Optimization of the Inner Relation Function of QPLS Using Genetic Algorithm

Hiroshi Yoshida and Kimito Funatsu*

Department of Knowledge-Based Information Engineering, Toyohashi University of Technology, Tempaku, Toyohashi, 441, Japan

Received April 21, 1997[⊗]

The quadratic partial least squares (QPLS) method is a nonlinear version of the PLS method. In the method a quadratic inner relation function is used instead of a linear function in the PLS method. When the QPLS method is applied, the inner relation function should be optimized. Hitherto, a mathematical optimization method is employed, however, which is not always useful. In this paper we proposed a new hybrid method that integrates the genetic algorithm (GA) and the QPLS method (GA-QPLS), in which GA is introduced to optimize the inner relation function instead of the mathematical method. The hybrid method is applied to chemical data that show a nonlinearity, and then its performance is investigated and compared with that of the conventional QPLS method. It is shown that the hybrid method can overcome shortcomings of the mathematical method and lead a significant improvement over the conventional QPLS method.

1. INTRODUCTION

Discovering and establishing relationships between the chemical structures of molecules and their activities or properties are always interesting research topics for chemists. The quantitative description of the relations is the so-called quantitative structure−activity/property relationships (QSAR/QSPR). The main purpose of the study is to obtain meaningful models, which can be applied to the prediction of activities or properties of new compounds. Hitherto, the Hansch approach using classical multiple linear regression (MLR) has been used in QSAR/QSPR studies.[1] The method expresses a biological activity or property as a linear combination of independent variables, yielding an explicit form of the independent variables. However, it has no ability to give a robust model in the cases where the variables (the structural descriptors that describe the features of the chemical structures) are correlated, or the number of samples is smaller than that of the descriptors.

Wold et al. proposed a versatile multivariate method of partial least squares (PLS), which now plays an important role in the QSAR/QSPR studies and analytical chemistry.[2−6] A predictive model can be obtained using the PLS method even though the descriptors are correlated or the number of the descriptors is greater than that of the samples. More recently the PLS method is successfully applied to a three-dimensional QSAR approach of comparative molecular field analysis (CoMFA).[7] Although the PLS method is useful, its major restriction is that only linear relation can be extracted from data; therefore, it is not suitable to tackle nonlinear problems.

Recently the opportunity of dealing with nonlinear data is increased following the extension of the QSAR applications. Basically there two ways to model nonlinear data: (1) expanding independent variables to include nonlinear terms such as quadratic and cross-product terms, then the model is constructed by MLR or PLS [The shortcoming of this method is that the model includes many nonlinear terms, which will dominate the data, often resulting in bad predictions. Quite recently Berglund et al. proposed a new method to overcome the shortcoming, in which the model obtained includes implicitly the nonlinear terms.[8]] and (2) using nonlinear regression methods such as the quadratic PLS method and Neural Networks (NN). In the QPLS method, a quadratic inner relation function is used instead of a linear relation function in the PLS method.[9−11] NN is also a nonlinear method, inspired by the current understanding of neurophysiology that tries to mimic the human brain. Nowadays NN is intensively used in the QSAR/QSAR studies.[12−17]

By the way, when the latter, that is to say, nonlinear methods are employed, the model equation should be optimized. In the case of NN the networks should be learned, which is also called optimized. Hitherto, the backpropagation (BP) algorithm is widely used as a training algorithm, which is criticized for its slow convergence speed. Moreover, it is prone to get trapped in local optima. Much effort has been spent to remedy the shortcomings of the BP algorithm from the theoretical and practical points of view.[14−17] On the other hand, in the case of the QPLS method, the quadratic inner relation function is used, which should be optimized. The optimization is performed based on a linearization of the function, so that the optimization is not always carried out properly. Besides, the convergence would be influenced with initial guesses. Hitherto, very little attention has been paid to its disadvantage, although only its advantage has been discussed. In this way, it is quite important to validate whether models obtained are the optimal ones. And at the same time the quality of the model plays a crucial role in the QSAR/QSPR studies.

Recently a genetic algorithm (GA) has been an object of attention as an optimization method.[18−22] GA is a heuristic optimization method that is inspired of evolutionary principals in nature. In GA a number of solutions to a particular problem are represented as chromosomes, and then the chromosomes are evolved based on the concepts such as crossover or mutation. The best evolved chromosome would have a very high probability to the solution. GA has been successfully applied to many optimization problems such as the optimization of NN instead of the BP algorithm.[18] In

* Corresponding author.
[⊗] Abstract published in *Advance ACS Abstracts,* October 15, 1997.

