Journal of Molecular Graphics and Modelling 29 (2010) 188–196

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

QSAR modeling of peptide biological activity by coupling support vector machine with particle swarm optimization algorithm and genetic algorithm

Xuan Zhou [a,b], Zhanchao Li [a], Zong Dai [a], Xiaoyong Zou [a,*]

[a] School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, PR China
[b] School of Pharmacy, Guangdong Pharmaceutical University, Guangzhou 510006, PR China

ARTICLE INFO

Article history:
Received 3 March 2010
Received in revised form 26 May 2010
Accepted 13 June 2010
Available online 18 June 2010

Keywords:
Quantitative structure–activity relationship
Particle swarm optimization algorithm
Genetic algorithm
Support vector machine
Peptide

ABSTRACT

A novel method coupling particle swarm optimization algorithm (PSO) and genetic algorithm (GA) was proposed to optimize simultaneously the kernel parameters of support vector machine (SVM) and determine the optimized features subset. By coupling GA with PSO, the particles produced in each generation in PSO algorithm were processed by crossover and mutation of GA, and then the particles could keep diversity to escape from local optima and find the global optima quickly and accurately. In order to evaluate the proposed method, four peptide datasets were employed for the investigation of quantitative structure–activity relationship (QSAR). The structural and physicochemical features of peptides from amino acid sequences were used to represent peptides for QSAR. The correlation coefficients (R) of training set of the four datasets were 1.0000, 0.9508, 1.0000, 0.9995, the R of test set of the four datasets were 0.9922, 0.9687, 0.9022, 0.7404, respectively. The root-mean-square errors (RMSEs) of training set of the four datasets were 0.0000, 0.0986, 0.0000, 0.0203, the RMSEs of test set of the four datasets were 0.2522, 0.2782, 0.9625, 0.2928, respectively. A protein dataset, which consists of 277 proteins, was also employed to evaluate the current method for predicting protein structural class, and the good results of overall success rate were obtained. The results indicated that the proposed method might hold a high potential to become a useful tool in peptide QSAR and protein prediction research.

1. Introduction

Peptides are very important in all living systems. They act as hormones, enzyme inhibitors, antibodies, olfaction and taste receptors, antimicrobial compounds or agents, and other biological functions. Due to their high activity, high selectivity and little side effects, peptides have attracted considerable pharmacological interest [1,2].

With the development of a peptide library, thousands of different peptides were designed, synthesized, and then subjected to a range of experimental screening procedures and biological assays. However, the experimental methods were time consuming and expensive, while the computational methods such as QSAR can be helpful for reducing the load of experiments.

The QSAR research has been widely applied to the prediction of biological activity or physicochemical properties of compounds in chemical, environmental, and pharmaceutical fields [3], it can be helpful for investigating the mechanisms of drug actions and transforming structure for drug design [4–6]. The basic assumption in QSAR is that the biological activity within a set of compounds is related to the structural variation of the compounds, i.e., the biological activity can be modeled as a function of molecular structure. For peptides, a precise amino acid sequence is required for their particular function or biological activity, and a QSAR model can indicate the variation in biological activity correlated with the change in peptide sequence and achieve improved peptide activity by modifying the peptide sequence.

To get a good QSAR model, the description of properties of peptide is quite important. Kidera et al. [7] first coded the natural amino acids through 10 orthogonal factors derived from 188 reported properties. The work was followed by Hellberg et al. [8–11] who developed principal properties, or z-scores, for each of 20 natural amino acids and a series of unnatural amino acids. The z-scores were extracted by principal component analysis (PCA) from the collections of experimental data, such as HPLC retention times, pK_a, NMR-derived properties, and other measurable variables related to hydrophobicity, size, and electronic features. Through z-scores and multivariate statistical regressions, successful models were provided in QSAR studies for peptide active on oxytocin, bradykinin, and substance P receptors. Similar results were obtained by Cocchi and Johansson [12] with another parametrization of amino acid side chains. In this approach, t-scores, derived from PCA of the interaction energies calculated by program GRID [13], turned out to be effective when applied in a QSAR study of a set of dipeptide ACE inhibitors. Collantes and Dunn [14] showed that two computable

* Corresponding author. Tel.: +86 20 84114919, fax: +86 20 84112245.
E-mail addresses: ceszxy@mail.sysu.edu.cn, veego_z@hotmail.com (X. Zou).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.06.002

3D-descriptors of isotropic surface area (ISA) [15] and electronic charge index (ECI) [16] may be effectively applied as side-chain descriptors. Zaliani and Gancia [17] developed new descriptors called MS-WHIM indexes, which were a collection of 36 statistical indexes aimed at extracting and condensing steric and electrostatic 3D-properties of a molecule. The reported descriptors mentioned above provided evidence that calculated structurally derived properties can be used to generate robust description for residues in a peptide sequence.

The modeling techniques are also critical to a good QSAR model. Multifarious modeling techniques have been widely employed to QSAR research, such as multiple linear regressions (MLR) [18], partial least squares analysis (PLS) [18], principal component analysis (PCA) [19], artificial neural networks (ANN) [20] and SVM [21–24].

Based on the structural risk minimization principle, an excellent machine learning method of SVM was first reported by Vapnik et al. [25]. Compared with other machine learning systems, SVM has many attractive features, including the absence of local minima, its speed and scalability and its ability to condense information contained in the training set [26]. As a new and powerful modeling tool, SVM has been extensively used to QSAR research. However, when SVM is utilized to QSAR modeling, two problems are encountered, namely the choice of optimal features subset and the set of kernel parameters.

It is well known that large numbers of features fed to SVM can increase computational complexity [27], suffer from the curse of dimensionality and the risk of over-fitting. On the contrary, a few features that are not relevant to biological activity can result in bad generalization performance and accuracy. Consequently, the selection of optimized features subset is necessary to speed up computation and to improve the generalization performance of SVM.

The set of kernel parameters should be optimized so that the performance of SVM can be brought into full play. These parameters affect more or less the performance of SVM, including the penalty constant C and the parameters in the kernel function (width parameter r of radial basis function, etc.).

It is important to provide adequate solutions for the choice of optimal features subset and the set of kernel parameters. However, SVM does not offer the option of a free choice of the optimal features subset and the set of the kernel parameters. A number of different heuristic algorithms, such as PSO [27], ant colony optimization algorithm (COA) [28], artificial immunization algorithm (AIA) [29], and GA [30], have been applied for feature selection, while the kernel parameters are usually selected by experience when a SVM system is constructed. However, features subset choice and kernel parameters setting should be considered together to achieve the highest regression accuracy, because the features subset choice influences kernel function and corresponding parameters setting and vice versa [31].

