Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2007) 269–281
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Scores of generalized base properties for quantitative sequence-activity modelings for *E. coli* promoters based on support vector machine

Guizhao Liang [a], Zhiliang Li [a,b,*]

[a] *College of Bioengineering, Chongqing University, Chongqing 400030, PR China*
[b] *College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400030, PR China*

Received 30 June 2006; received in revised form 18 November 2006; accepted 10 December 2006
Available online 15 December 2006

Abstract

A novel base sequence representation technique, namely SGBP (scores of generalized base properties), was derived from principal component analysis of a matrix of 1209 property parameters including 0D, 1D, 2D and 3D information for five bases such as A, C, G, T and U. It was then employed to represent sequence structures of *E. coli* promoters. Variables which were used as inputs of partial least square (PLS) and support vector machine (SVM) were selected by genetic arithmetic-partial least square. All samples were divided into train set which was applied to develop quantitative sequence-activity modelings (QSAMs) and test set which was used to validate the predictive power of the resulting models according to D-optimal design. Investigation on QSAM by PLS showed properties of base of position −42, −34, −31, −33, −41, −46 and −29 may yield more influence on strengths, which has thus pointed us further into the direction of strong promoters. Parameters of SVM were determined by response surface methodology. Satisfactory results indicated that the simulative and the predictive abilities for the internal and external samples of QSAM by SVM were better than those of PLS. Those results showed that SGBP is a useful structural representation methodology in QSAMs due to its many advantages including plentiful structural information, easy manipulation, and high characterization competence. Moreover, SGBP-GA-SVM route for sequences design and activities prediction of DNA or RNA can further be applied.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Scores of generalized base properties (SGBP); Quantitative sequence-activity modeling (QSAM); Support vector machine (SVM); DNA; Promoter

1. Introduction

Since the DNA double helix structure was advocated by Jim Watson and Francis Crick, essence of life has been understood by molecular level. Subsequently, triplet codons and codons table have been determined, gene as coded proteins in DNA have been confirmed, the central rule that genetic information is from DNA to RNA, and then to protein has been known well, operon model for genes expression and regulation has been designed, and semi-conservative duplication for DNA has been made sure, which have constituted a perfect picture from storage, expression, regulation, and replication of genetic information to accomplishment of complicated functions. With the accomplishment of human genome project, the number of sequences and structures for nucleic acids and proteins is increased by exponent. Ten of millions base pairs sequences and hundred of thousands protein sequences have been determined. So, it is not practical to determine their structures and functions only by experiment. A great number of sequences are analyzed and picked out only by previous data and experience. Then questions that we will deal with are resorted to experiment. As a result, limited manpower and material resources are used. Those tasks are accomplished only by computer-aided methodology. Some achievements on it have been obtained [1–5]. However, those were achieved based on arithmetic modification and statistic probability. Quantitative structural characterization methods related to bioactivity are relatively few.

QSAM provides a new tool for those questions to be dealt with. Those studies were initiated by Sneath [6]. Consequently, sequences of *E. coli* promoter and lysogenization were characterized using binary code, and satisfying results were obtained based on PLS by Mulligan et al. [7] and Jonsson et al. [8]. Technique for sequence structural representation and modeling is the key to quantitatively predicting bioactivity. In

* Corresponding author at: College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400030, PR China. Tel.: +86 23 65106677.
E-mail addresses: sdqdlgz@163.com (G. Liang), liszyx@126.com (Z. Li).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.12.004

present study, we mainly focus on sequence representation and modeling techniques which are crucial to QSAM. SGBP (scores for generalized base properties) was derived from principal component analysis of a matrix of 1209 property parameters of five bases including A, C, G, T and U. It was then employed to character structures of promoters of *E. coli*. SGBP scales which are easily manipulated contain a great deal of information related to base sequences. Multiple linear regression, principal component regression, and partial least square can be used to develop linear models. ANN can be applied to build nonlinear models. Generally, favorable results can be obtained by ANN, but over-fitting may arise. In addition, because its structures are not easily determined, the results from ANN are instable. A new machine learning arithmetic-support vector machine (SVM) has widely been investigated because of its high generality since 1990s of last century [9–11]. There has been a lot of interest in studying SVM in the field of machine learning due to its remarkable generalization performance in recent years. It can deal with nonlinear, high dimension and local least plights using structural risk minimization principle but not using traditional empirical risk minimization principle. Over-fitting of data can be theoretically avoided using SVM algorithm. Because there are no rigorous rules to be followed to select the parameters of SVM, we tentatively used response surface methodology to select the parameters. The results from SVM were compared with those from PLS. Satisfying results from SVM showed that SVM can be further used to construct QSAMs.

2. Theory and methods

2.1. SGBP scales generation

DNAs or RNAs are linear macromolecules which are made up of five-membered deoxyribose rings combined with A, C, G, and T or U by phosphodiester bond. So, various positions and properties of bases will lead to different functions of sequences. Herein, a total of 1209 various property descriptors is collected to describe DNA or RNA sequence structural diversities. The descriptors categories are as follows: (a) 0D-29 constitutional descriptors related to atom and bond counts, molecular weight, and sum of atomic properties [12]; (b) 1D-81 descriptors including functional groups counts, atom centered fragments, and molecular properties [12,13]; (c) 2D-514 descriptors including topological descriptors obtained from molecular graph including MEDV and MHDV descriptors, topological descriptors, work and path counts descriptors, connectivity index descriptors, information index descriptors, auto-correlations descriptors from the molecular graph, edge adjacency indices, Burden eigenvalue, topological charge indices descriptors, and eigenvalue-based indices descriptors [14–19]; (d) 3D-585 descriptors including Randic molecular profile descriptors derived from the distance distribution moments of the geometry matrix, geometric descriptors based on molecular geometry, RDF descriptors obtained by radial basis functions centred on different interatomic distances, MoRSE descriptors predicted by summing atom weights viewed by a different angular scattering function, WHIM descriptors based on the calculation of principal component axes calculated from a weighted covariance matrix obtained by the molecule geometrical coordinates, and GETAWAY descriptors predicted from the leverage matrix obtained by the centred atomic coordinates [20–26].

Because molecular descriptor variables may be highly correlated with each other, principal component analysis (PCA) was employed to find linear combinations of features that capture the variation between different kinds of bases. PCA, as a useful tool for dimensionality reduction technique, whose main idea is to project the data from a high dimensional space onto a lower dimensional space, has been widely used in many aspects such as data compression, pattern recognition, time series prediction as well as multivariate process monitoring [27]. The system is chosen such that the greatest variance is captured by the first axis, or the first ''principal component''. Successive principal components capture progressively less variance. Each component is a linear combination of some of the initial features. From these 1209 parameters, we removed redundant parameters for which the magnitude of the correlation coefficient with another parameter was greater than 0.90. The remaining 41 independent parameters were kept. The first four principal components (PCs) obtained from PCA account for 99.9998% of variable dispersion. That is to say, the first four PCs scores can explain the most information in the original data matrix (5×41). So, original data matrix (5×41) can be replaced by these four PCs scores matrix (5×4). Here, these four score vectors are tentatively called scores for generalized base properties (SGBP) (Table 1).

The loadings reflect the relative contribution of each variable to the four SGBP vectors (Table 2). The first principal component reflects these variables information of third component symmetry directional WHIM index (weighted by atomic masses) with a largest positive coefficient, structural information content (neighborhood symmetry of 1-order), K global shape index (weighted by atomic electrotopological states), lag 2 (weighted by atomic polarizabilities) of Moran autocorrelation with a largest negative coefficient, Torsion energy, and sum of atomic van der Waals volumes (scaled on carbon atom).

