 2 was used to classify compounds as active and inactive; according to this threshold, compounds in classes 3 and 2 were labeled as active, compounds in classes 1 and 0 as inactive. The data set was randomly divided into a training set of 1008 compounds and a test set of 600 compounds.

2.2. Molecular diversity analysis

Relative molecular diversity was analyzed using calculated pairwise Tanimoto coefficient (T_c) of a data set. Tanimoto coefficients and average T_c values were calculated

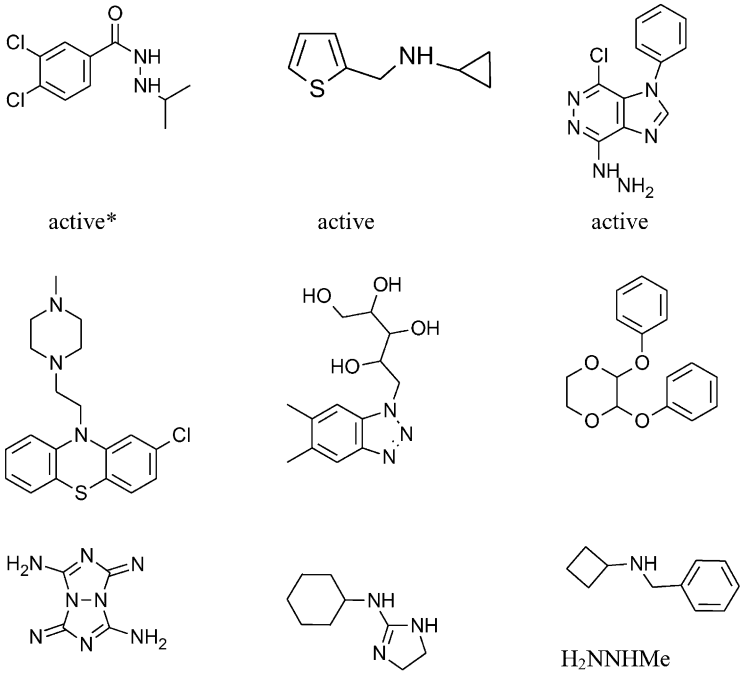

Fig. 3. Representative structures of MAO inhibitors (compounds marked with (∗) were classified as active in the analysis).

using the MACCS type of fingerprints implemented in MOE [18]. The MACCS key is a bit string where each bit position refers to the presence or absence of a unique structural pattern. The Tanimoto coefficient for comparison of two molecules was calculated as follows:

$$T_c = \frac{B_c}{B_1 + B_2 - B_c}$$

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

MDL drug data report (MDDR) [19] was used as a reference database in the diversity analysis. A total of 85,949 compounds with molecular weight less than 700 and containing no metal, Si, and B elements were extracted from MDDR. A set of 138 estradiol analogs were extracted from the ER ligand data set as a congeneric set of compounds.

2.3. Molecular descriptors

All molecular descriptors used in this analysis were coded and calculated using the 2000.02 version of MOE [20]. Out of 398 molecular descriptors that can be calculated from MOE, 150 descriptors were carefully and intuitively selected as a variable pool. These descriptors include physicochemical properties such as log P [21], Kier connectivity indexes, shape indexes, and E-state indexes [22,23], SS-key type descriptors [24], and other descriptors developed by Chemical Computing Group, Inc. [25]. In this study, 1D and 2D molecular descriptors were used, which have been shown to perform well in SAR analysis [19,24,26]. In addition, Kier's shape indices contain implicit 3D structural information. Explicit 3D molecular descriptors were not considered to avoid bias of the analysis due to predicted conformational effects and lengthy calculations.