The GA can be used to select simultaneously the features subset and optimize kernel parameters of SVM [32]. This algorithm has a wide range of applications, however, it generally has good search accuracy but poor search precision [33]. Although GA approaches the globally optimal solution, it commonly fails to find the optimum solution [33]. Moreover, the operational speed of GA is always very slow.

Other algorithms such as PSO can also be used to select simultaneously the features subset and optimize kernel parameters of SVM [34]. However, in PSO algorithm, the search might lead to premature convergence and plunge into a local optimum.

In order to utilize fully the advantages of GA and PSO, a novel method coupling PSO and GA was presented to optimize simultaneously the kernel parameters of SVM and to determine the optimized features subset. The proposed method was evaluated by predicting activities of four peptide datasets and protein structural class of one protein dataset, and the results indicated that it was a useful tool for the investigation of QSAR and protein prediction research.

2. Materials and methods

2.1. Datasets

Four datasets used in the present work were taken from the literatures [35–38], the amino acid sequences of the peptides, their corresponding experimental and predicted biological activity values were listed in the Table S1–S4 (see from the supplementary information).

Dataset-1 [35] consists of 24 melittin omission analogues. Melittin, a predominant peptide isolated from honeybee venom, is known for its marked cytolytic activity. The sequence of melittin is: *GIGAVLKVLTTGLPALISWIKRKRQQ*, each position of melittin sequence was separately omitted to generate the complete series of 24 omission analogues. Their activities were expressed as log HD50 (HD50: "Hemolytic Dose" necessary to lyse 50% of the cells). The analogues of 5, 10, 15 and 20 were selected as test set and the other 20 analogues were taken as training set.

Dataset-2 [36] consists of 45 sauvagine analogues. Sauvagine is a good corticotrophin by releasing factor receptor 2 (CRF2R). The sequence of sauvagine is: *QGPPISIDLSLELLRKMIEIEKQEKEKQQ*, the analogues were optimized with amino acids at positions 11 and 12 (all with proline at position 10). Their activities were expressed as logEC50 of CRF2R. The analogues of 10, 16, 22, 30, 35 and 44 were selected as test set and the others were taken as training set.

Dataset-3 [37] consists of 101 antibiotic peptides called CAMEL-s. These compounds represent derivatives from the hybrid polypeptide CAMEL0 previously created by the respective fusion of the C- and N-terminus sequences of natural peptides cecropin and melittin. Their activities were expressed as antibacterial potencies. The test set and training set are same as the literature [37]. The samples of 1–91 are training set, and the samples of 92–101 are test set.

Dataset-4 [38] consists of 117 peptides that bind to the class I major histocompatibility complex molecule HLA-A*0201. The binding affinity data were expressed as pIC50 (−log IC50) in terms of molar concentration. The test set and training set is same as the literature [38]. The samples of 1–90 are training set, and the samples of 91–117 are test set.

2.2. Descriptor calculation and preprocess

The collection of molecular descriptors plays an important role in the QSAR study. In this work, 11 structural and physicochemical features of peptide from amino acid sequences were used to describe peptide, including amino acid composition, dipeptide composition, autocorrelation, composition, transition and distribution, sequence order and pseudo-amino acid composition. These features, which were usually used as descriptors in predicting protein structural classification [28], were expanded to peptide QSAR by characterizing peptides. The features were summarized in Table S5 (see from the supplementary information), which can be computed by the PROFEAT web server [39]. PROFEAT is accessible at http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi. The numbers of obtained descriptors of datasets-1–4 were 1491, 1497, 1481 and 1475, respectively.

These features from amino acid sequences were usually used in predicting protein structural classification. In previous study of peptide QSARs, the descriptors of peptide were always the structural and physicochemical properties of single amino acid, not related to the amino acid sequence, and then the descriptors were more suitable for peptides with short sequence. However, the pep-

tides in this work have relative long amino acid sequences, so the descriptors from amino acid sequences were chosen.

In order to get the set of informative descriptors, the following pre-processing steps were taken: the constant descriptors for all peptides in one dataset were discarded. As a result, datasets-1–4 consist of 906, 1070, 711, 702 descriptors, respectively. Before modeling, these descriptors values were subjected to the following conversion:

$$X(i)_{new} = \frac{X(i) - X_{\min}}{X_{\max} - X_{\min}}$$

where $X(i)$ is descriptor value for the peptide of i, $X(i)_{\text{new}}$ is the conversion value, $X_{\max}$ and $X_{\min}$ are the maximum and minimum value, respectively. Therefore, each new descriptor value is in the range of 0 and 1.

2.3. Support vector machine

The SVM as a novel type of learning machine is gaining rapid popularity due to its remarkable generalization performance [40]. The basic idea of SVM is to map the original data into a higher dimensional feature space via a kernel function and then to do classification in this space by constructing an optimal separating hyperplane. The SVM was initially developed for binary classification problems, and now SVM can also be utilized to solve nonlinear regression estimation by the introduction of ε-insensitive loss function. A detailed depiction to the theory of SVM for classification and regression can be referred to the literatures [41,42].

The public available LIBSVM software [43] can be downloaded freely from http://www.csie.ntu.edu.tw/~cjlin/libsvm, which was used to process the SVM regression. The radial basis function was selected as the kernel function due to its effectiveness and speed in training process. The kernel parameters including the penalty constant C, the parameters in kernel function (width parameter σ of radial basis function), and ε of the ε-insensitive loss function, were optimized with features subset simultaneously by coupling GA and PSO with SVM.

2.4. GA

Holland [44] began his work on GA at the beginning of the 1970s. The basic idea of GA [30] is as follow: the genetic pool of a given population potentially contains the solution, or a better solution, to a given adaptive problem. This solution is not "active" because the genetic combination on which it relies is split between several subjects. Only the association of different genomes can lead to the solution. No subject has such a genome, but new genetic combination occurs during reproduction (inherit a "good gene" from parents), crossover and mutation. The new genetic solutions can greatly improve the capability of the algorithm to approach, and eventually find the optimum.

2.5. PSO

In 1995, the continuous version of PSO algorithm was proposed as a heuristic swarm intelligent global optimization technique by Kennedy and Eberhart [45]. For the PSO algorithm each individual as a particle represents a potential solution in the multidimensional search space. In each generation, the updated position of every particle benefits from the experience of itself and other particles of the swarm, namely, the velocity of each particle is changed toward the personal best position (P_i) and the global best position (P_g).

Subsequently, Kennedy and Eberhart also presented binary version of PSO algorithm for discrete combinatorial optimization problem in 1997 [46]. In the PSO algorithm, the position of every particle is restricted to 0 and 1 binary search space and the velocity represents the probability that the position of each dimension take the value 1 or 0. The velocity updating equation remains unchanged and the velocity of every dimension is mapped to the interval [0,1] by a sigmoid function.