The second principal component mainly relates to unweighted signal 29 of 3D-MoRSE which is a descriptor of molecule representation of structure based on electron diffraction, Electronic energy, lag 6 (weighted by atomic van

Table 1
Four PC solution scores for the 41 selected base properties

Bases	$SGBP_1$	$SGBP_2$	$SGBP_3$	$SGBP_4$
A	−3.9505	4.0764	−1.1507	1.2426
C	4.3677	1.0541	1.5173	3.2084
G	−2.7552	−4.8467	1.1540	1.4321
T	0.4217	0.8763	3.3983	−4.0915
U	1.9163	−1.1601	−4.9190	−1.7917
Eigenvalues	11.5312	10.8331	10.1758	8.4599
Variance explained (%)	28.1249	26.4223	24.8189	20.6337
Cumulative variance explained (%)	28.1249	54.5472	79.3661	99.9998

Table 2
Loadings from PCA of the descriptor matrix (5 × 41) for bases

No.	Base properties	PC_1	PC_2	PC_3	PC_4
1	Average molecular weight	0.007	−0.120	−0.283	−0.056
2	Sum of atomic van der Waals volumes (scaled on carbon atom)	−0.240	−0.028	0.169	0.065
3	Sum of Kier–Hall electrotopological states	−0.102	−0.256	0.036	−0.135
4	Mean atomic polarizability (scaled on carbon atom)	−0.105	−0.046	−0.289	−0.017
5	Mean electrotopological state	0.211	−0.104	−0.091	−0.184
6	Number of multiple bonds	−0.215	0.097	−0.189	−0.006
7	Aromatic ratio	−0.090	0.132	−0.265	−0.023
8	First Zagreb index by valence vertex degrees	−0.203	−0.208	−0.053	−0.054
9	Reciprocal hyper-detour index	0.156	−0.078	0.145	−0.228
10	E-state topological parameter	−0.091	−0.168	0.161	−0.199
11	2-Path Kier alpha-modified shape index	−0.092	−0.087	0.279	−0.055
12	Kier flexibility index	0.064	−0.048	0.273	−0.142
13	Kier benzene-likeliness index	0.097	0.118	0.240	−0.137
14	Sum of topological distances between N and O	−0.056	−0.285	0.075	0.057
15	Structural information content (neighborhood symmetry of 1-order)	0.253	−0.109	−0.111	−0.025
16	Lag 3 (weighted by atomic van der Waals volumes) of Moran autocorrelation	−0.014	−0.160	−0.166	0.228
17	Lag 6 (weighted by atomic van der Waals volumes) of Moran autocorrelation	−0.058	0.226	0.055	0.212
18	Lag 2 (weighted by atomic polarizabilities) of Moran autocorrelation	−0.262	0.108	−0.003	−0.098
19	Lag 4 (weighted by atomic polarizabilities) of Moran autocorrelation	0.123	−0.039	−0.064	−0.301
20	Lowest eigenvalue *n*. 1 of Burden matrix (weighted by atomic polarizabilities)	−0.127	0.106	0.242	0.107
21	Radial Distribution Function-3.0 (weighted by atomic masses)	−0.155	0.193	0.106	−0.156
22	Radial Distribution Function-3.0 (weighted by atomic polarizabilities)	−0.102	0.056	0.180	−0.247
23	Signal 21 (unweighted) of 3D-MoRSE	0.015	0.085	0.052	−0.325
24	Signal 22 (unweighted) of 3D-MoRSE	0.185	0.024	−0.059	0.258
25	Signal 29 (unweighted) of 3D-MoRSE	0.023	0.249	0.005	−0.195
26	Signal 27 (weighted by atomic masses) of 3D-MoRSE	−0.108	−0.260	−0.084	0.085
27	Signal 18 (weighted by atomic van der Waals volumes) of 3D-MoRSE	0.165	−0.215	−0.106	−0.092
28	Signal 16 (weighted by atomic Sanderson electronegativities) of 3D-MoRSE	0.080	0.126	0.201	0.202
29	Third component symmetry directional WHIM index (weighted by atomic masses)	0.289	0.050	0.024	0.020
30	First component shape directional WHIM index (weighted by atomic van der Waals volumes)	−0.069	−0.199	0.221	0.049
31	K global shape index (weighted by atomic electrotopological states)	0.241	0.065	0.164	0.040
32	R maximal autocorrelation of lag 5 (weighted by atomic van der Waals volumes)	0.078	−0.276	−0.071	0.079
33	Number of urea derivatives	0.212	0.054	0.083	0.212
34	Number of acceptor atoms for H-bonds (N O F)	−0.229	−0.145	0.032	0.136
35	Moriguchi octanol–water partition coefficient (log *P*)	−0.130	0.217	−0.114	−0.139
36	The seventh weight molecular holographic distance vector ($MHDV_7$)	0.105	0.135	0.062	0.274
37	HOMO	−0.167	−0.133	0.151	0.173
38	Total energy	0.218	0.183	−0.090	0.026
39	Electronic energy	−0.144	0.242	−0.087	0.078
40	Dipole moment	0.030	−0.086	0.258	0.166
41	Torsion energy	−0.257	0.122	−0.051	0.077

der Waals volumes) of Moran autocorrelation, and Moriguchi octanol–water partition coefficient (log *P*) with relatively high positive loading coefficients. A large negative loading coefficient occurs for sum of topological distances between N and O, R maximal autocorrelation of lag 5 (weighted by atomic van der Waals volumes), signal 27 (weighted by atomic masses) of 3D-MoRSE, and sum of Kier–Hall electrotopological states.

The third principal component basically concerns with 2-path Kier alpha-modified shape index and Kier flexibility index which belong to topological descriptors, and dipole moment which is an electronic descriptor. These properties vary inversely with mean atomic polarizability (scaled on carbon atom), average molecular weight, and aromatic ratio, which represent constitutional features.

By and large, the fourth principal component refers to variables with high coefficients on the seventh weight molecular holographic distance vector ($MHDV_7$) and unweighted signal 22 of 3D-MoRSE. MHDV advocated by our group [15] is an electrotopological descriptor computed as electronic interaction based on 13 atomic types, atomic attributes and relative bond-length. The $MHDV_7$ is the electronic interaction vector between C– and >N–, >P– ("–", ">" and "<" refer to chemical bond connected with a non-hydrogen atom, two non-hydrogen atoms, and two non-hydrogen atoms, respectively). Reversely, it involves the variables with large negative loading coefficient on unweighted signal 21 of 3D-MoRSE, lag 4 (weighted by atomic polarizabilities) of Moran autocorrelation.

2.2. Promoter sequence data and its parametrization

Promoters, as specific DNA sequences, play an essential role in genetic information transferring [28]. Relations between promoter sequences and their strengths were extensively

Table 3
Thirty-eight *E. coil* DNA promoter sequences and their observed and predicted strengths