S0095-2338(97)00026-7 CCC: $14.00

the QSAR/QSPR studies, hybrid approaches that integrates GA and modeling methods have begun to be used.[20−22] Leadri et al. have integrated GA and PLS for feature selection, in which a more predictive model is obtained than that of the classical method.[20,21]

In this paper a new hybrid method that integrates GA and QPLS (GA-QPLS) is proposed, in which GA is applied to optimize the quadratic inner relation function. The hybrid method is applied to the modeling of autoignition temperature (AIT) that shows a nonlinearity. Then its performance is investigated and compared with that of the conventional QPLS method. The hybrid method can overcome the shortcomings of the conventional method, and the potential of the hybrid method is demonstrated by the development of a improved QSPR model.

2. MATERIALS AND METHODS

2.1. Materials. The GA-QPLS method proposed in this work is applied to chemical data to investigate its performance and to compare with that of the conventional QPLS method. It is necessary that the method be applied to nonlinear data to do that. Tetteh et al. have investigated autoignition temperature (AIT) data of 85 organic compounds by means of NN.[16] Therefore it would be thought that the AIT data show a nonlinearity and are suitable to be investigated by means of nonlinear methods. They use six structural descriptors, which are physicochemical parameter P_c, geometric parameter P_A, topological index ${}^0X$, and electric parameters Q_T, I_{ald} and I_{ket}, respectively. Therefore a six dimensional vector is described as follows.

$$\mathbf{x} = (P_c, P_A, {}^0X, Q_T, I_{ald}, I_{ket})$$

Details of the parameters are given in Table 1, where I_{ald} and I_{ket} indicate aldehyde and ketone functional groups, respectively. For compounds with the functional groups, values of one are assigned and vice versa.

2.2. Methods. 2.2.1. Partial Least Squares (PLS). The PLS method is a well-known multivariate method, which can give structural requirements for dependent variables. Therefore, the method is often used especially in the QSAR studies up to now. A brief introduction of the method will be given below, for a detailed explanation of the method, see ref 3. In the case of the PLS method, independent variables **X** and dependent variables **y** can be decomposed as follows.

$$\mathbf{X} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}'_a + \mathbf{E} \tag{1}$$

$$\mathbf{y} = \sum_{a=1}^{A} \mathbf{u}_a q_a + \mathbf{f} \tag{2}$$

$$\mathbf{u} = b\mathbf{t} + \mathbf{h} \tag{3}$$

$$\mathbf{t} = \mathbf{X}\mathbf{w} \tag{4}$$

where $\mathbf{t}_a$ and $\mathbf{u}_a$ are the ath latent variables and $\mathbf{p}_a$ and q_a are the ath loadings of **X** and **y**, respectively. **E** and **f** are residuals. A is the number of the PLS components. The PLS method uses a link function which relates latent variables to dependent variables. Its coefficient and residual are b and **h**, respectively.

Table 1. Parameters Used for the Modeling

P_c	critical pressure [1.013*10^5 Pa]
P_A	parachor at 20 °C [(cm^3 mol^{-1})(dyn cm$-$1)$^{1/4}$]
${}^0X$	zeroth order molecular connectivity index
Q_T	sum of negative atomic charges [au]
I_{ald}	descriptor for aldehydes
I_{ket}	descriptor for ketones

2.2.2. Quadratic PLS (QPLS). The QPLS method is an extension of the linear PLS method to deal with nonlinear problems.[9] In the case of the PLS method, the link function $g(\mathbf{t})$ is a linear function as mentioned above; on the other hand, in the case of the QPLS method the link function is a quadratic function as shown in eqs 5 and 6.

$$\mathbf{u} = g(\mathbf{t}) + \mathbf{h} \tag{5}$$

$$g(\mathbf{t}) = c_0 + c_1\mathbf{t} + c_2\mathbf{t}^2 \tag{6}$$

where c_0, c_1, and c_2 are coefficients of the link function, respectively. In the original algorithm, an initial weight vector **w** is determined by means of the linear PLS method, then the coefficients c_0, c_1, and c_2 are determined by the least squares method. In the optimization process of the QPLS model, **w** is updated by a linearization of the quadratic inner relation function. The convergence of the process is checked by means of **w**.

$$\frac{||\mathbf{w}_{new} - \mathbf{w}_{old}||}{||\mathbf{w}_{old}||} \leq 10^{-10} \tag{7}$$

The optimization process is repeated until this criterion is fulfilled or the number of iterations exceeds 50 or 100. However the number of the iterations is not always enough, being clearly problem-dependent. Therefore it can be thought that there are two crucial problems in the conventional QPLS algorithm as follows.