2.4. Binary QSAR analysis

A detailed description of the binary QSAR methodology has been given in previous publications [4–6]. Briefly, for each molecule in a data set, a set of molecular descriptors is computed, which is then transformed into a set of decorrelated and normalized set of variables, and the probability distribution is estimated based on Bayes' Theorem. The binary QSAR analysis procedure is summarized in Fig. 4, models were cross-validated using a leave-one-out procedure [27]. Quality of a binary QSAR model was measured as follows: let m_0 represent the number of active compounds, m_1 the number of inactive compounds, c_0 the number of active compounds correctly labeled by the QSAR model, and c_1 the number of inactive compounds correctly predicted by the QSAR model. Three parameters of performance were

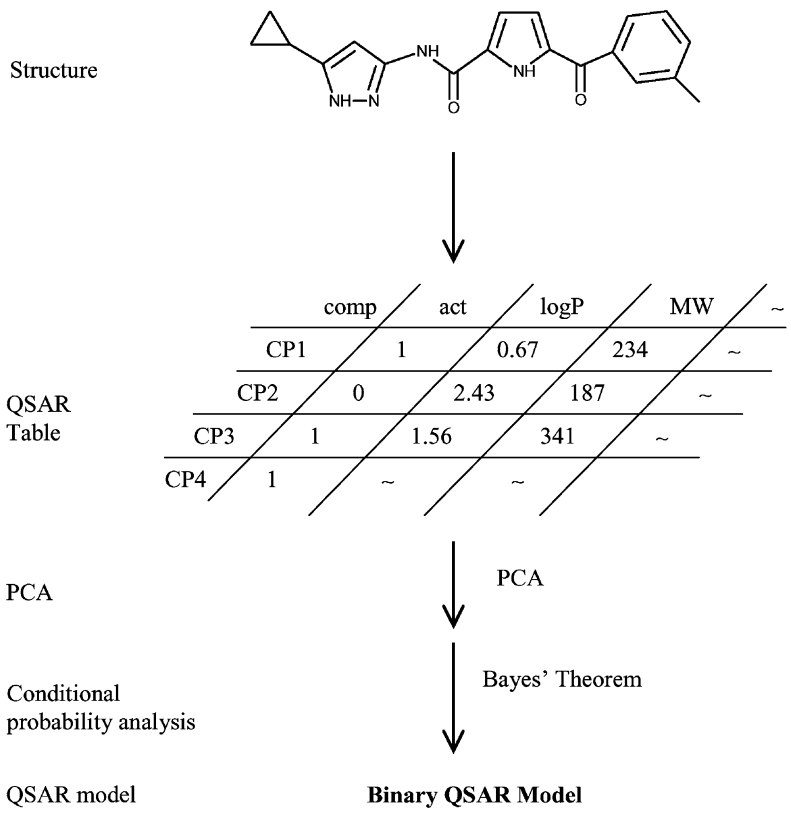

Fig. 4. Flow chart of binary QSAR analysis implemented in MOE.

calculated: (1) accuracy on active compounds, c_0/m_0; (2) accuracy on inactive compounds, c_1/m_1; (3) overall accuracy on all of the compounds, $(c_0 + c_1)/(m_0 + m_1)$.

2.5. Variable selection

GAs are a class of algorithms inspired by the process of natural evolution in which species having a high fitness under some conditions can prevail and survive to the next generation; the best species can be adapted by crossover and/or mutation in the search for better individuals. GAs have attracted much attention as a means to approach various optimization problems in many fields including QSAR analysis [28,29]. In this study, variable selection is achieved by using a GA developed in our group. In this method, a chromosome and its fitness in the species represent a set of molecular descriptors and the cross-validated predictive accuracy of the derived binary QSAR model, respectively. Binary QSAR analysis is used to evaluate solution fitness; the best variable combinations are selected using a cross-validated statistical score. Fig. 5 shows the GA-based variable selection scheme. This algorithm consists of five basic steps: (1) create initial population. 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 non-selection. In the process of creating the initial population, half of the chromosomes are randomly generated, then the second half is the complement of the first half. Using this scheme, every descriptor in the variable pool gets sampled. (2) Evaluate the fitness of the initial population using binary QSAR. The fitness of each chromosome is evaluated by the cross-validated predictive accuracy of the binary QSAR model. (3) Select a reproductive population. The chromosome with the highest fitness is chosen as the best chromosome. If two chromosomes have the same fitness, the one with fewer variables is chosen as the best chromosome. The best chromosome is protected and will survive to the next generation and only to be replaced with a chromosome gives better fitness or a chromosome gives equal fitness but with fewer variables. The chromosomes with highest fitness are selected from the population in an arbitrary proportion as a reproductive population. (4) Create a new population by uniform crossover of the reproductive population. It has been shown that uniform crossover is more effective than either one- or two-point crossover [30]. In uniform crossover, a random mask is generated, which consists of a random vector of 1's and 0's of the same length as the parent chromosomes. Uniform crossover is an operation that takes each bit in the parents and assigns the bit from one parent to one offspring and the bit from the other parent to the other offspring according to the random mask. A new population is generated by uniform crossover of pairs of the reproductive population. (5) Create next generation by combining and mutating the reproductive population and the new population. The reproductive population and new population are combined and mutated with a mutation rate of 5%. The best chromosome in the reproductive population is kept from the mutation process. The cycle is repeated until the number of generations reaches a given maximum. The final model obtained is further refined by removing descriptors which do not affect predictive accuracy significantly.