The binary and continuous PSO can be combined to optimize features subset and kernel parameters simultaneously. To implement this proposed method, each particle was encoded to a string that was composed of two parts: binary and decimal coding systems. To the two parts, the velocity and position of each particle was updated according to the modified binary and continuous PSO, respectively. The binary coding system consists of binary bits for the selection of descriptors. A bit "0" implies that the corresponding descriptor was excluded from the features subset, otherwise, the descriptor was included. The decimal coding systems include three real numbers denoting kernel parameters (C, σ, and ε). The features subset and kernel parameters were decided by the 5-fold cross-validation.

2.6. PSO–GA–SVM

We tried to couple GA with SVM (GA–SVM) to optimize features subset and kernel function. The result indicated that the convergence speed was too slow and the number of descriptors optimized was not ideal. The GA is a general algorithm, and its solution is always satisfactory but not the optimum.

We also tried to couple PSO with SVM (PSO–SVM) to optimize features subset and kernel function. The result indicated that the convergence speed was very fast and the model accuracy was not ideal. In PSO algorithm, the diversity of particle may reduce quickly and the search might lead to premature convergence and plunge into a local optimum after some iteration, the velocity of particle may be converged near the maximum velocity ($V_{\max}$) or ($-V_{\max}$) along with algorithm execution and the search might not escape from local optima.

Considering the characteristics of the two algorithms, we conceived that if the particles produced in each generation in PSO algorithm were processed by crossover and mutation of GA, the particles may keep diversity to escape from local optima. Therefore, a method was proposed by coupling GA with PSO to maintain particle multiformity and large movement of velocity for PSO algorithm. In this method, PSO was coupled with GA to optimize simultaneously the kernel parameters of SVM and to determine the optimized features subset. The PSO–GA–SVM procedure is described as follows:

Step 1 Produce all the initial strings of PSO randomly with an appropriate size of population.
Step 2 Run SVM and calculate the fitness values of each particle in the population by fitness function. If the iteration number comes up to the predefined maximum iteration number, the process is stopped with the output of results, otherwise, go to the next step.
Step 3 Update velocity and position of each member based on PSO. While, select a given percentage of the personal best position of the particle that has the best fitness values in the current generation. The global best position is also selected. The selected member as a part of the next generation is used as parent particle to produce new particle in the next step.
Step 4 Under the guidance of genetic algorithm, produce a given percentage of new members of the next generation by mating operation based on the parents. Five random selected positions are assigned to crossover operator of the binary coding part. For decimal coding part, crossover operator is as follow:

$$P_{i\,new} = P \times P_i + (1 - P) \times P_g$$

where P is the random number of (0, 1).

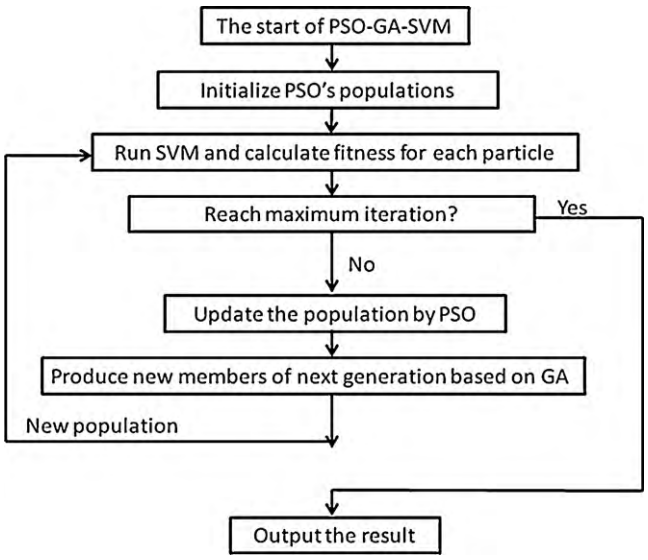

Fig. 1. The chart of PSO–GA–SVM scheme.

Similarly, velocity of new individual is also produced to binary and decimal coding part.

Step 5 Go back to the second step to run SVM and calculate the fitness values of the renewed population.

The whole procedure of PSO–GA–SVM was illustrated in Fig. 1.

The population size of PSO was 30, the father population number was 15, the mutation rate was 3 out of 30 chromosomes, the number of mutation position of every chromosome was 5, and the termination condition was the iteration number of 10,000.

A good fitness function is the key to assess the performance of each particle and to obtain high regression accuracy. Two objectives must be considered for designing fitness function, including the maximization of regression accuracy and the minimization of number of selected descriptors. On the basis of these requirements, a predefined fitness function is presented as follows:

$$\text{fitness} = p \times \text{RMSE} + \frac{(1-p)\times n}{N}$$

$$\text{RMSE} = \sqrt{\frac{\sum (x_{\text{pre}} - x_{\text{exp}})^2}{m}}$$

where RMSE (root-mean-squared error) represents the regression accuracy of SVM based on 5-fold cross-validation (x_{pre} represents the predicted values, x_{exp} represents the experimental values, m is the number of samples), n is the number of selected descriptors, N is the number of informative descriptors by pre-processing, and p is the weighting coefficient controlling the tradeoff between the regression accuracy and the number of selected descriptors. A large value of p may result in more selected descriptors, which can prevent us from finding crucial physicochemical factors. On the contrary, a small value of p is beneficial to the less selected descriptors, but may lead to low regression accuracy. Consequently, an appropriate value of p is vital to accuracy and interpretability of model.

3. Results and discussion

3.1. Analysis of the convergence processes for three methods

The convergence results of the three algorithms (PSO–SVM, GA–SVM, PSO–GA–SVM) for the dataset-1 were discussed as follows.

The convergence situation of GA–SVM, PSO–SVM and PSO–GA–SVM were illustrated in Fig. 2 (a–c). Fig. 2a showed that when the iteration number of GA–SVM reached 100,000, the fitness was inclined to constant and converged, indicating the operational speed was too slow. The iteration number of convergence of PSO–SVM was about 1500, the fast convergence rate might make the model plunge into local optima (Fig. 2b). It was illustrated from Fig. 2c that the iteration number of convergence of PSO–GA–SVM was about 3000.

Obviously, PSO–GA–SVM as improved PSO algorithm had much faster operational speed than GA–SVM, however, slower convergence rate than PSO–SVM. The results may be caused by the mutation and crossover of GA, which raised particle diversity in PSO. By coupling GA with PSO, the convergence curve has a more moderate decline in the gradient than that of PSO, indicating the diversity of particle did not reduce quickly. Therefore, the search may not lead to premature convergence and plunge into a local optimum easily, and the cases of other three datasets were similar.