(1) The initial guess derived from the linear PLS method is not always suitable.

(2) The optimization of the quadratic inner relation function is not always carried out due to the linearization of the function, causing local optimum problems.

If the QPLS method is employed, the optimization of the QPLS model is a crucial problem, where none of the other optimization methods have been applied effectively. Therefore it is desired that a powerful optimization method be applied to the QPLS method.

2.2.3. Genetic Algorithm (GA). GA is one of the optimization methods, which is is so called after its similarity of the evolutionary process in nature. The fundamental difference between the process of evolution and GA is whether a nucleotide sequence of DNA is used or a bit string is used to describe a chromosome. Basically there are five basic steps in GA: 1, coding; 2, evaluation; 3, reproduction; 4, crossover; 5, mutation. Parameters to be optimized are coded into a bit string (chromosome). Many chromosomes are created in a population, so that many possible solutions can be taken into consideration at the same time. The chromosomes are then evolved by the process of GA. The major strength of GA is namely its ability to explore a huge multidimensional space and give a better chance of finding the global optimum even in the presence of local optima. In

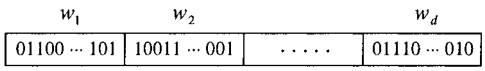

Figure 1. Representation of a chromosome for a *d*-dimensional vector **w**.

the following section, a brief overview of each step of GA and strategies used in this work are explained.

2.2.4. Integration of GA and QPLS. In paragraph 2.2.2 the shortcomings of the conventional QPLS method are pointed out. When the QPLS method is employed, the inner relation function should be optimized. In this work GA is introduced to optimize the function, in other words two parameters (**c** and **t**) that govern the function are coded as a chromosome, and then the chromosome is evolved based on the idea of GA. When GA is applied to optimize the parameters, there are two possible ways as follows.

(1) The latent variable **t** and the coefficients of the inner relation function **c** are determined at the same time. It means that **t** and **c** are coded as a chromosome, and then the chromosome is evolved.

(2) The only latent variable **t** is coded as a chromosome, and then the chromosome is evolved. Finally **c** is determined by the least squares method.

When the parameters to be optimized are coded, the length of the chromosome of the case 1 is longer than that of the case 2. It is considered that optimization of a longer chromosome would not be easy, a shorter chromosome more desirable. In this work, therefore the latter is used for the optimization. The dimensionality of **t** is *n*, which is equal to the number of samples. When the number of the samples increases, the length of the corresponding chromosome becomes also longer, which is not desirable from the practical point of view. However, **t** is a linear combination of **X** and **w**, in which the dimensionality of **w** is *d*. The dimensionality is usually not so large; therefore, it is a better way to optimize **w** instead of **t**.

2.2.4.1. Coding. In this work, the weight vector **w** is coded as a chromosome. The chromosome has a length given by the number of parameters in the problem multiplied by the number of bits needed to code each parameter. The chromosome is composed of *d* genes, which is equal to the number of elements of **w**, and each element is composed of a bit string of length ten.[23] Therefore the *d*-dimensional vector **w** can be coded as a bit string of length *d**10 as shown in Figure 1. Then the initial population of a given size *N* is created, in which structures of each chromosome are created in a totally random way. As mentioned before, in the conventional QPLS algorithm, only one initial guess **w** is determined by means of the PLS method. Therefore the initial guess is not always suitable, which would influence the convergence of the optimization. In the case of the GA-QPLS method, however, many initial guesses can be taken into consideration at the same time; therefore, there is the advantage that the method would not be influenced with the initial guess.