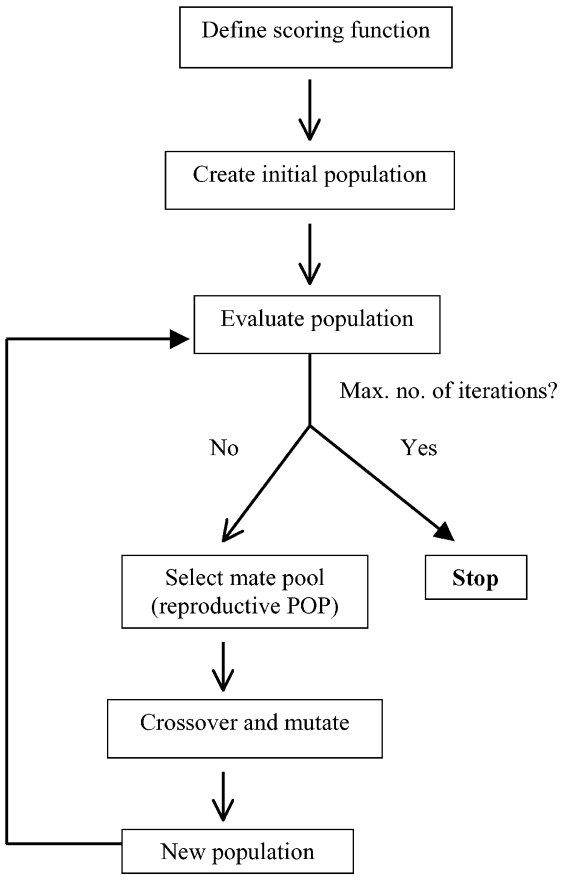

Fig. 5. Scheme of GA-based optimization process.

3. Results and discussion

3.1. Diversity analysis

The calculated average Tanimoto coefficients of different data sets are presented in Table 1. Compounds from MDDR has a calculated average T_c value of 0.39 ± 0.11 (average

Table 1
Average Tanimoto coefficients of compounds analyzed

Compounds	Number of compounds	Average T_c
ER ligands	400	0.40 ± 0.18
CA II inhibitors	280	0.54 ± 0.17
MAO inhibitors	1008	0.28 ± 0.13
MDDR	85949	0.39 ± 0.11
Estradiol analogs	138	0.78 ± 0.12

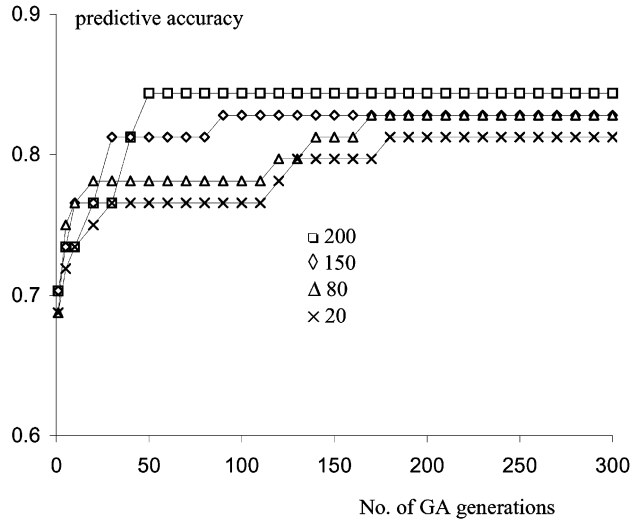

Fig. 6. Effect of initial population size on the cross-validated predictive accuracy of GA-based binary QSAR.