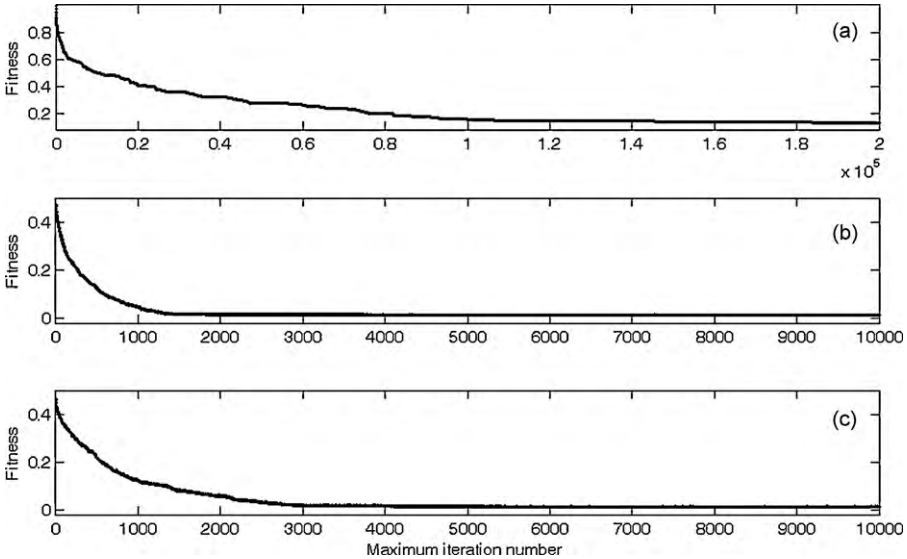

Fig. 2. Minimum fitness value *versus* number of iteration: (a) GA–SVM; (b) PSO–SVM; (c) PSO–GA–SVM.

Table 1
Comparison of different methods of four datasets.

Method	Statistic	Dataset-1		Dataset-2		Dataset-3		dataset-4	
		Train	Test	Train	Test	Train	Test	Train	Test
PSO–GA–SVM	*R*	1.0000	0.9922	0.9508	0.9687	1.0000	0.9022	0.9995	0.7404
	RMSE	0.0000	0.2522	0.0986	0.2782	0.0000	0.9625	0.0203	0.2928
	n	12		8		14		17	
PSO–SVM	*R*	0.9430	0.9892	1.0000	0.7121	1.0000	0.7796	0.8046	0.3839
	RMSE	0.2706	0.1744	0.0000	0.3824	0.0000	1.5665	0.4246	0.4829
	n	5		10		12		3	
GA–SVM	*R*	0.9967	0.9963	0.9927	0.9205	0.9989	0.7843	0.9918	0.6226
	RMSE	0.0885	0.1107	0.0393	0.2983	0.0965	1.3911	0.0921	0.4040
	n	30		36		88		84	
PLS	*R*	0.9999	0.9819	0.9616	0.9605	0.9864	0.7133	0.9823	0.5101
	RMSE	0.0103	0.1656	0.2011	0.6386	0.3340	1.8127	0.1334	0.8501
	n	12		8		14		17	
BP-ANN	*R*	1.0000	0.9280	0.9635	0.7343	1.0000	0.7065	0.9997	0.4597
	RMSE	0.0000	0.5880	0.5450	2.1203	0.0000	0.7677	0.0164	0.6052
	n	12		8		14		17	
ANN [33]	*R*					0.9273	0.8096		
	RMSE					0.6910	1.5871		
G/PLS [34]	*R*							0.9434[a]	0.7141[a]
								0.9110[b]	0.7348[b]
								0.9000[c]	0.7616[c]

R: correlation coefficient; RMSE: root-mean-square error; *n*: the number of optimized features.
[a] Descriptor based.
[b] Single encoding binary
[c] Single + vicinal encoding binary.

3.2. Modeling results comparison with different methods

The evaluation of PSO–GA–SVM and other comparative methods for QSAR of four peptide datasets were listed in Table 1, including correlation coefficient (*R*) between experimental values and predicted values, RMSE and the number of optimized descriptor (*n*).

PSO–GA–SVM has good performances of finding the global optimum by combining PSO and GA. Therefore, the model accuracy was good both for training set and test set, and the appropriate number of optimized descriptors was also obtained.

From Table 1, the best model accuracy was obtained by PSO–GA–SVM. The results of *R* of training set of the four datasets were 1.0000, 0.9508, 1.0000, 0.9995, the *R* of test set of the four datasets were 0.9922, 0.9687, 0.9022, 0.7404, respectively. The RMSE of training set of the four datasets were 0.0000, 0.0986, 0.0000, 0.0203, the RMSE of test set of the four datasets were 0.2522, 0.2782, 0.9625, 0.2928, respectively. The obtained results of datasets-1–3 were satisfied for both training set and test set. For dataset-4, the predicted accuracy for training set was good, but the predicted accuracy for test set was not so ideal. The results may be caused by the fact that the peptides in dataset-4 consist of only nine amino acids. The shorter peptides chain are, the lesser similarity peptides in one dataset have, therefore the untrained peptides are more difficult to be predicted. Therefore, this method is not very appropriate for peptides with chain too short.

In fact, the model accuracy of GA–SVM was comparable to the model accuracy of PSO–GA–SVM, the results of *R* of training set of the four datasets were 0.9967, 0.9927, 0.9989, 0.9918, the *R* of test set of the four datasets were 0.9963, 0.9205, 0.7843, 0.6226, respectively. However, the numbers of descriptor optimized by GA–SVM of the four datasets were 30, 36, 88 and 84, which were much more than that of PSO–GA–SVM (12, 8, 14 and 17).

The model accuracy of PSO–SVM for training set was acceptable. The results of *R* of the four datasets were respective 0.9430, 1.0000, 1.0000, 0.8046, but the *R* of test set of the four datasets were respective 0.9892, 0.7121, 0.7796 and 0.3839, which were obviously poorer than that of PSO–GA–SVM. The results may be caused by the fact that too fast convergence rate leaded to over-fitting, and the obtained results were local optima instead of global optima. The number of descriptors optimized by PSO–SVM was less than that of PSO–GA–SVM on the whole, but at the expense of model accuracy.

The PSO–GA–SVM was compared with the traditional method of PLS. In PLS method, the number of principal component has a great influence on the result, so the number of optimized descriptors of PSO–GA–SVM was chosen as the number of principal component of PLS. The results of *R* of training set of the four datasets were 0.9999, 0.9616, 0.9864, 0.9823, the *R* of test set of the four datasets were 0.9819, 0.9605, 0.7133, 0.5101, respectively. The model accuracy of PLS was poorer than that of PSO–GA–SVM, and more importantly, the principal components of PLS were complex combination of descriptors, not like the definite descriptors optimized by PSO–GA–SVM.