No.	Name	Sequences	Obsd	$Cald_{PLS}$	$Cald_{SVM}$
1	D/E20	ACTGCAAAAATAGTTTGACACCCTAGCCGATAGGCTTTAAGATGTACCCAGTTCGATGAGAGCGATAA	1.748	1.810	1.743
2	H207	TTAAAAAATTCATTTGCTAAACGCTTCAAATTCTCGTATAATATACTTCATAAATTGATAAACAAAAA	1.740	1.739	1.735
3	G25	GAAAAATAAAATTCTTGATAAAATTTTCCAATACTATTATAATATTGTTATTAAAGAGGAGAAATTAA	1.278	1.307	1.273
4	A1	ATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCATCGAGAGGGACACGGCGA	1.881	1.881	1.876
5	A2	GAAAAACAGGTATTGACAACATGAAGTAACATGCAGTAAGATACAAATGCCTAGGTAACACTAGCAGC	1.301	1.263	1.306
6	L	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAGGACGCACTGAC	1.568	1.469	1.563
7	CON	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACCATAAGGAGGTGGATCCGGC	0.602	0.645	0.607
8	LAC	AGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTGGTATGTTGTGTGGAATTGTGAGCGGATAACAA	0.756	0.706	0.751
9	LAC/UV5	AGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTGGTATAATGTGTGGAATTGTGAGCGGATAACAA	0.518	0.555	0.523
10	N25/03	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCAATTGTGAGCGGATAACAA	0.903	0.839	0.898
11	N25/ANTI	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATCCGGAATCCTCTTCCCG	0.432	0.428	0.437
12	N25/LAC	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCAAATTGTGAGCGGATAACA	0.903	0.864	0.908
13	CON/N25	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATAAATTTGAGAGAGGAGT	1.398	1.294	1.309
14	CON/PEX	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATAAAGGGTCGAGAGGAGT	1.204	1.156	1.209
15	CON/ANTI	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATCCGGAATCCTCTTCCCG	0.255	0.340	0.292
16	CON/D/E20	TTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACCCAGTTCGATGAGAGCGATAA	1.114	1.124	1.119
17	CON/TRP	TTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACGCAAGTTCACGTAAAAAGGGT	0.903	0.895	0.908
18	L-8A	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATAATGAGCACATCAGCAGGACGCACTGAC	1.672	1.462	1.573
19	L/CON	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATCAGCAGGACGCACTGAC	1.146	1.415	1.507
20	L/N25	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATAAATTTGAGAGAGGAGT	1.813	1.857	1.818
21	L/CON/N25	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATAAATTTGAGAGAGGAGT	1.813	1.803	1.783
22	N25/O5	GGATAACAATTTAGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.173	1.294	1.178
23	N25/USR	GGCTAAAAAACACGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.491	1.505	1.486
24	CON/O5	GGATAACAATTTAGTTGACATTTTTAAGCTTGGCGGTTATAATGTTACCATAAGGAGGTGGGAATTCC	1.173	1.094	1.178
25	L/N25DSR	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATAAATTTGAGAGAGGAGT	1.813	1.857	1.818
26	L/CON/N25DSR	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATAAATTTGAGAGAGGAGT	1.778	1.803	1.783
27	LS1	TCCGTCTCGACGGGTTGACACAAAAGCCACAAGGGGTTATAATGAGCACATAAACTTGAGAGAGGAAT	2.143	2.134	2.138
28	LS2	TGCGTATAGACAGTTTGACACAAAAGCCACAAGGTGTTATAATGAGCACATAAATTTGAGAGAGGAAT	2.217	2.180	2.212
29	N25	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.477	1.382	1.382
30	J5	TATAAAAACCGTTATTGACACAGGTGGAAATTTAGAATATACTGTTAGTAAACCTAATGGATCGACCT	0.954	0.936	1.096
31	A3	TGAAACAAAACGGTTGACAACATGAAGTAAACACGGTACGATGTACCACATGAAACGACAGTGAGTCA	1.342	1.441	1.415
32	TACL	TTCTGAAATGAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAA	1.230	1.182	1.142
33	N25/PEX	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAAGGGTCGAGAAGAGT	1.176	1.195	1.160
34	CON/03	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCAATTGTGAGCGGATAACAA	0.903	0.751	0.745
35	L-12T	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATACTGAGCACATCAGCAGGACGCACTGAC	1.398	1.422	1.4971
36	N25/O4	CATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCATAAATTTGAGAGAGGAGT	1.246	1.255	1.247
37	N25/AUSR	GGCTGTGCGGCACGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.301	1.380	1.339
38	L/N25USR	CATAAAAAATTTATTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAGGACGCACTGAC	1.763	1.530	1.725

studied [4,7,8,29–34]. Thirty-eight promoter sequences used in this study are presented in Table 3 [4,7,8,31,33]. These sequences include both original promoters of *E. coli* and recomposed promoters which are derived from promoters infected by λ bacteriophage. They are 68 bp DNA fragment from transcription start site position −49 to position +19, and include RNA polymerase site for Sextama (−35 region), double hilex liquated site for Pribnow (−10 region) and other linked and assistant sequence fragments. Few of the promoters have been consistently characterized with regard to their in vivo promoter strength. However, a system that allows this efficiency parameter to be determined relative to an internal standard has been developed and used to characterize promoters [8,35]. In this assay, the strength of the test promoter is expressed relative to that of the promoter for β-lactamase (P_{bla}) which is present on the same plasmid. Monitoring of the mRNA expressed from the promoter under study in relation to the standard, permits the relative promoter efficiency to be determined unbiased by translational effects or gene dosage. The logarithm of the promoter strength in P_{bla}-units is used in the present modeling because of the large variation in strength. We use SGBP scales to represent the 38 promoter sequences, and furthermore, construct QSAMs based on different statistics methods.

Promoter sequences are represented from position 5′ to position 3′. Accordingly, each base in the sequence is described by four SGBP variables according to varied base sitting position. Promoter sequences of different lengths result in different quantitative descriptors. So, a set of sequence varied in n positions can be characterized by the concatenation of $4 \times n$ vectors.

2.3. Variables selection

In QSAM studies, few variables as possible should be included in a model. On the one hand, not all the structural descriptors are relevant to biological activities. On the other hand, a QSAM containing few variables will be easy interpreted. So those redundant descriptors should be eliminated in order to promote its robustness and predictive capability especially when the number of variables is very large. Here variables selection is completed by genetic arithmetic-partial least square (GA-PLS) as a popular variables selection tool nowadays, which is a sophisticated hybrid approach that combines GA as a powerful optimization method with PLS as a robust statistical method for variables selection. GA mimics natural selection in nature. The principle of natural selection is that the species with a high fitness under some environmental conditions can prevail in next generation. The species with a low fitness cannot survive through selection. The best species may be reproduced by crossover together with random mutations of chromosomes in the surviving ones. In GA-PLS, the chromosome and its fitness in the species correspond to a set of variables and internal predictivity of the derived PLS model, respectively. GA includes five steps: (1) the initial population of chromosomes is created by setting all bits in each chromosome to a random value; (2) the fitness of each chromosome is evaluated by the internal predictivity of the PLS model derived from a binary bit pattern. The internal predictive performance of the model is expressed in terms of a cross-validation (CV) square of multiple correlation coefficient (R^2) value (hereafter, denoted by Q^2_{cv}) by the leave-one-out (LOO) procedure as follows:

$$Q^2_{cv} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (1)$$

where y_i and $\hat{y}_i$ represents the observed value and the predicted value of the dependent variable, respectively, $\bar{y}$ the mean observed value of the dependent variable and n is the number of samples; (3) the chromosome with the least number of variables and the highest fitness is marked as an informative chromosome; (4) GA manipulation including crossover, mutation, and replication is carried out; (5) the cycles of above four steps (from step 2 to 4) are repeated until the optimal chromosomes are achieved. A detail description about GA-PLS can be seen in Refs. [36,37].

2.4. Internal test and external validation

An excellent QSAM model should have both favorable estimation ability for any internal sample and outstanding predictive ability for any external sample. The most usual method to prove a model to have excellent internal predictive ability is a CV method. In the present work, LOOCV for internal validation criteria is used. Predictive performance of the model can be assessed by the prediction values of Q^2_{cv}. External validation can only be achieved by splitting the total data set into a training set for establishing the QSAMs and a test set for evaluating model performance. The external prediction power of QSAMs can be evaluated by Q^2_{ext} as follows [38]:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_{tr})^2} \quad (2)$$

where y_i and $\hat{y}_i$ are the observed and predicted values over the test set of the dependent variable, respectively, $\bar{y}_{tr}$ the mean value of the dependent variable for the training set and n is the number of test set.

There are many approaches for creating training and test sets including straightforward random selection, clustering techniques, and D-optimal, etc. Some results showed that external validation by D-optimal is often superior to that by others [38,39]. Thus, D-optimal splitting methodologies for the purpose of external validation are adopted here. The arithmetic select the samples that maximize the $|X'X|$ determinant, where X is the information matrix of independent variables (descriptors), or of independent and dependent variables (descriptors and response). The points maximizing the $|X'X|$ determinant are spanned across the whole area occupied by representative points. These points constitute training set and the points not selected are used as test set. This arithmetic guarantees a well-balanced structural diversity and representativity of the entire data space (descriptors plus response). Principles for D-optimal can be seen in Ref. [40]. Twenty-eight

DNA sequences were treated as training set which was utilized to construct QSAMs and the remaining 10 sequences were regarded as test set in order to validate the predictive power of the models.