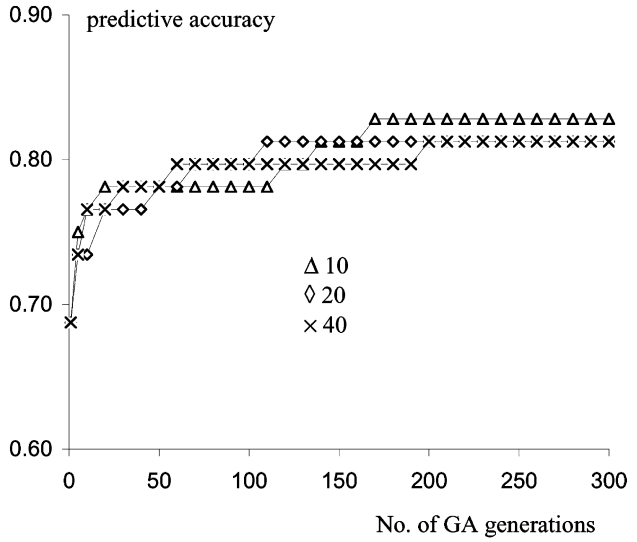

Fig. 7. Effect of reproductive population size on the outcome of GA-based binary QSAR.

$\pm$S.D.), while 138 estradiol analogs have a average T_c value of 0.78 ± 0.12. Compared to the calculated average T_c values of the two reference sets of compounds, the diversity of ER ligands (0.40 ± 0.18), CA II inhibitors (0.54 ± 0.17), and MAO inhibitors (0.28 ± 0.13) is high.

3.2. Selection of GA parameters

Two important parameters, the size of an initial population and the size of a reproductive population, can affect the outcome and the calculation speed of GA-based binary QSAR analysis. We investigated the effects of the sizes of initial population and reproductive population on the general outcome of the GA using the ER ligands. Fig. 6 illustrates the effect of four different initial population sizes on the binary QSAR analysis of ER ligands. The GA with an initial population size of 200 rapidly converges ($\sim$50 generations) and reached an optimal binary QSAR model in a reasonable number of GA generations. Deciding upon the size of the initial population is very difficult. There are some trade-offs between initial population size and the number of generations needed to converge or to reach an optimal QSAR model. Generally speaking, a large initial population provides the GA with a nice sampling of the search space. As a result of this study, an initial population of 200 was subsequently used.

Deciding how many chromosomes to keep as a reproductive population in each GA generation is somewhat arbitrary. On the one hand, letting only a few chromosomes survive to the next generation limits the available genes in the offspring. On the other hand, keeping too many chromosomes allows bad performers to contribute their characteristics to the next generation. Furthermore, the computational time of GA-based binary QSAR analysis is directly dependent on the size of the reproductive population. Fig. 7 shows the effects of different sizes of reproductive population on GA-based binary QSAR analysis of ER ligands. From this figure, we can see that there were no significant differences in the final predictivity of the binary QSAR models for the three reproductive population sizes. However, the computation time for a reproductive population size of 40 chromosomes is four times longer than the one with a reproductive population size of 10 chromosomes. Therefore, in this study, a reproductive population size of 10 chromosomes was used.

Mutation rate is another important factor in GA. It can introduce traits not in the parent population and can prevent the GA from converging prematurely. Typically, mutation rate is of the order of 1–5%. In this study, a mutation rate of 5% was arbitrarily chosen.