The PSO–GA–SVM was also compared with back propagation artificial neural networks (BP-ANN) method. The descriptors optimized by PSO–GA–SVM were used as the input parameters. The results of *R* of test set of the four datasets were respective 0.9280, 0.7343, 0.7065, and 0.4597, which were much poorer than that of PSO–GA–SVM.

Out of the four datasets, dataset-3 was reported in the literature [37]. The QSAR approach was based on the artificial neural network algorithm by utilizing the “inductive” descriptors. The results of *R* of training set and test set were 0.9273 and 0.8096, which were obviously not as good as the results of PSO–GA–SVM. Dataset-4 was reported in the literature [38]. Compared with the result of literature, the proposed method has better predicted accuracy for training set and similar predicted accuracy for test set.

From the mentioned above, it was shown that the PSO–GA–SVM method was the optimum choice. The predicted values and the experimental values of four datasets were listed in Table S1–S4. From the tables, the predicted values can reproduce the experimental values well. The correlation diagrams between experimental values and predicted values of four datasets were shown in Fig. 3 (a–d) (training sets) and Fig. 4 (a–d) (test sets). In these figures, the points were distributed closely and randomly on both sides of

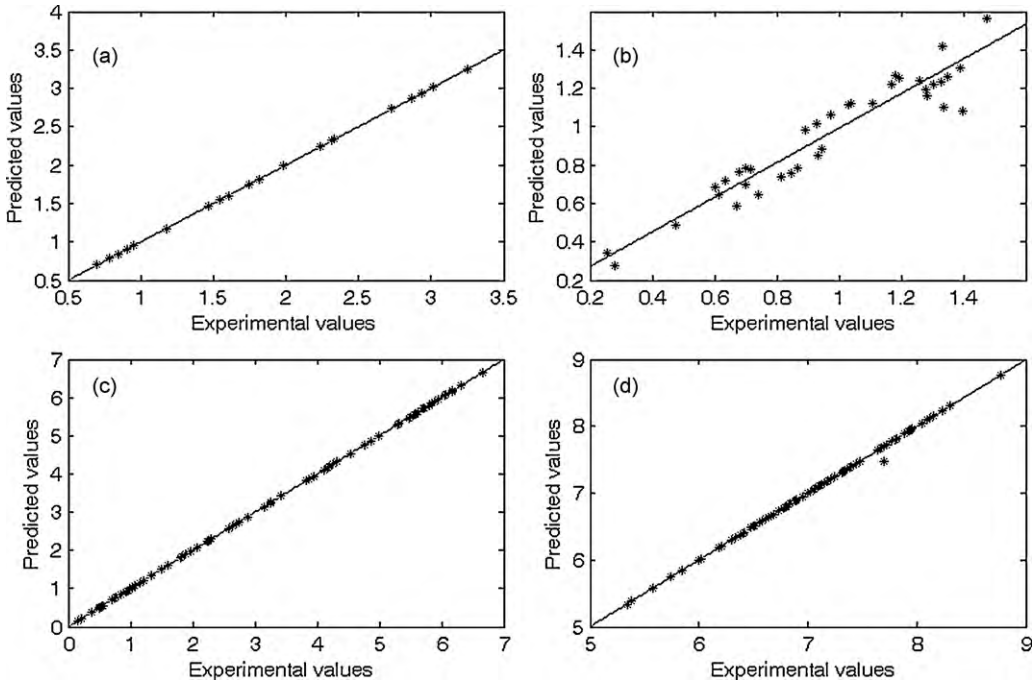

Fig. 3. Plot of experimental *versus* predicted values of training set based on PSO–GA–SVM: (a) dataset-1; (b) dataset-2; (c) dataset-3; (4) dataset-4.

the regression line, indicating good predicted accuracies and no systematic errors in the method.

The method to define the test set in QSAR research is a complex problem. In this paper, the test sets were selected by the amino acid sequences or the activities of peptides with regular intervals. However, to model a more robust QSAR, there are still some pitfalls. For the further validation of our method, the melittin analogues dataset was investigated to process the cross-prediction [47], which was constructed by probing all LSO (leave-4-out) iterations. The prediction results were illustrated in the Fig. 5.

As shown in Fig. 5, the R values of various test sets generated by LSO iterations mostly distributed near 1, which is consistent with the results by the original test set selection ($R = 0.9922$). The number of lower R values decreases dramatically with the decline of R values. The low R values are difficult to be avoided for any predictive method, in that if the objects selected into the training set were easier fitted into the model, then the remaining group may provide a worse fit since only the worse molecules are available for the test set [47].

From above, it can be concluded that the proposed method of PSO–GA–SVM can predict the activities of melittin analogues well even if the test sets were selected by cross-prediction. It can be expected that the other three datasets can get similar results.

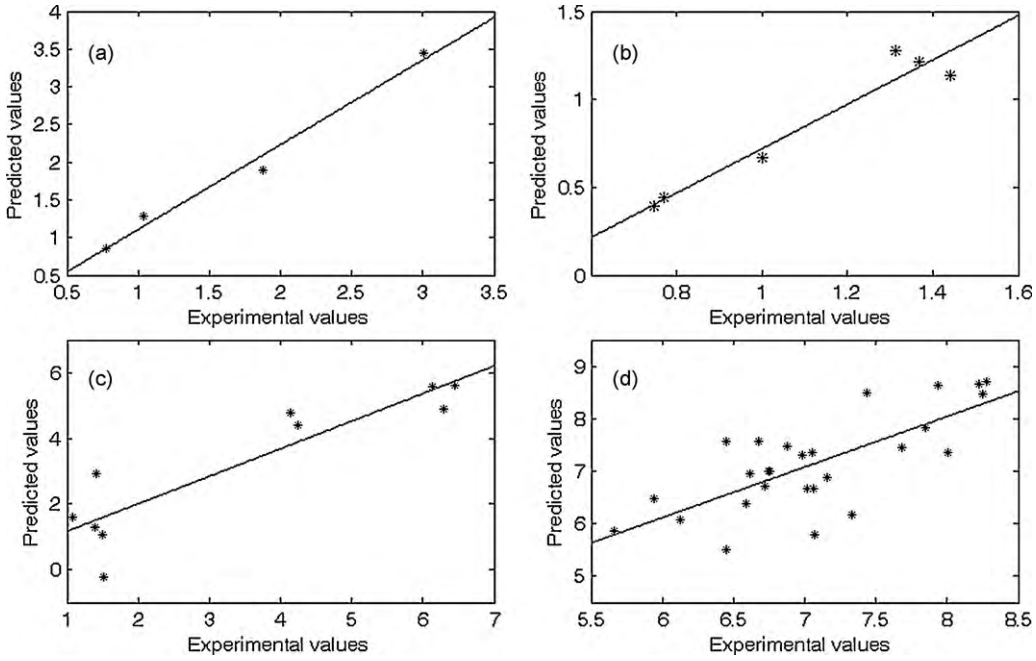

Fig. 4. Plot of experimental *versus* predicted values of test set based on PSO–GA–SVM: (a) dataset-1; (b) dataset-2; (c) dataset-3; (4) dataset-4.