2.2.4.2. Evaluation of Fitness. Fitness of each chromosome is evaluated by a fitness function. The scheme of the evaluation is shown in Figure 2. (1) A chromosome is decoded into the decimal to calculate the weight vector **w**. (2) The latent variable vector **t** is calculated using **X** and **w**. (3) The coefficients of the inner relation function **c** are determined by means of the least squares method, then the inner function is constructed, and finally $\hat{y}$ is estimated. In

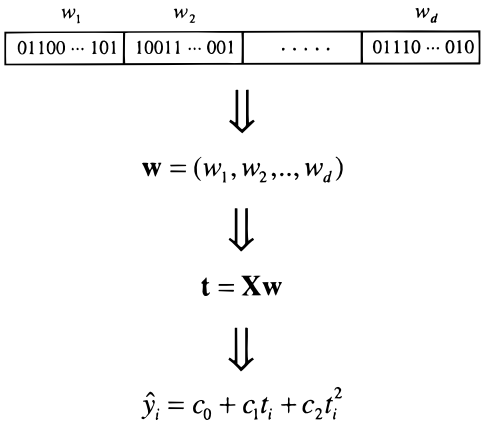

Figure 2. Evaluation of a chromosome.

this way, *an abstract chromosome* being a bit string is converted into *a real number* $\hat{y}$. If $\hat{y}$ is close to *y*, it can be said that the corresponding chromosome is well evaluated. Therefore, the reciprocal of *residual squared error* (*RSE*) is introduced as the fitness function. The definition of *RSE* is given as follows.

$$RSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8}$$

If the chromosome has a higher fitness value, it has also a higher possibility to survive and spread its feature to the next generation.

2.2.4.3. Reproduction. When the fitness values of each chromosome are calculated, the only chromosomes that fitness values are better than the average one can survive. The other chromosomes are eliminated from the population. A new population of *N* chromosomes is created as the next generation. If the chromosome has a higher fitness value, it has also a higher possibility to survive and spread its feature as mentioned above. It means that chromosomes with high fitness values are created more often.

2.2.4.4. Crossover. A pair of chromosomes is selected at random, and then an exchange of genes takes place to create a pair of new chromosomes. It is expected that the new chromosomes (offsprings) would have better fitness values than those of their parents, by which a wide variety of candidate solutions can be explored. After having produced the offsprings, they will take the place of their parents. There are some possible ways to produce offsprings; uniform crossover is employed in this study.

2.2.4.5. Mutation. All the possible solutions that can be exploed are dependent on the initial population; therefore a deadlock situation would occur unless some random changes would take place. In general a certain chromosome is selected at random, then some bits are also selected at random. If the bits are 0s they are changed to 1s and vice versa. Then the fitness values of each chromosome in the next generation are again evaluated, and a series of the above-mentioned operations is repeated until the number of the generations reaches the desired one or the target fitness value is obtained.

The programs used in this study are developed in our laboratory and run on a Gateway-2000. These programs are written in C language, and the QPLS algorithm is modified according to the Wold's algorithm with a minor correction.[5]

Table 2. Results of PLS

	no. of components					
	1	2	3	4	5	6
R^2	0.355	0.485	0.558	0.658	0.860	0.917
sd	81.6	72.9	67.5	59.4	38.1	29.3
Q^2	0.257	0.444	0.492	0.620	0.787	0.902

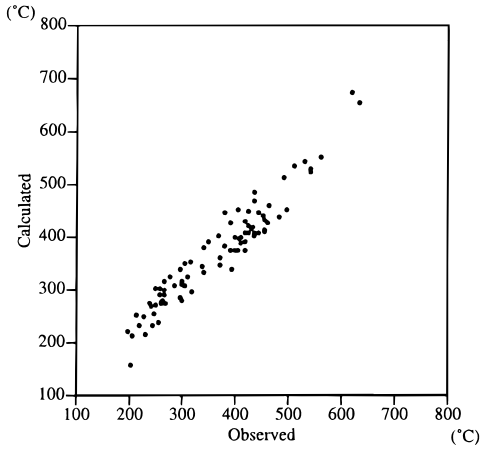

Figure 3. Observed-calculated plot (PLS).

3. RESULTS

In this study 85 AIT data are used for the modeling. The purpose of this study is to show the validity of the GA-QPLS method. Therefore it is necessary that the method be applied to nonlinear data. At first the PLS method was employed to investigate the nonlinearity. Prior to the PLS analysis, PCA was performed and eigenvalues obtained were investigated. As a result, it could be thought that the NIPALS algorithm was suitable for the modeling. The results are given in Table 2, where R^2 is a squared conventional correlation coefficient, Q^2 is a squared predictive coefficient and *sd* is a standard deviation, respectively. Q^2 is calculated by cross-validation using the leave-one-out procedure.[24] In this case, it is difficult to decide the optimal number of the PLS components, because the model precision continues to increase in proportion to the number of the components, the optimal number is determined to be six. Theoretically the PLS model converges toward the MLR model when the PLS component tends toward the dimensionality of **X**. In this case, the model obtained is equal to the MLR model, causing the critical problems of MLR as mentioned above. Therefore the model should be validated statistically, i.e., the bootstrap method.[25] However the validation is omitted because the purpose of the PLS analysis is only to investigate the nonlinearity of the data.