3.3. Analysis of ER ligands

Fig. 8 shows a plot of the cross-validated predictive accuracy of binary QSAR models versus the number of GA generations of a single run. With the GA-based variable selection method, a binary QSAR model with a predictive accuracy of 80% on actives, 93% on inactives, and an overall predictive accuracy of 91% was obtained within 50 GA generations. The cross-validated predictive accuracy was 78% on actives, 93% on inactives, and 90% for all the compounds. Twenty-eight molecular descriptors were used in this binary QSAR model. External validation of a QSAR model derived with GA-based variable selection is necessary [31]. The derived binary QSAR model was applied to the test set of compounds not included in the training set. Six out of seven active compounds (86%) were correctly predicted, and 51 out of 56 inactive compounds (91%) were correctly labeled by the binary QSAR model. The overall predictive accuracy for the test set compounds was 90%. The

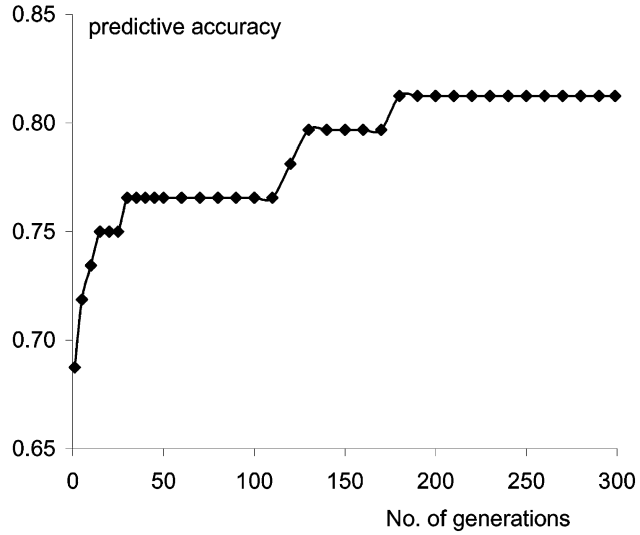

Fig. 8. Plot of predictive accuracy vs. number of GA generations of ER ligands.

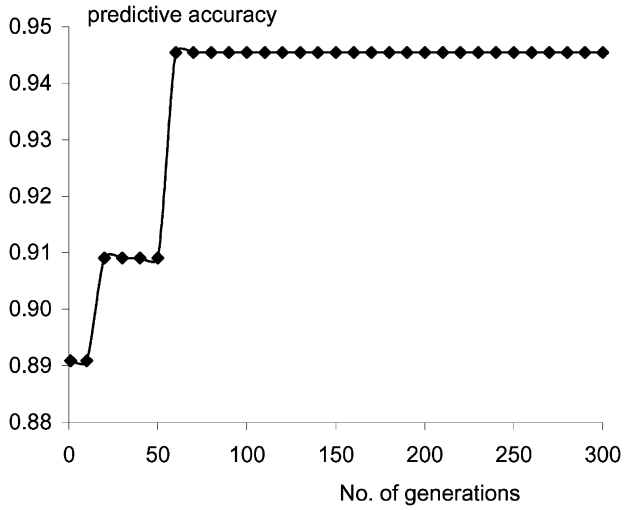

Fig. 9. Plot of predictive accuracy vs. number of GA generations of CA II inhibitors.

predictive results for the derived binary QSAR are summarized in Table 2. The predictive accuracy for non-cross-validated, cross-validated, and test set are very consistent indicating that the derived binary QSAR model is highly predictive and robust. The total number of combinations of different molecular descriptors (initial population + 50 GA generations) was 1200. If all the possible combinations of 150 molecular descriptors were investigated, the number of combinations to be explored would be $2^{150} - 1$, which is a astronomical number and impossible to enumerate fully. Therefore, the GA-based variable selection method greatly enhanced the applicability of binary QSAR analysis.

3.4. Analysis of CA II inhibitors

The results of the binary QSAR analysis of CA II inhibitors are summarized in Table 3 and Fig. 9. Similar to the analysis of ER ligands, within 60 GA generations, a binary QSAR model with a predictive accuracy of 94% on

Table 2
Predictive accuracy of the optimal binary QSAR model of ER ligands

Compound		No. of compounds		Predictive accuracy[a]	
		Observed	Predicted		
Training set	Active[b]	64	51	80 (78)	91 (90)
	Inactive[c]	336	311	93 (93)	
Test set	Active	7	6	86	90
	Inactive	56	51	91	

[a] Values in parenthesis are cross-validated accuracy.
[b] RBA $\geq$ 50%.
[c] RBA $<$ 50%.