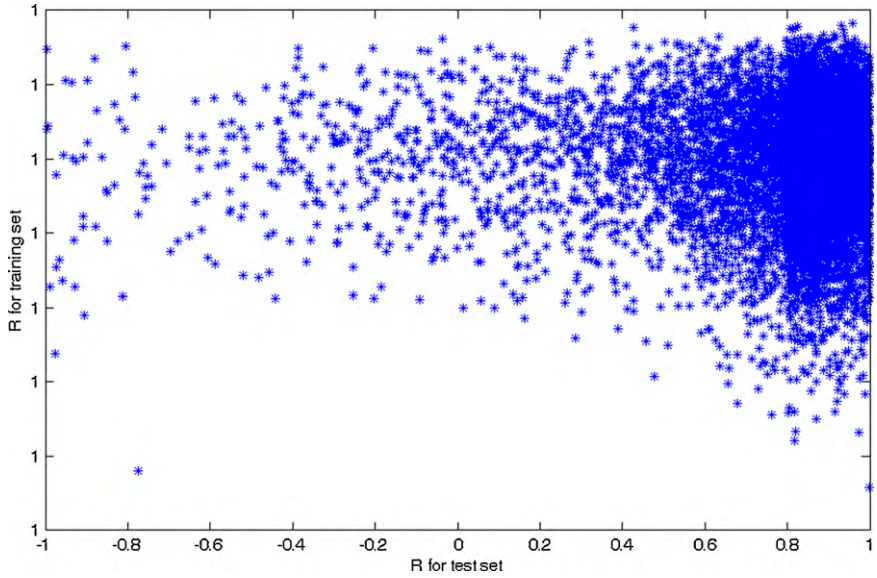

Fig. 5. Plot of prediction accuracies (R) of training sets *versus* test sets by cross-prediction.

3.3. Analysis of the optimized features subset

The optimized features subsets of four datasets by GA–SVM, PSO–SVM and PSO–GA–SVM were summarized in Table 2. From Table 2, it was shown that the optimized features subsets were various with different methods, however the orders of features that contributed to the prediction of peptide activity were similar. The optimized descriptors were mainly distributed in the features of Moran autocorrelation, Moreau–Broto autocorrelation, Geary autocorrelation, distribution and quasi-sequence-order in four datasets, therefore these features have greater contribution to peptide activity than other features.

Autocorrelation described the level of correlation between two objects (protein or peptide sequences) in terms of their specific structural or physicochemical properties [48], which were defined based on the distribution of amino acid properties along the sequence [49]. Three types of autocorrelation were analyzed: the first was Moreau–Broto autocorrelation [50], which has been used for predicting transmembrane protein types [51] and protein secondary structural contents [52], the second was Moran autocorrelation [53], which has been applied for predicting protein helix contents [54], the third was Geary autocorrelation [55], which has been used for analyzing allele frequencies and population structures [56]. Distribution features represent the amino acid distribution patterns of a specific structural or physicochemical property along a protein or peptide sequence [57]. Quasi-sequence-order descriptors were derived from both the Schneider–Wrede [58] physicochemical distance matrix and the Grantham [59] chemical distance matrix between each pair of the 20 amino acids. Not only these features can be applied for predicting protein successfully, but also they have great contributions to QSAR of peptides. We can expect that if a new encoding scheme can integrate with autocorrelation, distribution and quasi-sequence-order, it may be of great significance in terms of predicting the activity of peptide.

For the three methods of GA–SVM, PSO–SVM and PSO–GA–SVM, the features subset of GA–SVM was the largest one, which included the features subset of PSO–GA–SVM and PSO–SVM on the whole. Considering the model accuracy of three methods, it can be concluded that the optimized descriptors by PSO–SVM plunged into local optimum to some extent, and the optimized descriptors by GA–SVM did not find the core descriptors, while the descriptors optimized by PSO–GA–SVM were the most important factors contributed to peptide activity.

For dataset-1, the features subset optimized by PSO–GA–SVM included four Geary autocorrelations of one hydrophobicity, one polarizability and two steric, two Moreau–Broto autocorrelations of flexibility and free energy in water, two Moran autocorrelations of residue accessible surface area in tripeptide and steric, one pseudo-amino acid composition, one composition of normalized vdW volumes, one distribution of hydrophobicity, one quasi-sequence-order descriptors based on normalized Grantham chemical distance. The results suggested that these descriptors were important to melittin analogues activity, including hydrophobicity, polarizability, steric, flexibility, free energy in water, residue accessible surface area in tripeptide, vdW volumes and Grantham chemical distance.

For dataset-2, there were three Moreau–Broto autocorrelations of residue accessible surface area in tripeptide, residue volume and relative mutability, three Moran autocorrelations of hydrophobicity, flexibility and free energy in water, two Geary autocorrelations of hydrophobicity and residue accessible surface area in tripeptide. The results suggested that these descriptors were important to sauvagine analogues activity, including hydrophobicity, flexibility, free energy in water, residue accessible surface area in tripeptide, residue volume and relative mutability.

For dataset-3, there were three Moran autocorrelations of two free energies in water and one relative mutability, one Geary autocorrelation hydrophobicity, one composition of normalized vdW volumes, two transitions of hydrophobicity and secondary structure, three distributions of two charge and one hydrophobicity, one sequence-order-coupling numbers based on normalized Grantham chemical distance, one quasi-sequence-order descriptors based on Schneider–Wrede distance, one pseudo-amino acid composition, one dipeptide composition. The results suggested that these descriptors were important to CAMEL-s activity, including hydrophobicity, free energy in water, relative mutability, vdW volumes, secondary structure, charge, Grantham chemical distance and Schneider–Wrede distance.

For dataset-4, there were one dipeptide composition, two M–B autocorrelations of one polarizability and one relative mutability, one Moran autocorrelation polarizability, two Geary autocorrelations of residue accessible surface area in tripeptide, one composition of solvent accessibility, four distributions of

Table 2
Results of the selection of the best feature subset.