2.5. PLS modeling

PLS is mainly used for modeling linear regression between multi-dependent variables and multi-independent variables. It has many advantages which ordinary multiple linear regression does not have. For instance, it can avoid harmful effects in modeling duo to multicollinearity, and is particularly fit for regressing when the number observation is less than the number of variables, etc. In addition, PLS regression combines basic functions of regressing model, PCA and canonical correlation analysis [41–44].

PLS also has the desirable property that the precision of the model parameters is improved with the increasing number of relevant variables and observations. The PLS regression algorithm consists of outer relations (X and Y block individually) and an inner relation linking both blocks:

$$x_{ik} = \sum_{a=1}^{A} t_{ia} p_{ak} + e_{ik} \tag{3}$$

$$x_{im} = \sum_{a=1}^{A} u_{ia} c_{am} + g_{im} \tag{4}$$

The t and u latent variables are correlated through the inner relation given below which leads to the estimation of the y from the x.

$$\hat{u} = bt \tag{5}$$

2.6. SVM modeling

SVM has successfully been applied to recognize face, speech, handful words, rubbish mails, protein space structure and gene and so on [9,45–50]. The application of kernel function, i.e., $K(x, x_i) = \Phi(x)\Phi(x_i)$, is a crucial technology for SVM. It can effectively deal with many problems such as dimensional puzzledom, calculation complexity, etc. At present, familiar kernel functions are mainly: linear kernel function: $K(x, x_i) = xx_i$; polynomial kernel function: $K(x, x_i) = (\alpha_1 x x_i + \alpha_2)^{\mathrm{p}}$; radial basis kernel function: $K(x, x_i) = \exp(-\gamma||x - x_i||^2)$; sigmoid kernel function: $K(x, x_i) = \tan h\ (\alpha_1 x x_i + \alpha_2)$. In this study, SVM with radial basis function (RBF) kernel function: $K(x, x_i) = \exp(-\gamma||x - x_i||^2)$ is used to develop QSAMs. Detailed information about kernel function is given in Refs. [9,11,47,50].

It is important to select parameters for SVM regression in the modeling. Because no standard rules are followed, we tentatively use response surface methodology [51] which can provide highly efficient design of experiments for the optimal process settings to achieve peak performance to select the parameters based on Q^2_{ext} of exterior validation in this study. If Q^2_{ext} of one model is equal to that of another model, parameters are further confirmed according to Q^2_{cv}.

2.7. Software used

One thousand two hundred and nine descriptor parameters of five bases are calculated by Dragon program (free download at http://www.disat.unimib.it/chm/), Chem3D 2005, and GITMHDV program which is programmed using True basic language by our group. GA-PLS, D-optimal, PLS and SVM are carried out by Matlab 7.0. Response surface methodology is accomplished by Version 7 of Design-Expert software.

3. Results and discussion

3.1. QSAM analysis from SGBP-GA-PLS

Every base in 38 pieces of promoter sequences was identified by four SGBP descriptors. Every promoter sequence was represented by 272 (68×4) SGBP descriptors. GA-PLS was subjected to the data set to eliminate the autocorrelation vectors and to optimize their descriptive power. The values of empirical parameters influencing the performance of GA-PLS were determined by experience from the series of GA-PLS studies. Parameters were set as follows: the number of populations was 200, the maximum number of generations was 200, the generation gap was 0.8, the crossover frequency was 0.5, the mutation rate was 0.005, and the fitting function was Q^2_{cv}. An optimal model was obtained from all trained models (No. 10th in Table 4). There were 93 SGBP descriptors selected in the model. A 80.6% of variable dispersion was explained by three principal components through the LOOCV procedure. Thirty-eight samples were divided into 28 training samples (1–28 in Table 3) and 10 testing samples (29–38 in Table 3) using D-optimal. 97.4% of variance was captured using three principal components through PLS based on 28 training samples. Root-mean-square error of estimation was

Table 4
Features for variable selected and results based on GA-PLS

No.	The number of variables	The number of principal components	Q^2_{cv}
1	93	2	0.775
2	99	2	0.752
3	101	2	0.755
4	102	3	0.736
5	98	3	0.790
6	94	3	0.723
7	98	3	0.745
8	97	3	0.766
9	104	3	0.753
10	93	3	0.806
11	101	4	0.801
12	92	4	0.791
13	98	4	0.788
14	94	5	0.782
15	101	5	0.773
Original model	272	2	0.758

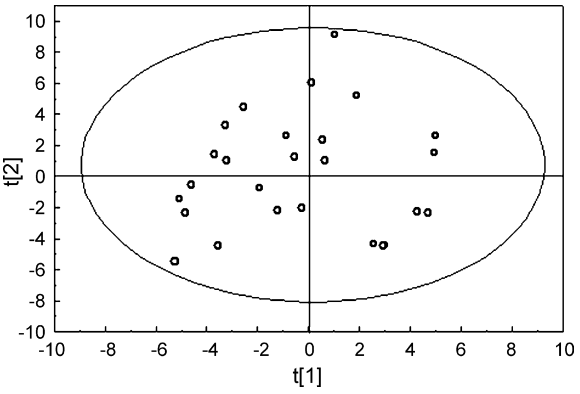

Fig. 1. Score for PLS.

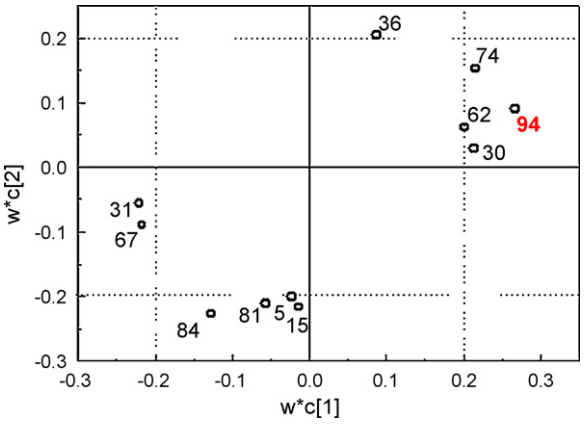

Fig. 3. Loading for PLS.

RMS_{cu} = 0.089. 85.9% of variance was cumulatively explained by the LOOCV procedure.

Score for PLS (Fig. 1) shows that their high dimensional property of the independent variables may be similar to each other when the two samples relatively approach. From it, the samples dots are evenly distributed and all the samples are in confident intervals of *Hotelling* T^2 ellipse. The distance to the PLS model in the *X* space is described in Fig. 2 in order to investigate efficiency on recombination. It can be seen the normalized distance to *X* for the samples of 2, 3, 4, 5, 8, 16 and 17 overrun the critical value of 1.28 (significance level = 5%), which shows that the results of seven samples characterized by three principal components are worse than average results of others. It can be seen from loading for PLS (Fig. 3) that loadings of v_{30}, v_{62} and v_{74} are higher (>0.200) than those of other variables in the first principal component. And they are positively correlated with *Y* (strengths). These corresponding variables are unweighted signal 29 of 3D-MoRSE, electronic energy, lag 6 (weighted by atomic van der Waals volumes) of Moran autocorrelation, and Moriguchi octanol–water partition coefficient (log *P*). Loadings of v_{31} and v_{67}, i.e., the third principal component of base position of −42 and −33, are less than −0.200 shows they are negatively correlation with *Y*. It can be known from analysis above those variables are representation of constitutional features on mean atomic polarizability (scaled on carbon atom), average molecular weight, and aromatic ratio. In loadings of the second principal component, v_{36}, i.e., the fourth principal component of base position of −41, shows the greatest contribution (0.200) to it, especially the corresponding variables are the $MHDV_7$ and unweighted signal 22 of 3D-MoRSE. On the contrary, v_5, v_{15}, v_{81} and v_{84} are negatively correlated with *Y* (less than −0.200). Those corresponding variables include the first principal component of base position of −48 which includes lag 2 (weighted by atomic polarizabilities) of Moran autocorrelation, Torsion energy, and sum of atomic van der Waals volumes (scaled on carbon atom), etc., the third principal component of position of −46 which reflects mean atomic polarizability (scaled on carbon atom), average molecular weight, and aromatic ratio, etc., the first principal component of base position of −29 which covers lag 2 (weighted by atomic polarizabilities) of Moran autocorrelation, Torsion energy, and sum of atomic van der Waals volumes (scaled on carbon atom), and the fourth principal component of base position of −29 which refers to those variables on signal unweighted 21 of 3D-MoRSE and lag 4 (weighted by atomic polarizabilities) of Moran autocorrelation, etc. Therefore, it can be estimated these bases will produce more influence on strengths, which brings a hopeful idea for designing strong promoters and predicting their strengths. Strong sequence can also be obtained by changing these bases. Fig. 4 indicates that the calculated strengths are preferably close to observed strengths, which shows the model has favorable simulative for internal samples and predictive ability for external samples.