Table 3
Predictive accuracy of the optimal binary QSAR model of CA II inhibitors

Compound		No. of compounds		Predictive accuracy[a]	
		Observed	Predicted		
Training set	Active[b]	218	206	94 (90)	95 (91)
	Inactive[c]	50	48	96 (94)	
Test set	Active	45	41	91	93
	Inactive	12	12	100	

[a] Values in parenthesis are cross-validated accuracy.
[b] $IC_{50} \leq 1\,\mu M$.
[c] $IC_{50} > 1\,\mu M$.

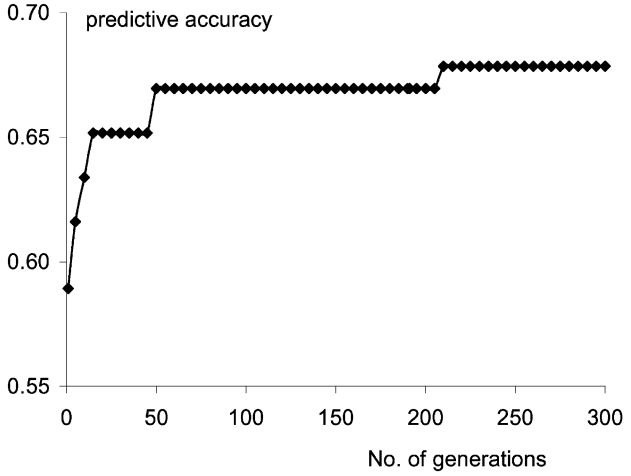

Fig. 10. Plot of predictive accuracy vs. number of GA generations of MAO inhibitors.

actives, 96% on inactives, and 95% for all the compounds, was derived. The cross-validated accuracy was 90% for actives, 94% for inactives, and 91% for all the compounds. Twenty-four molecular descriptors were used in this binary QSAR model. The derived binary QSAR model was tested with a set of 57 CA II inhibitors not included in the training set. Forty-one out of 45 active compounds (91%) were correctly predicted, and all 12 inactive compounds (100%) were correctly predicted. The overall predictive accuracy for the test set of compounds was 93% consistent with the cross-validation result. A total of 1400 combinations of molecular descriptors were explored. The result also indicates the usefulness of the GA-based variable selection method.

3.5. Analysis of MAO inhibitors

This is the most diverse set of compounds of the three data sets analyzed based on the calculated average Tanimoto coefficient (0.28 ± 0.13). Fig. 10 and Table 4 summarize the results of the binary QSAR analysis of MAO inhibitors. Because of the size and probably the diversity of this set of compounds, it took more GA generations than the previous two data sets to obtain a reasonable binary QSAR model (around 200 GA generations). The total number of sets of molecular descriptors explored for this set of compounds was 4200. The binary QSAR model has a predictive accuracy of 75% for actives, 86% for inactives, and 85% for all the compounds. The predictive accuracy for the test set of compounds was 64% for actives, and 86% for inactives. A further investigation had found that the binary QSAR model had a much higher predictive accuracy for the most active compounds (class 3) (predictive accuracy of 91%) and least active compounds (class 0) (predictive accuracy of 86%) than compounds with activity around the binary threshold value (predictive accuracy of 61% for class 2 compounds and 78% for class 1 compounds). Similar results were observed for the test set of compounds (see Table 4). The predictive accuracy was 74% for class 3, 52% for class 2, 72% for class 1, and 87% for class 0 compounds in the test set. This kind of boundary effect was also observed in previous studies [5,6]. Because the binary threshold value was arbitrary, compounds having activity value near the binary threshold value could be placed in either active or inactive fields which could be due to experimental errors. A second possible explanation for the boundary effect is that the structural difference between most active compounds (class 3) and least active compounds (class 0) is more distinct than that between compounds having activity near the binary threshold value (classes 1 and 2).