The plot of the calculated against observed AIT values is shown in Figure 3. It cannot be said that the AIT data show a nonlinearity because good agreement is obtained between the calculated and observed values. Therefore the nature of the data is investigated in more detail. The PLS plot of the *y*-score(*u*) against *x*-score(*t*) for the first component is shown in Figure 4. From the plot at least the data are not correlated; some complicated factors would be included. Therefore it can be thought that the AIT data does not show the linearity and are suitable to be used for the evaluation of the potential of the QPLS and GA-QPLS method.

Next the QPLS method was employed for the modeling. The AIT data show the nonlinearity, so that it is expected

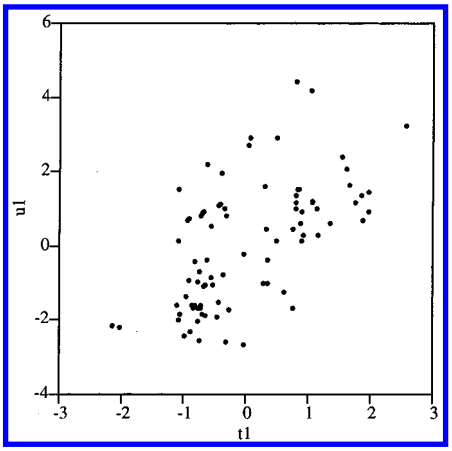

Figure 4. U1-t1 plot (PLS).

Table 3. Results of QPLS

	no. of components					
	1	2	3	4	5	6
R^2	0.038	0.247	0.323	0.776	0.844	0.844
sd	99.6	88.1	83.6	48.1	40.2	40.2
Q^2	−0.122	0.027	0.278	0.314	0.661	0.844

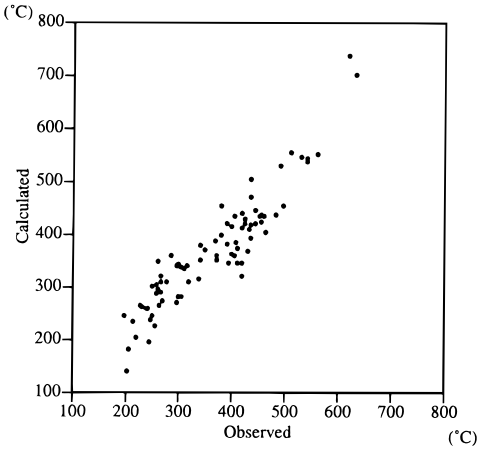

Figure 5. Observed-calculated plot (QPLS).

that the nonlinear approach will be successfully performed. In the original QPLS algorithm, the number of the iterations is set to be 50 or 100, which would be problem-dependent as mentioned above. In our program, the maximum number of the iterations is set to be 1000; however, when the algorithm is employed, the calculation could not reach the convergence even after the iterations. The results are given in Table 3.

In this case, the model precision also increases gradually in proportion to the number of the QPLS components. When the number of the components is five, though R^2 does not improve anymore, Q^2 still improves, the optimal number is determined to be also six. The plot of the calculated against observed AIT values is shown in Figure 5. It is interesting to compare the results of the PLS method with those of the QPLS method. In the case of the PLS analysis, the model precisions of R^2 and Q^2 are 0.917 and 0.902, respectively. On the other hand, in the QPLS analysis R^2 and Q^2 are 0.844 and 0.844. Though the QPLS method can take a quadratic nonlinear relation into consideration for the modeling, the results of the method are worse than those of the PLS method. Especially as for the predictivity, Q^2 is minus when the number of the components is one. This means that the

Table 4. Parameters Used for the Optimization

no. of chromosomes	90
probability of crossover	50%
probability of mutation	2%

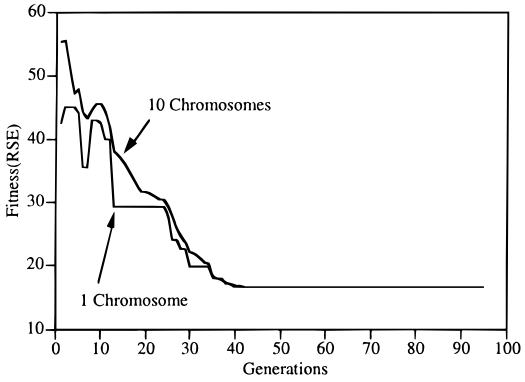

Figure 6. Optimization process of the inner relation function.