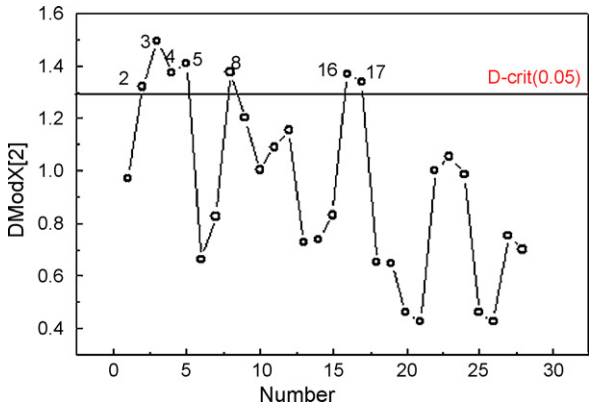

Fig. 2. Distance in *X* space modeled by PLS.

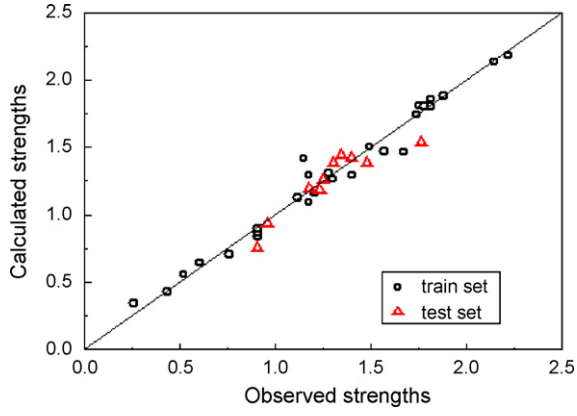

Fig. 4. Regression between observed and calculated strengths by PLS.

3.2. QSAM analysis from SGBP-GA-SVM

As for SVM regression with RBF kernel, the important difference from classical RBF is that each basic function has a corresponding support vector. The support vectors and the output weights are all determined by arithmetic. Parameters of SVM involve regularized coefficient C, ε of ε-insensitive loss function, and the kernel style K and its corresponding variables. C is a regulative parameter which controls the tradeoff maximizing the margin and minimizing the training error between maximal interval and minimum training error. Insufficient stress will be placed on fitting the training data if C is too small. Reversely, fitting will be beyond if C is too large. C should becomingly be given relatively large value in order to make the training process stabilization. The optimal value for ε depends on the type of noise present in the usually unknown data. Even if a great deal of knowledge for the noise is available to select an optimal value for ε, there is the practical consideration of the number of resulting support vectors. ε-insensitivity prevents the entire training set meeting boundary conditions and so allows for sparse possibility in the dual formulation's solution. The γ controls amplitude of Gaussian function, that is, generality capability of SVM [52]. So, choosing appropriate value for ε is critical from theory. Because unimpressive modeling and predictive ability of models by original 272 variables as input of SVM were obtained, so 93 variables selected by GA-PLS were as the input for SVM, also availability of GA-PLS for SVM would be probed into.

Single factor was alternately used to determine the bound of all parameters. At first, $\varepsilon = 0.050$ and $\gamma = 0.0078$ were rooted, then models were trained at C = 0.5, 1.0, 1.5, 2.0, 5.0, 8.0, 10.0, 15.0, 20.0, 50.0, 100.0, 300.0, 500.0, 700.0 and 900.0, respectively. The C versus Q^2_{ext} by SVM is indicated in Fig. 5. Although the relatively larger value of C can make learning process stable, Q^2_{ext} keeps invariableness (0.694) with improvement of C from 20 to 900. There is the highest Q^2_{ext} = 0.834 when C is 1.5. Then C was determined at 1.0–8.0. Next, C and γ were fixed with $C = 2.0$ and $\gamma = 0.0078$, then ε of 0.001, 0.005, 0.010, 0.020, 0.030, 0.040, 0.050, 0.060, 0.070, 0.085 and 0.095 were considered, respectively. The ε versus Q^2_{ext} by SVM is given in Fig. 6. It can be seen from it that Q^2_{ext} keeps almost indistinctive change from 0.001 to 0.050. Then

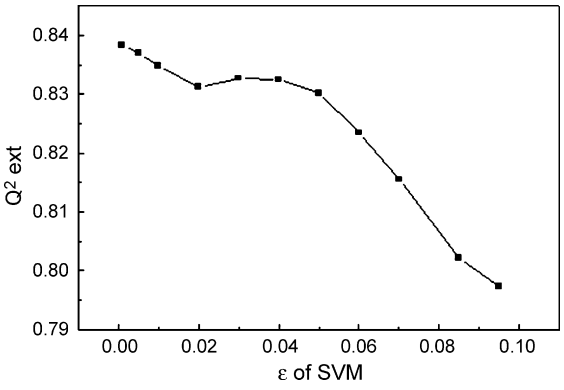

Fig. 6. The ε vs. Q^2_{ext} by SVM ($C = 2.0$, $\gamma = 0.0078$).

$\varepsilon = 0.005$–0.085 was selected in order to be convenient for operation. When $C = 2.0$ and $\varepsilon = 0.050$, then $\gamma = 0.0024$, 0.0027, 0.0032, 0.0038, 0.0045, 0.0055, 0.0069, 0.0089, 0.0118 and 0.0165 were taken into account, respectively. The curve of γ versus Q^2_{ext} shows (Fig. 7) the relatively peak Q^2_{ext} is 0.843 when γ is 0.0055. Then models were trained at γ from 0.0027 and 0.0118. Through the above process, the conditions with C = 1.0–8.0, ε = 0.005–0.085, and γ = 0.0027–0.0118 were given. Successively twenty predictive experiments were designed to optimize C, ε and γ using a response surface methodology based on central composite design (Six center points are designed) (Table 5). Results of exterior training showed there was the highest Q^2_{ext} of 0.858 when $C = 8.0$, $\varepsilon = 0.005$, and $\gamma = 0.0027$ (the 20th model).

The quadratic equation about C, ε, γ and Q^2_{ext} determined by response surface methodology was:

$$Q^2_{ext} = 0.481 + 0.054C + 2.865\varepsilon + 36.822\gamma - 0.177C\varepsilon - 3.787C\gamma - 25.068\varepsilon\gamma - 0.001C^2 - 23.710\varepsilon^2 - 965.408\gamma^2 \quad M1$$

A Fisher F value was 6.29. Probability $> F$ was 0.0041 shows that the M1 was significant at 0.05 level. we used M1 to optimize the parameters in order to acquire the possiblely maximal Q^2_{ext} (Table 6). The relatively highest Q^2_{ext} was 0.853 (No. 11th), which was less than 0.858 derived from the 20th model in Table 5. Therefore, the parameters of SVM which

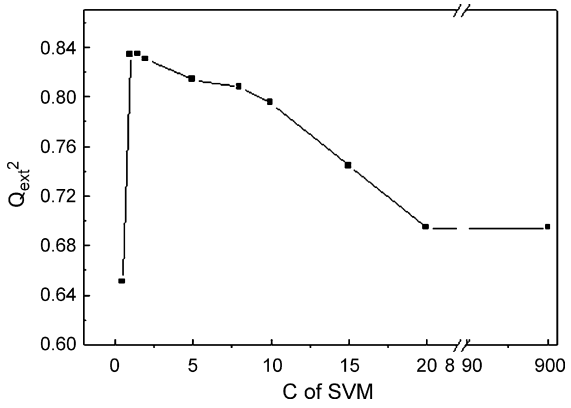

Fig. 5. The C vs. Q^2_{ext} by SVM ($\varepsilon = 0.050$, $\gamma = 0.0078$).