3.6. Comparison with binary QSAR models without using GA-based variable selection

Two binary QSAR models have been published previously by one of the authors [5,6]. In previous binary QSAR analyses of CA II inhibitors [5] and ER ligands [6], 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

Table 4
Predictive accuracy of the optimal binary QSAR model of MAO inhibitors

Compound	MAO activity	Binary classification	No. of compounds		Predictive accuracy[a]		
			Observed	Predicted			
Training set	3	Active	53	48	91	75 (65)	85 (83)
	2	Active	59	36	61		
	1	Inactive	67	52	78	86 (85)	
	0	Inactive	829	715	86		
Test set	3	Active	34	25	74	64	84
	2	Active	27	14	52		
	1	Inactive	47	34	72	86	
	0	Inactive	492	430	87		

[a] Values in parenthesis are cross-validated accuracy.

Table 5
Comparison of binary QSAR models with and without GA-based variable selection

Binary QSAR model			Compounds		
			ER ligands	CA II inhibitors	MAO inhibitors
Predictive accuracy	Actives	With GA	80 (77)[a]	95 (90)	75 (65)
		Without GA	71 (66)	97 (91)	61 (54)
	Inactives	With GA	93 (93)	96 (94)	86 (85)
		Without GA	93 (92)	89 (76)	90 (89)
	Overall	With GA	91 (90)	95 (91)	85 (83)
		Without GA	90 (88)	96 (89)	87 (85)
Number of descriptors used		With GA	28 (21)[b]	24 (23)	38 (36)
		Without GA	150 (26)	150 (35)	150 (42)

[a] Value in parenthesis is cross-validated predictive accuracy.
[b] Value in parenthesis is the number of PCAs used in the model.

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. In the present study, identifying indicator variables was not attempted.

Binary QSAR models derived without GA-based variable selection (using all 150 molecular descriptors) are summarized in Table 5. The binary QSAR models derived using GA-based variable selection method used far fewer molecular descriptors, which is very important for using the derived model in virtual screening. The speed of virtual screening is dependent on the number of molecular descriptors used in the binary QSAR model. These models also have higher predictive accuracy, especially cross-validated predictive accuracy. This suggests that these models are more robust and predictive than those models derived using the full set of 150 descriptors.

4. Conclusion

In this investigation, we have developed a GA for variable selection in binary QSAR analysis. The algorithm has been applied in the binary QSAR analyses of three diverse sets of compounds, ER ligands, carbonic anhydrase II inhibitors, and MAO inhibitors. The results indicate that the GA-based variable selection method increased the performance of the binary QSAR method.

Acknowledgements

The authors would like to thank Dr. Yvonne Martin of Abbott for providing the MAO inhibitor set.

References

[1] C. Hansch, R.M. Muir, T. Fujita, P.P. Maloney, E. Geiger, M. Streich, The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients, J. Am. Chem. Soc. 85 (1963) 2817–2824.
[2] T. Fujita, J. Iwasa, C. Hansch, A new substituent constant, π, derived from partition coefficients, J. Am. Chem. Soc. 86 (1964) 5175–5180.
[3] C. Hansch, D. Hoekman, H. Gao, Comparative QSAR: toward a deeper understanding of chemicobiological interactions, Chem. Rev. 96 (1996) 1045–1075.
[4] P. Labute, Binary QSAR: a new method for the determination of quantitative structure–activity relationships, in: R.B. Altman, A.K. Dunker, L. Hunter, T.E. Klein, K. Lauderdale (Eds.), Proceedings of the Pacific Symposium on Biocomputing'99 World Scientific, New Jersey, pp. 444–455.
[5] H. Gao, J. Bajorath, Comparison of binary and 2D QSAR analyses using inhibitors of human carbonic anhydrase II as a test case, Mol. Diver. 4 (1999) 115–130.
[6] H. Gao, C. Williams, P. Labute, J. Bajorath, Binary-QSAR analysis of estrogen receptor ligands, J. Chem. Inf. Comput. Sci. 39 (1999) 164–168.
[7] J.M. Sutter, S.L. Dixon, P.C. Jurs, Automated descriptor selection for quantitative-structure–activity relationships using generalized simulated annealing, J. Chem. Inf. Comput. Sci. 35 (1995) 77–84.
[8] J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, Microchem. J. 47 (1993) 60–66.
[9] S.S. So, M. Karplus, Evolutionary optimization in quantitative structure–activity relationships: an application of genetic neural networks, J. Med. Chem. 12 (1996) 9–20.
[10] D. Rogers, A.J. Hopfinger, Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships, J. Chem. Inf. Comput. Sci. 34 (1994) 854–866.
[11] T.J. Hou, J.M. Wang, N. Liao, X.J. Xu, Applications of genetic algorithms on the structure–activity relationship analysis of some cinnamamides, J. Chem. Inf. Comput. Sci. 39 (1999) 775–781.
[12] K. Hasegawa, T. Kimura, K. Funatsu, GA strategy for variable selection in QSAR studies: enhancement of comparative molecular binding energy analysis by GA-based PLS method, Quant. Struct. Act. Relat. 18 (1999) 262–272.
[13] K. Hasegawa, Y. Miyashita, K. Funatsu, GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists, J. Chem. Inf. Comput. Sci. 37 (1997) 306–310.