QPLS model is not predictive at all. It can be thought that the QPLS algorithm is subject to the initial guess that is initially determined by the PLS method, and the QPLS model is not optimized properly, being easy to get trapped in local optima. The conventional QPLS method could not model the nonlinear data successfully and its shortcomings are shown strikingly. Hitherto, little attention has been paid to those shortcomings, in addition there are not any promising mathematical optimization methods to overcome the shortcomings. The results also show that the optimization of the model itself is an important task when nonlinear methods are employed.

Finally the GA-QPLS method was employed to overcome the shortcomings, the optimization of the quadratic inner relation function. When GA is applied to the optimization, three parameters (number of chromosomes, probability of crossover and mutation) have to be defined in advance. Hitherto, no study of optimization of the parameters has been performed, so that there is not a fixed rule but a rule of thumb to obtain the optimal parameters. In general the probability of crossover is relatively high, i.e., about 50%, while the probability of mutation is relatively low i.e., about 2 or 3%. In this work, several runs are performed to obtain the trends of the parameters for the optimization. The parameter used are give in Table 4.

The plot of how the inner relation function is optimized is given in Figure 6. In the plot, the *RSE* value of the best chromosome (individual) and the mean *RSE* value of the best ten individuals are shown. In the case of the best individual, at the sixth generation a new individual being well evaluated is created. In this study, the individuals of generation $k + 1$ completely take the place of the individuals of generation k, so that no overlapping of individuals between generations is allowed, good individuals would be lost. The plot is a typical case that the best individual at the sixth generation is lost. No new individual being better than the previous generation is created even if crossover and mutation are introduced between the 13th and 34th generations. It can be seen that the value alternates sensitively due to crossover and mutation. Next the *RSE* value of the mean of the best ten individuals is investigated. The value is gradually decreasing, i.e., the fitness value is increasing. At the 95th generation, it is considered that the whole individuals does not improve any more and the optimization is carried out.

Table 5. Results of GA-QPLS

	no. of components					
	1	2	3	4	5	6
R^2	0.803	0.929	0.938	0.943	0.946	0.946
sd	40.1	27.1	25.2	24.2	23.7	23.7
Q^2	0.790	0.902	0.897	0.901	0.903	0.903

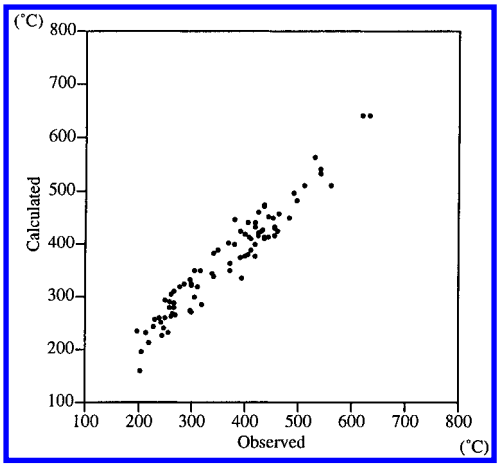

Figure 7. Observed-calculated plot (GA-QPLS).

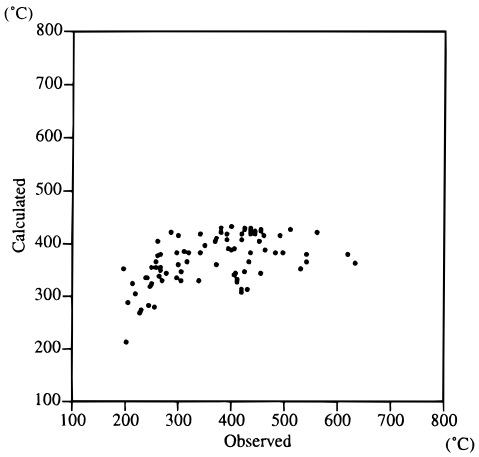

Figure 8. Observed-calculated plot (QPLS).

Then it is considered that the best individual at the fianl generation is the optimal weight vector **w**.