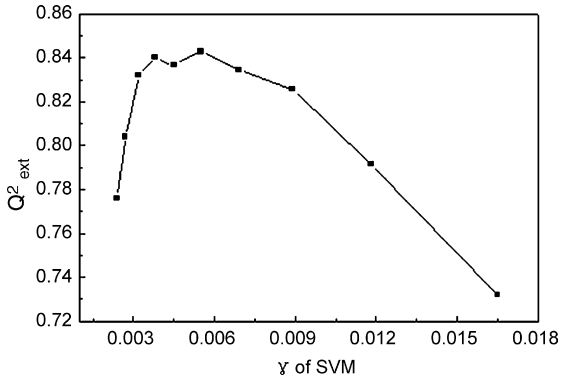

Fig. 7. The γ vs. Q^2_{ext} by SVM ($C = 2.0$, $\varepsilon = 0.050$).

Table 5
Features and results of response surface methodology design for SVM

No.	Factor 1 (C)	Factor 2 (ε)	Factor 3 (γ)	Q^2_{cv}	Q^2_{ext}
1	8.0	0.085	0.0027	0.821	0.781
2	1.0	0.085	0.0118	0.839	0.817
3	4.5	0.045	0.0073	0.851	0.826
4	4.5	0.005	0.0073	0.849	0.826
5	8.0	0.005	0.0118	0.850	0.789
6	4.5	0.085	0.0073	0.832	0.772
7	4.5	0.045	0.0073	0.844	0.826
8	4.5	0.045	0.0027	0.858	0.849
9	1.0	0.005	0.0118	0.832	0.814
10	8.0	0.085	0.0118	0.827	0.746
11	1.0	0.085	0.0027	0.701	0.663
12	1.0	0.045	0.0073	0.851	0.824
13	4.5	0.045	0.0118	0.828	0.785
14	4.5	0.045	0.0073	0.847	0.826
15	1.0	0.005	0.0027	0.621	0.589
16	8.0	0.045	0.0073	0.832	0.822
17	4.5	0.045	0.0073	0.835	0.826
18	4.5	0.045	0.0073	0.859	0.826
19	4.5	0.045	0.0073	0.861	0.826
20	8.0	0.005	0.0027	0.883	0.858

were determined according to the 20th model in Table 5 were as follows: C was 8.0, ε was 0.005, and kernel function was $K(x, x_i) = \exp(-0.0027||x - x_i||)^2$. Square of multiple correlation coefficient for modeling $(R^2_{cu}) = 0.980$, root-mean-square error for modeling $(RMS_{cu}) = 0.073$, $Q^2_{cv} = 0.883$, square of multiple correlation coefficient for external validation

Table 6
Features and results of based on M1 by response surface methodology

No.	C	ε	γ	Q^2_{cv}	Q^2_{ext}
1	7.9	0.0119	0.0035	0.855	0.850
2	7.9	0.0501	0.0030	0.849	0.844
3	7.8	0.0290	0.0062	0.859	0.844
4	7.9	0.0406	0.0049	0.851	0.841
5	7.8	0.0235	0.0061	0.852	0.847
6	7.9	0.0283	0.0058	0.850	0.846
7	8.0	0.0362	0.0053	0.851	0.843
8	8.0	0.0399	0.0044	0.857	0.842
9	8.0	0.0191	0.0045	0.855	0.849
10	7.8	0.0270	0.0052	0.849	0.848
11	7.9	0.0093	0.0031	0.869	0.853
12	7.9	0.0503	0.0028	0.852	0.848
13	7.8	0.0236	0.0063	0.850	0.846
14	7.8	0.0412	0.0056	0.848	0.838
15	7.5	0.0203	0.0047	0.850	0.848
16	7.7	0.0317	0.0062	0.849	0.843
17	7.8	0.0375	0.0059	0.850	0.840
18	7.8	0.0339	0.0063	0.851	0.840
19	7.4	0.0203	0.0040	0.854	0.846
20	7.4	0.0240	0.0051	0.853	0.849
21	2.1	0.0381	0.0118	0.827	0.798
22	2.1	0.0381	0.0118	0.831	0.821
23	1.0	0.0503	0.0118	0.841	0.823
24	1.0	0.0508	0.0118	0.830	0.822
25	1.0	0.0472	0.0118	0.836	0.823
26	1.0	0.0503	0.0118	0.838	0.823
27	1.0	0.0508	0.0118	0.829	0.822
28	1.0	0.0472	0.0118	0.830	0.823

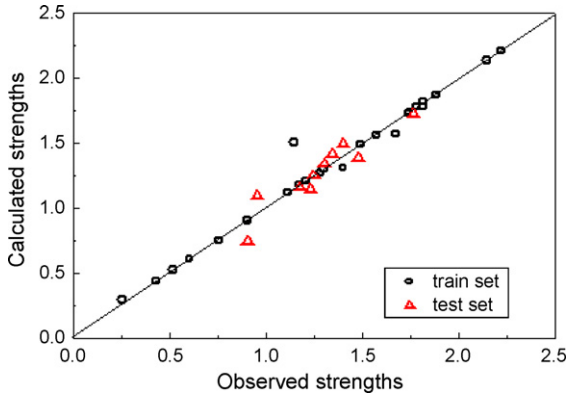

Fig. 8. Regression between observed and calculated activities by SVM.

$(R^2_{ext}) = 0.855$, $Q^2_{ext} = 0.858$, and root-mean-square error for external validation $(RMS_{ext}) = 0.089$ were obtained, respectively. Regression between observed and calcualted activities by SVM (Fig. 8) indicates it can provide a satisfying result to predict promoter strengths.

SVM, as a new statistical learning algorithm, cherishes many unique features. Particularly, it can handle large datasets and exhibits remarkable resistance to over-fitting. SVMs condense information in the training set by using a very small number of samples with support vectors (SVs) to provide sparse representation. It is believed that these SVs contain all the information needed for modeling. In most cases the number of SVs is smaller than the total number of training samples. In our method, the ratio of SVs to training samples is 100.0%, which means no any training samples could be safely removed. We presumed that a relatively small quantity of training samples may account for reasons why 100.0% training samples have been treated as SVs.

3.3. Comparison among different QSAMs

Modeling statistics results from PLS and SVM are gathered in Table 7. It can be seen from it satisfying results are acquired using two modeling methods. By comparison, exterior predictive capability as well as modeling and interior predictive power of SVM are obviously better than those of PLS. It can be concluded that SGBP-GA-SVM is appropriate

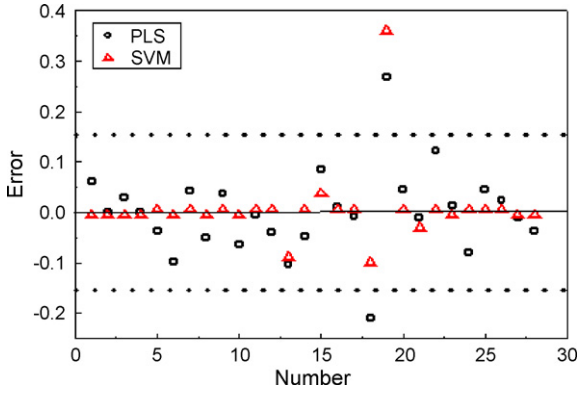

Fig. 9. Predictive errors for samples of train set.

Table 7
Comparison for different QSAMs

Model	Data size	A[a]	C	ε	γ	R^2_{cu}	RMS_{cu}	Q^2_{cv}	R^2_{ext}	Q^2_{ext}	RMS_{ext}
PLS	28/10[b]	2	–	–	–	0.974	0.089	0.859	0.808	0.812	0.103
SVM	28/10	–	8.0	0.005	0.0027	0.980	0.073	0.883	0.855	0.858	0.089

[a] The number of principal components.
[b] The two numbers separated by slashes denote the numbers of compounds in the train, test sets, respectively.