[14] H. Kubinyi, Variable selection in QSAR studies. Part I. An evolutionary algorithm, Quant. Struct. Act. Relat. 13 (1994) 285–294.
[15] H. Kubinyi, Variable selection in QSAR studies. Part II. Highly efficient combination of systematic search and evolution, Quant. Struct. Act. Relat. 13 (1994) 393–401.
[16] B.T. Luke, Evolutionary programming applied to the development of quantitative structure–activity relationships and quantitative structure–property relationships, J. Chem. Inf. Comput. Sci. 34 (1994) 1279–1287.
[17] H. Gao, J.A. Katzenellenbogen, R. Garg, C. Hansch, Comparative QSAR analysis of estrogen receptor ligands, Chem. Rev. 99 (1999) 723–744.
[18] R.D. Brown, Y.C. Martin, Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.
[19] MDL Drug Data Report 99.2, MDL Information Systems, Inc., 1999.
[20] Chemical Computing Group Inc. MOE 1998.03, 1255 University Street, Montreal, Que., Canada, H3B 3X3.
[21] C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, ACS Professional Reference Book, American Chemical Society, Washington, DC, 1995.
[22] L.B. Kier, L.H. Hall, The nature of structure–activity relationships and their relation to molecular connectivity, Eur. J. Med. Chem. 12 (1997) 307–312.
[23] L.B. Kier, Indexes of molecular shape from chemical graphs, Med. Res. Rev. 7 (1987) 417–440.
[24] L. Xue, J. Godden, H. Gao, J. Bajorath, Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis, J. Chem. Inf. Comput. Sci. 39 (1999) 699–704.
[25] A. Lin, QuaSAR-descriptors, J. Chem. Comput. Group. http://www.chemcomp.com, 8 February 20001.
[26] W. Ajay, P. Walters, M.A. Murcko, Can we learn to distinguish between ‘drug-like’ and ‘nondrug-like’ molecules? J. Med. Chem. 41 (1998) 3314–3324.
[27] R.D. Cramer, J.D. Bunce, D.E. Patterson, I.E. Frank, Cross-validation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies, Quant. Struct. Act. Relat. 7 (1988) 18–25.
[28] R.L. Haupt, S.E. Haupt (Eds.), Practical Genetic Algorithms, Wiley, New York, 1998.
[29] G. Syswerda, Genetic algorithms and their applications, in: L. Davis (Ed.), Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, 1991, pp. 332–349.
[30] G. Syswerda, Uniform crossover in genetic algorithms, in: J.D. Schaffer (Ed.), Proceedings of the Third International Conference on Genetic Algorithms, Los Altos, Morgan Kaufmann, CA, 1989, pp. 2–9.
[31] R. Leardi, Application of genetic algorithms to feature selection under full validation conditions and to outlier detection, J. Chemom. 8 (1994) 65–79.