Feature	Number of optimized descriptors											
	Dataset-1			Dataset-2			Dataset-3			Dataset-4		
	GA–SVM	PSO–SVM	PSO–GA–SVM	GA–SVM	PSO–SVM	PSO–GA–SVM	GA–SVM	PSO–SVM	PSO–GA–SVM	GA–SVM	PSO–SVM	PSO–GA–SVM
Amino acid composition	2			2			2			1		
Dipeptide composition				6			15	1	1	31		1
Normalized Moreau–Broto autocorrelation	5	1	2	3		3	11	1		12		2
Moran autocorrelation	3	1	2	6	1	3	16	2	3	9		1
Geary autocorrelation	8		4	4	6	2	17	2	1	4		2
Composition	3	2	1				1	1	1	2		1
Transition	1			3	1		4	1	2	2		
Distribution	3		1	7			13	2	3	13	3	4
Sequence-order-coupling number	2						2		1	2		1
Quasi-sequence-order descriptors	2	1	1	4	2		7	2	1	7		4
Pseudo-amino acid composition	1		1	1					1	1		1
Total number of descriptors	30	5	12	36	10	8	88	12	14	84	3	17

hydrophobicity, one sequence-order-coupling numbers based on Schneider–Wrede distance, four quasi-sequence-order descriptors of three based on normalized Grantham chemical distance and one based on Schneider–Wrede distance, one pseudo-amino acid composition. The results suggested that these descriptors were important to peptides activity in dataset-4, including hydrophobicity, polarizability, relative mutability, residue accessible surface area, solvent accessibility, Schneider–Wrede distance and Grantham chemical distance.

3.4. Prediction of protein structural classes

As an effective QSAR tool for peptides, the current method was also expected as a good method for protein structure prediction, because protein is essentially peptide with long amino acid sequence. To evaluate further the proposed method, a protein dataset was also employed to predict protein structural classes, details about this section were discussed in "supplementary information".

4. Conclusion

A novel method coupled GA with PSO was utilized to model adaptively SVM, in which the features subset was selected and the kernel parameters were optimized simultaneously. The descriptors of peptide, the structural and physicochemical features from amino acid sequence, were first used to investigate QSAR of peptide. The method was also applied to the determination of protein structural class. The results indicated that the proposed method had the ability to achieve good prediction. It can be anticipated that the approach might hold a high potential to become a useful tool in peptide QSAR and protein prediction research.

Acknowledgments

We gratefully acknowledge to the financial support by the National Natural Science Foundation of China (No. 20975117, 20805059), the Natural Science Foundation of Guangdong Province (No. 7003714), and the Ph.D. Programs Foundation of Ministry of Education of China (No. 20070558010).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2010.06.002.

References

[1] J. Jira'cek, A. Yiotakis, B. Vincent, A. Lecoq, F. Checler, V. Dive, Development of highly potent and selective phosphinic peptide inhibitors of zinc endopeptidase 24-15 using combinatorial chemistry, J. Biol. Chem. 270 (1995) 21701–21706.
[2] M. Marraud, A. Aubry, Crystal structures of peptides and modified peptides, Biopolymers 40 (1996) 45–83.
[3] X.J. Yao, A. Panaye, J.P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu, B.T. Fan, Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression, J. Chem. Inf. Comput. Sci. 44 (2004) 1257–1266.
[4] S. Qin, H.X. Liu, J. Wang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Quantitative structure–activity relationship study on a series of novel ligands binding to central benzodiazepine receptor by using the combination of heuristic method and support vector machines, QSAR Comb. Sci. 26 (2007) 443–451.
[5] M.A. Chamjangali, M. Beglari, G. Bagherian, Prediction of cytotoxicity data (CC50) of anti-HIV 5-pheny-l-phenylamino-1H-imidazole derivatives by artificial neural network trained with Levenberg–Marquardt algorithm, J. Mol. Graph. Model. 26 (2007) 360–367.
[6] A. Mohajeri, B. Hemmateenejad, A. Mehdipour, R. Miri, Modeling calcium channel antagonistic activity of dihydropyridine derivatives using QTMS indices analyzed by GA-PLS and PC-GA-PLS, J. Mol. Graph. Model. 26 (2008) 1057–1065.
[7] A. Kidera, Y. Konishi, M. Oka, T. Ooi, H. Scheraga, A Statistical analysis of the physical properties of the 20 naturally occurring amino acids, J. Protein Chem. 4 (1985) 23–55.