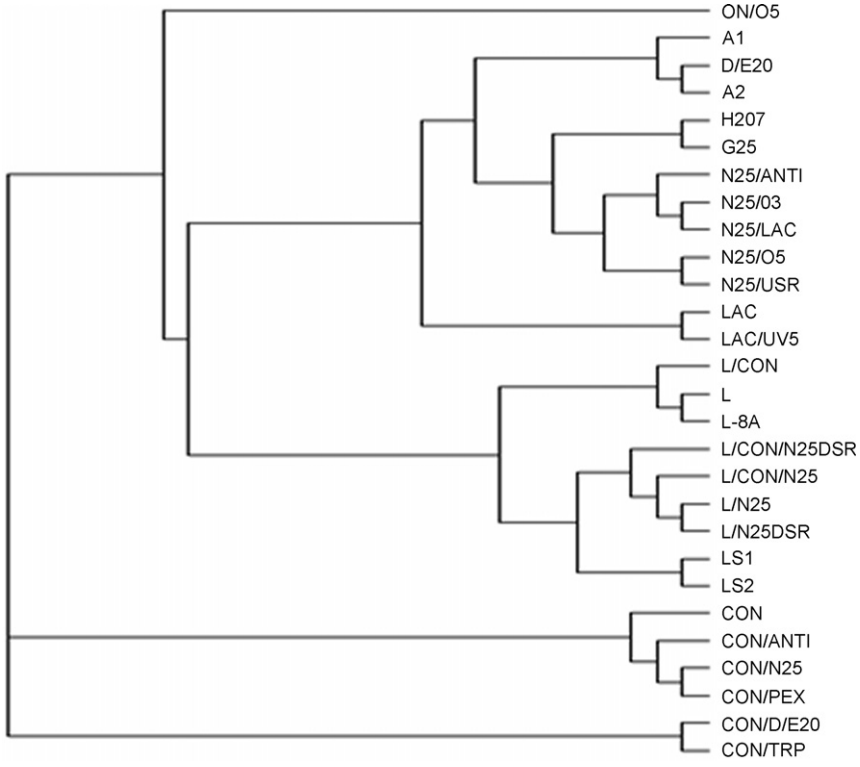

Fig. 10. Phylogenetic tree for train set.

for constructing favorable QSAM for DNA promoter strengths and sequences.

Plot of predictive errors for samples of train set (Fig. 9) show that absolute values of errors for L-8A (see the 18th sample in Table 3) and the L/CON sample (see the 19th sample in Table 3) predicted by PLS, L/CON predicted by SVM are more than 0.150. Why predictive errors for two samples are greater than those for other samples? Thereinafter, homologous analysis may answer this question. It can be concluded from phylogenetic tree for all samples of training set (Fig. 10) there is the largest similarity index of 98.5% between L-8A and L/CON. Base of position −12 of L-8A, that is G, is substituted by T (in L/CON sequence) (change of the number of rings is called as transversion). Furthermore, we find there is similarity index of 98.5% between L-8A and L. Because transversion from C (L) to A (L-8A) of position −8 base results in the change of the number of rings and the possible conformation, observed strength of L-8A (1.672) is more than that of L (1.568). We presume that PLS model having not preferably identified the action came into being the relatively large error. On the contrary, SVM model can identify this action, so relatively little error was obtained. L/CON shows similarity of 97.1% with L. Although it does not cause the change of the number of rings, substitute of position −12 base of L, that is G, by T (in L/CON sequence) and position −8 base of L, that is C, by A (in L/CON sequence) may induce the change of conformation. As a result, observed strength of L/CON is relatively little. It is speculated

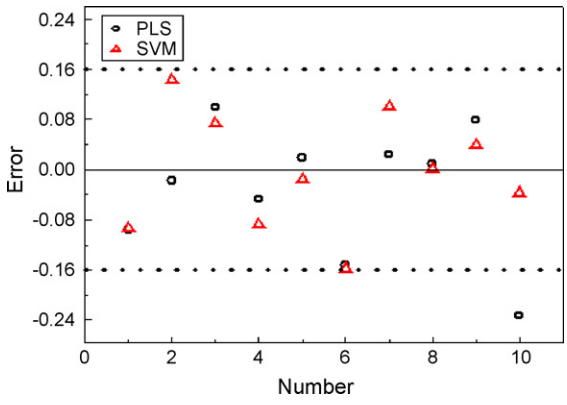

Fig. 11. Predictive errors for samples of test set.

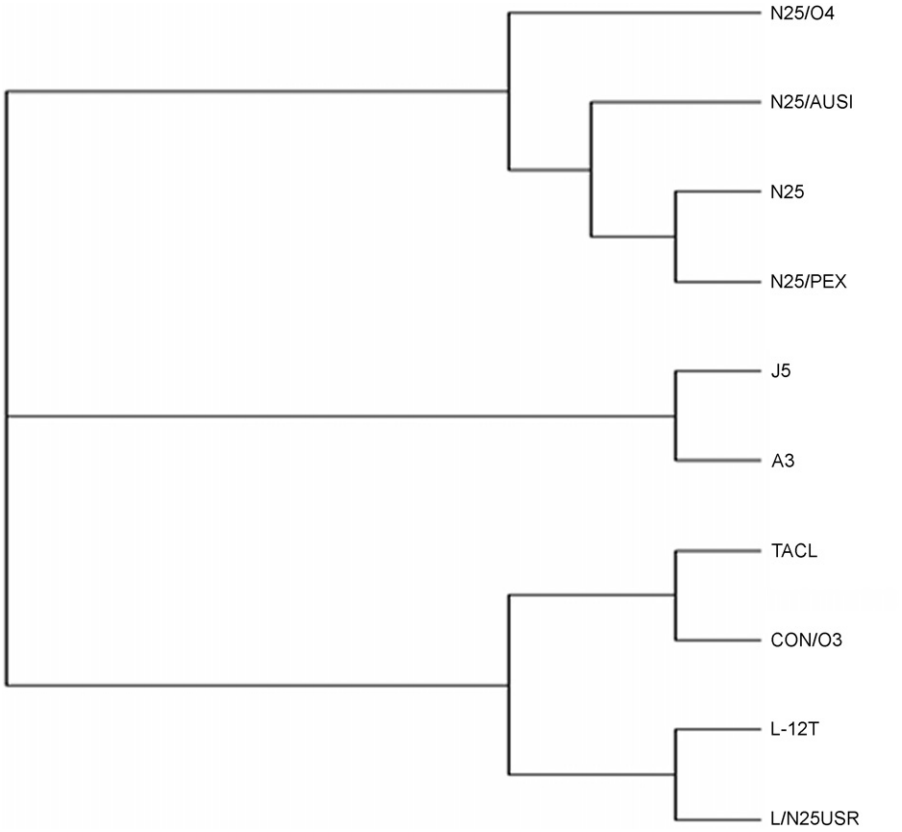

Fig. 12. Phylogenetic tree for test set.

that error for L/CON by PLS and SVM model was relatively larger due to extreme acuteness to this action. The curve of predictive errors for samples of test set is depicted in Fig. 11. From it, it can be seen that absolute value for errors of the L/N25USR (see the 10th sample in Table 3) predicted by PLS goes beyond 0.150. So, phylogenetic tree for all samples of test set is described in Fig. 12. It can be seen that there is the largest similarity index of 98.1%. By comparison, the base of position −12 in L-12T is T, but the base of position −12 in L/N25USR is G. Because change of the number of rings due to the transversion of T substituted by G results in the diversification of conformation of L/N25USR, observed strength of L/N25USR is obviously greater than that of L-12T. We suppose that PLS model is insensitive to this action caused relatively greater error.