The procedure is also repeated for the other five components, and then the GA-QPLS model is constructed. The results of the modeling are given in Table 5. It is obvious that the results are improved dramatically in comparison with those of the conventional QPLS method. The optimal number is determined to be two; in addition the model precision is better than that of the conventional QPLS model. The plot of the calculated against observed values is shown in Figure 7. The plot of the QPLS model with two components is given in Figure 8, too. It is also clear from the plots that good agreement is obtained by introducing GA. In this work, the GA-QPLS method does not allow overlapping of chromosomes between generations so that good chromosomes would be lost. Recently a modified GA is proposed, in which good chromosomes are maintained between generations. If the idea is taken into consideration in the GA-QPLS method, the optimization is carried out more effectively.

In the GA-QPLS method, the idea of GA is introduced to optimize the inner relation function. Many initial guesses

Table 6. Coefficients of the Inner Relation Function (QPLS)

compd	c_0	c_1	c_2
1	−0.087	0.114	0.082
2	−0.218	−0.387	0.130
3	0.104	−0.341	−0.129
4	0.097	−3.627	−3.232
5	0.037	−39.201	−1223.159
6	0.001	0.132	−0.018

Table 7. Coefficients of the Inner Relation Function (GA-QPLS)

compd	c_0	c_1	c_2
1	0.433	4.657	−1.893
2	−0.071	−0.171	0.356
3	0.056	0.012	−0.093
4	−0.037	0.008	0.114
5	−0.022	0.016	0.017
6	−0.001	−0.646	5.809

are taken into consideration so that the method can explore many possible solutions simultaneously, each of which explores different regions in parameter space. Moreover candidate solutions are mated to explore more wide different regions. As a result, it is shown that the hybrid method can lead a significant improvement over the conventional QPLS method.

4. DISCUSSION

It is effective to employ nonlinear methods to deal with nonlinear data. However, the optimization of the model obtained is required. In the case of the conventional QPLS method, the quadratic inner relation function is optimized by the linearization of the function, and then the shortcomings are shown strikingly. Therefore, it is interesting to investigate how the inner relation function is optimized. In this case the coefficients of the function are investigated.

At first the coefficients of the conventional QPLS method are given in Table 6. It is obvious that the coefficients of the fourth and fifth component model are quite large, which are underlined. Therefore it can be said that the inner relation function is not optimized properly.

Next, the coefficients of the GA-QPLS method are investigated. The coefficients are given in Table 7. The coefficients of the fourth and fifth component mode are not quite large compared with those of Table 6. The coefficients of the sixth component that are underlined are relatively larger than those of the fifth one. Judging from the explained variance, the variance does not improve any more when the number of the components is more than five. Therefore it can be said that the sixth component model would be derived from the residual of the model. It means that no meaningful models can be obtained from the residual, which should be reasonable. In this way the validity of the proposed GA-QPLS method can be shown.

In this work, the quadratic inner relation function is optimized successfully by means of GA. What is more, the GA-QPLS method has an advantage, in which it can be also expanded to optimize not only arbitrary polynomial functions but also nonlinear functions. Therefore it can be said that the hybrid method is a powerful and versatile method. Recently many nonlinear PLS methods (NLPLS) have been proposed, in which the mathematical descriptions are complicated, and it is also difficult to optimize inner relation functions of the methods. For example, in the Spline PLS (SPL-PLS) method initial guess is also determined by PLS, causing the same problem of the QPLS method. Therefore few applications of those methods have appeared in the QSAR/QSPR studies.[26−29] If the methods are integrated with GA likewise, the integrated methods would overcome some shortcomings that the conventional method cannot deal with, and then the methods would be quite useful. GA will be used more and more as an optimization method in situations where classical methods cannot be applied.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support of Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