[8] S. Hellberg, M. Sjostroem, S. Wold, The prediction of bradykinin potentiating potency of pentapeptides. An example of a peptide quantitative structure–activity relationship, Acta Chem. Scand. Ser. B 40 (1986) 135–140.
[9] S. Hellberg, M. Sjostroem, B. Skagerberg, S. Wold, Peptide quantitative structure–activity relationships, a multivariate approach, J. Med. Chem. 30 (1987) 1126–1135.
[10] S. Wold, L. Eriksson, J. Jonsson, M. Sjostroem, S. Hellberg, B. Skagerberg, C. Wikstroem, Principal property values for six non-natural amino acids and their application to a structure–activity relationship for oxytocin peptide analogues, Can. J. Chem. 65 (1987) 1814–1820.
[11] J. Jonsson, L. Eriksson, S. Hellberg, M. Sjostroem, S. Wold, Multivariate parametrization of 55 coded and non-coded amino acids, Quant. Struct. Act. Relat. 8 (1989) 204–209.
[12] M. Cocchi, E. Johansson, Amino acids characterization by GRID and multivariate data analysis, Quant. Struct. Act. Relat. 12 (1993) 1–8.
[13] P.J. Goodford, A Computational procedure for determining energetically favorable binding sites on biologically important macromolecules, J. Med. Chem. 28 (1985) 849–857.
[14] E.R. Collantes, W.J. Dunn III, Amino acids side chain descriptors for quantitative structure–activity relationship studies of peptide analogues, J. Med. Chem. 38 (1995) 2705–2713.
[15] J. Fauchere, V. Pliska, Hydrophobic parameters of amino acid side chain from the partitioning of N-acetyl-amino-acid amides, Eur. J. Med. Chem. 18 (1983) 369–375.
[16] R. Wolfenden, L. Andersson, P.M. Cullis, C.C.B. Southgate, Affinities of amino acid side chains for solvent water, Biochemistry 20 (1981) 849–855.
[17] A. Zaliani, E. Gancia, MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies, J. Chem. Inf. Comput. Sci. 39 (1999) 525–533.
[18] A.K. Saxena, P. Prathipati, Comparison of MLR, PLS and GA-MLR in QSAR analysis, SAR QSAR Environ. Res. 14 (2003) 433–445.
[19] M. Sun, Y.G. Zheng, H.T. Wei, J.Q. Chen, M. Ji, QSAR studies on 4-anilino-3-quinolinecarbonitriles as Src kinase inhibitors using robust PCA and both linear and nonlinear models, J. Enzyme Inhib. Med. Chem. 24 (2009) 1109–1116.
[20] O. Deeb, B. Hemmateenejad, ANN-QSAR model of drug-binding to human serum albumin, Chem. Biol. Drug Des. 70 (2007) 19–29.
[21] J.P. Doucet, F. Barbault, H.R. Xia, A. Panaye, B. Fan, Nonlinear SVM approaches to QSPR/QSAR studies and drug design, Curr. Comput. Aid. Drug Des. 3 (2007) 263–289.
[22] U. Norinder, Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection, Neurocomputing 55 (2003) 337–346.
[23] C.W. Yap, Z.R. Li, Y.Z. Chen, Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods, J. Mol. Graph. Model. 24 (2006) 383–395.
[24] V.V. Zernov, K.V. Balakin, A.A. Ivaschenko, N.P. Savchuk, I.V. Pletnev, Drug discovery using support vector machines. The case studies of drug-likeness. Agrochemical-likeness, and enzyme inhibition predictions, J. Chem. Inf. Comput. Sci. 43 (2003) 2048–2056.
[25] C. Cortes, V. Vapnik, Support-vector networks, Mach Learn. 20 (1995) 273–297.
[26] C. Chen, X.B. Zhou, Y.X. Tian, X.Y. Zou, P.X. Cai, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Anal. Biochem. 357 (2006) 116–121.
[27] Q. Shen, W.M. Shi, W. Kong, B.X. Ye, A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification, Talanta 71 (2007) 1679–1683.
[28] R.K. Sivagaminathan, S.A. Ramakrishnan, Hybrid approach for features subset selection using neural networks and ant colony optimization, Expert Syst. Appl. 33 (2007) 49–60.
[29] S.F. Yuan, F.L. Chu, Fault diagnosis based on support vector machines with parameter optimization by artificial immunization algorithm, Mech. Syst. Signal Pr. 21 (2007) 1318–1330.
[30] H.M. Jalali, A. Kyani, Application of genetic algorithm kernel partial least square as a novel nonlinear feature selection method: activity of carbonic anhydrase II inhibitors, Eur. J. Med. Chem. 42 (2007) 649–659.
[31] C.L. Huang, J.F. Dun, A distributed PSO–SVM hybrid system with feature selection and parameter optimization, Appl. Soft Comput. 8 (2008) 1381–1391.
[32] Z.C. Li, X.B. Zhou, Y.R. Lin, X.Y. Zou, Prediction of protein structure class by coupling improved genetic algorithm and support vector machine, Amino Acids 35 (2008) 581–590.
[33] P.N. Stefan, Artificial neural networks and genetic algorithm in QSAR, J. Mol. Struct. (Theochem.) 622 (2003) 71–83.
[34] C.L. Huang, H.C. Liao, M.C. Chen, Prediction model building and feature selection with support vector machines in breast cancer diagnosis, Expert Syst. Appl. 34 (2008) 578–587.
[35] S.E. Blondelle, R.A. Houghten, Hemolytic and antimicrobial activities of the twenty-four individual omission analogues of melittin, Biochemistry 30 (1991) 4671–4678.
[36] J.I. Robert, W. Feng, T. Michelle, D. Elizabeth, B.B. Mary, L. Frank, T.H. Richard, W.M. Adam, Discovery of corticotropin releasing factor 2 receptor selective sauvagine analogues for treatment of skeletal muscle atrophy, J. Med. Chem. 48 (2005) 262–265.
[37] C. Artem, J. Bojana, Application of 'inductive' QSAR descriptors for quantification of antibacterial activity of cationic polypeptides, Molecules 9 (2004) 1034–1052.
[38] R.S.P. Raghuvir, K.M. Alpeshkumar, A.K. Santosh, C.C. Evans, Encoding type and position in peptide QSAR: application to peptides binding to class I MHC molecule HLA-A*0201, QSAR Comb. Sci. 26 (2007) 189–203.
[39] Z.R. Li, H.H. Lin, L.Y. Han, L. Jiang, X. Chen, Y.Z. Chen, PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, Nucleic Acids Res. 34 (2006) W32–W37.
[40] H.X. Liu, R.S. Zhang, X.J. Yao, M.C. Liu, Z.D. Hu, B.T. Fan, Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs, J. Chem. Inf. Comput. Sci. 44 (2004) 161–167.
[41] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[42] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other 5 Kernel-based Learning Methods, Cambridge University Press, Cambridge, 2000.
[43] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/*cjlin/libsvm (2001).
[44] J. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, 1975.
[45] J. Kennedy, R.C. Eberhart, Particle swarm optimization, in: Proceedings of the IEEE International Conference on Neural Network, vol. 4, IEEE Inc., Perth, 1995, pp. 1942–1948.
[46] J. Kennedy, R.C. Eberhart, A discrete binary version of the particle swarm algorithm, Int. Conf. Syst. Man Cybernet. 5 (1997) 4104–4108.
[47] J. Polanski, A. Bak, R. Gieleciak, T. Magdziarz, Modeling robust QSAR, J. Chem. Inf. Model. 46 (2006) 2310–2318.
[48] P. Broto, G. Moreau, C. Vandicke, Molecular structures: perception, autocorrelation descriptor and SAR studies, Eur. J. Med. Chem. 19 (1984) 71–78.
[49] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 28 (2000) 374.
[50] G. Moreau, P. Broto, Autocorrelation of molecular structures, application to SAR studies, Nouv. J. Chim. 4 (1980) 757–764.
[51] Z.P. Feng, C.T. Zhang, Prediction of membrane protein types based on the hydrophobic index of amino acids, J. Protein Chem. 19 (2000) 269–275.
[52] Z. Lin, X.M. Pan, Accurate prediction of protein secondary structural content, J. Protein Chem. 20 (2001) 217–220.
[53] P.A. Moran, Notes on continuous stochastic phenomena, Biometrika 37 (1950) 17–23.
[54] D.S. Horne, Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities, Biopolymers 27 (1988) 451–477.
[55] R.C. Geary, The contiguity ratio and statistical mapping, Incorpor. Stat. 5 (1954) 115–145.
[56] R.R. Sokal, B.A. Thomson, Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population, Am. J. Phys. Anthropol. 129 (2006) 121–131.
[57] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, S.H. Kim, Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification, Proteins 35 (1999) 401–407.
[58] G. Schneider, P. Wrede, The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site, Biophys. J. 66 (1994) 335–344.
[59] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864.