REFERENCES AND NOTES

(1) Hansch, C.; Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc*. **1964**, *86*, 1616−1626.
(2) Dunn, III, W. J.; Wold, S.; Edlund, E.; Hellberg, S.; Gasteiger, J. Multivariate Structure-Activity Relationships between Data from a Battery of Biological Tests and an Ensemble of Structure Descriptors: The PLS Method. *Quant. Struct.-Act. Relat*. **1984**, *3*, 131−137.
(3) Geladi, P.; Kowalski, B. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(4) Miyashita, Y.; Ohsako, H.; Takayama, C.; Sasaki, S. Multivariate Structure-Activity Relationships Analysis of Fungicidal and Herbicidal Thiolcarbamates Using Partial Least Squares Method. *Quant. Struct.-Act. Relat.* **1992**, *11*, 17−22.
(5) Hasegawa, K.; Yokoo, N.; Watanabe, K.; Hirata, M.; Miyashita, Y.; Sasaki, S. Multivariate Free-Wilson Analysis of α-Chymotrypsin Inhibitors Using PLS. *Chemon. Intell. Lab. Syst*. **1996**, *33*, 63−69.
(6) Lindberg, W.; Ohman, J.; Wold, S. Multivariate Resolution of Overlapped Peaks in Liquid Chromatography Using Diode Array Detection. *Anal. Chem*. **1988**, *58*, 299−303.
(7) Cramer, III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis(CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Protein. *J. Am. Chem. Soc*. **1988**, *110*, 5959−5967.
(8) Berglund, A.; Wold, S. INLR, Implicit Non-Linear Latent Variable Regression. *J. Chemom*. **1997**, *11*, 141−156.
(9) Wold, S.; Wold, N. K.; Skagerberg, B. Nonlinear PLS Modeling. *Chemom. Intell. Lab. Syst.* **1989**, *7*, 53−65.
(10) Kimura, T.; Miyashita, Y.; Funatsu, K.; Sasaki, S. Quantitative Structure-Activity Relationships of the Synthetic Substrates for Elastase Enzyme Using Nonlinear Partial Least Squares Regression. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 185−189.
(11) Hasegawa, K.; Kimura, T.; Miyashita, Y.; Funatsu, K. Nonlinear Partial Least Squares Modeling of Phenyl Alkylamines with the Monoamine Oxidase Inhibitory Activities. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1025−1029.
(12) Zupen, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH: New York 1993.
(13) Smits, J. R. M.; Melssen, W. J.; Buydens, L. M. C.; Kateman, G. Using Artificial Neural Networks for Solving Chemical Problems Part I. Multi-Layer Feed-Forward Networks. *Chemom. Intell. Lab. Syst*. **1994**, *22*, 165−189.
(14) Gakh, A. A.; Gakh, E. G.; Sumpter, B. G.; Noid, D. W. Neural Network-Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 832−839.
(15) Gemperline, P. J.; Long, J. R.; Gregoriou, V. G. Nonlinear Multivariate Calibration Using Principal Components Regression and Artificial Neural Networks. *Anal. Chem*. **1991**, *63*, 2313−2323.
(16) Tetteh, J.; Metcalfe, E.; Howells, S. L. Optimization of Radial Basis and Backpropagation Neural Networks for Modelling Auto-Ignition Temperature by Quantitative-Structure Property Relationships. *Chemom. Intell. Lab. Syst*. **1996**, *32*, 177−191.
(17) Wang, J. H.; Jiang, J. H.; Yu, R. Q. Robust Back Propagation Algorithm as a Chemometric Tool to Prevent the Overfitting to Outliers. *Chemom. Intell. Lab. Syst*. **1996**, *34*, 109−115.
(18) Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York, 1991.
(19) Hibbert, D. B. Genetic Algorithm in Chemistry. *Chemom. Intell. Lab. Syst*. **1993**, *19*, 277−293.
(20) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom*. **1992**, *6*, 267−281.

(21) Rimbaud, D. J.; Massart, D. L.; Leardi, R.; Noord, O. E. D. Genetic Algorithms as a Tool for Wavelength Selection in Multivariate Calibration. *Anal. Chem.* **1995**, *67*, 4295−4301.
(22) Hibbert, D. B. A Hybrid Genetic Algorithm for the Estimation of Kinetic Parameters. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 319−329.
(23) Prior to the present work, we applied GA to the optimization (learning) of neural networks, in which each weight and threshold is coded as a bit string. In that case the favorable number of bits needed to code is ten. It would be thought ten bits are needed for parameter estimation. Therefore in the present work, the element of **w** is also coded with a bit string of length ten.
(24) Wold, S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1978**, *22*, 397−405.
(25) Cramer, III, 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.
(26) Frank, I. E. A Nonlinear PLS. *Chemom. Intell. Lab. Syst.* **1990**, *8*, 109−119.
(27) Wold, S. Nonlinear Partial Least Squares Modelling II. Spline Inner Relation. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 71−84.
(28) Taavitsainen, V. M. Nonlinear Data Analysis with Latent Variables. *Chemom. Intell. Lab. Syst.* **1992**, *19*, 185−194.
(29) Hokuldsson, A. Quadratic PLS Regression. *J. Chemom.* **1992**, *6*, 307−334.

CI970026I