4. Conclusions and expectation

SGBP as a novel technique has been advocated to character *E. coli* promoter sequences. Results indicate structural information of 38 promoter sequences can be favorably represented by SGBP with many advantages such as adequate information and easy operation. It can further be used to represent DNA and RNA sequences. Favorable QSAMs about promoter sequences and strengths are constructed using PLS and SVM. Through investigation on QSAM by PLS, it is forecasted that properties of base of position −42, −34, −31, −33, −41, −46 and −29 have more influence on strengths, which brings a promising idea for designing sequence and predicting strength of promoters. Strong sequence can also be obtained by changing these bases. Because no standard rules are followed to determine parameters of SVM. Here the parameters are initialized by a response surface methodology by which the results are more creditable than those by a random or an experience. By comparison with results of PLS, better ones are obtained using SVM, which can avoid some shortcomings such as over training, low generality, excessive dependence on experience, instability, and falling into local minimum. QSAMs from this study can offer a way to design DNA sequences and to predict their functions, especially, applications on QSAMs based on SGBP-GA-SVM can be further popularized. Although excellent performance can be obtained by SVM which can be used to deal with what we wanted to tackle, because many parameters has to be selected before QSAM is built, especially, the selection of kernel function, principles and application skills for SVM must be deeply investigated in order to acquire good results and successful applications. For example, some researchers have dealt with some puzzledom based on kernel PCA and kernel PLS methods which can offset some shortcomings of SVM and give a possible key to dealing with them [53,54]. It is believed that new era of machine learning will be inaugurated with the development of SVM deep research. In this study, 38 pieces of promoter sequences with same lengths are investigated. However, as for DNA sequences with different lengths, auto cross covariances (ACCs) can be used to translate them into a

uniform set of independent variables which can also be interpreted, i.e., used to identify important features in the original sequences [55,56]. SGBP and SVM modeling technique for sequences of DNA and RNA are deeply in progress.

Acknowledgments

The work was supported by the State Chunhui Project Fund (No. 98-3-8), Fok Ying-Tung Educational Foundation (No. 98-7-6) and Chongqing University Innovation Fund (No. 03-5-6).

References

[1] H.M. Müller, S.E. Koonin, Vector space classification of DNA sequences, J. Theor. Biol. 223 (2003) 161–169.
[2] T. Biro, A. Czirok, T. Vicsek, A. Major, Application of vector space techniques to DNA, Fractals 6 (1998) 205–210.
[3] M. van Heel, A new family of powerful multivariate statistical sequence analysis techniques, J. Mol. Biol. 220 (1991) 877–887.
[4] B. Demeler, G.W. Zhou, Neural network optimization for *E. coli* promoter prediction, Nucl. Acids Res. 19 (1991) 1593–1599.
[5] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.
[6] P.H.A. Sneath, Relations between chemical structure and biological activity in peptides, J. Theor. Biol. 12 (1966) 157–195.
[7] M.E. Mulligan, D.K. Hawley, R. Entriken, R.W. McClure, *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity, Nucl. Acids Res. 12 (1984) 789–800.
[8] J. Jonsson, T. Norberg, L. Carlsson, C. Gustafsson, S. Wold, Quantitative sequence-activity model (QSAM)-tools for sequence design, Nucl. Acids Res. 20 (1993) 733–739.
[9] C. Cortes, V. Vapnik, Support vector networks, Mach. Learn. 20 (1995) 273–293.
[10] P.A. Flach, On the state of the art in machine learning: a personal review, Artif. Intell. 131 (2001) 199–222.
[11] A.V.D. Sάnchez, Advanced support vector machines and kernel methods, Neurocomputing 55 (2003) 5–20.
[12] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley–VCH, Weinheim, 2000.
[13] P. Ertl, B. Rohde, P. Selzer, Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties, J. Med. Chem. 43 (2000) 3714–3717.
[14] S.S. Liu, C.Z. Cao, Z. Li, A novel molecular distance-edge (MDE, λ) vector and the normal boiling point of alkanes, J. Chem. Inf. Comput. Sci. 38 (1998) 387–394.
[15] S.S. Liu, C.S. Yin, S.X. Cai, Z.L. Li, A novel MHDV descriptor for dipeptide QSAR studies, J. Chinese Chem. Soc. 48 (2001) 253–260.
[16] J. Gilvez, R. Garcia, M.T. Salabert, R. Soler, Charge indexes: new topological descriptors, J. Chem. Inf. Comput. Sci. 34 (1994) 520–525.
[17] G. Rucker, C. Rucker, Counts of all walks as atomic and molecular descriptors, J. Chem. Inf. Comput. Sci. 33 (1993) 683–695.
[18] C. Rucker, G. Rucker, Mathematical relation between extended connectivity and eigenvector coefficients, J. Chem. Inf. Comput. Sci. 34 (1994) 534–538.
[19] A.T. Balaban, D. Ciubotariu, M. Medeleanu, Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors, J. Chem. Inf. Comput. Sci. 31 (1991) 517–523.
[20] M.V. Diudea, D. Horvath, A. Graovac, Molecular topology. 15. 3D distance matrices and related topological indices, J. Chem. Inf. Comput. Sci. 35 (1995) 129–135.
[21] A.T. Balaban, From chemical topology to 3D geometry, J. Chem. Inf. Comput. Sci. 37 (1997) 645–650.
[22] M. Randic, A.F. Kleiner, L.M. DeAlba, Distance/distance matrices, J. Chem. Inf. Comput. Sci. 34 (1994) 277–286.
[23] J.H. Schuur, P. Selzer, J. Gasteiger, The coding of the three-dimensional structure of moleculesby molecular transforms and its application to structure-spectra correlations and studies of biological activity, J. Chem. Inf. Comput. Sci. 36 (1996) 334–344.
[24] J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, V. Steinhauer, Chemical information in 3D space, J. Chem. Inf. Comput. Sci. 36 (1996) 1030–1037.
[25] R. Todeschini, P. Gramatica, 3D-modelling and prediction by WHIM descriptors. Part 6. Application of WHIM descriptors in QSAR studies, Quant. Struct. -Act. Relat. 16 (1997) 113–119.
[26] V. Consonni, R. Todeschini, M. Pavan, Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors, J. Chem. Inf. Comput. Sci. 42 (2002) 682–692.
[27] D. Kim, I.-B. Lee, Process monitoring based on probabilistic PCA, Chemon. Intell. Lab. Syst. 67 (2003) 109–123.
[28] M.T. Marr, J.W. Roberts, Promoter recognition as measured by binding of polymerase to nontemplate strand oligonucleotide, Science 276 (1997) 1258–1260.
[29] M.E. Mulligan, J. Brosius, W.R. McClure, Characterization in vitro of the effect of spacer length on the activity of *Escherichia coli* RNA polymerase at the TAC promoter, J. Biol. Chem. 260 (1985) 3529–3538.
[30] M. Sandberg, M. Sjostrom, J. Jonsson, A multivariate characterization of tRNA nucleosides, J. Chemometr. 10 (1996) 493–508.
[31] M. Kobayashi, K. Nagata, A. Ishihama, Promoter selectivity of *Escherichia coli* RNA polymerase: effect of base substitutions in the promoter −35 region on promoter strength, Nucl. Acids Res. 18 (1990) 7367–7372.
[32] P.A. Szoke, T.L. Allen, P.L. deHaseth, Promoter recognition by *Escherichia coli* RNA polymerase: Effects of base substitutions in the −10 and −35 regions, Biochemistry 26 (1987) 6188–6194.
[33] D.G. Ayers, D.T. Auble, P.L. deHaseth, Promoter recognition by *Escherichia coli* RNA polymerase: role of the spacer DNA in functional complex formation, J. Mol. Biol. 207 (1989) 749–756.
[34] H. Kiryu, T. Oshima, K. Asai, Extracting relations between promoter sequences and their strengths from microarray data, Bioinformatics 21 (2005) 1062–1068.
[35] M. Lanzer, H. Bujard, Promoters largely determine the efficiency of repressor action, Proc. Natl. Acad. Sci. 85 (1988) 8973–8977.
[36] 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.
[37] K. Hasegawa, K. Funatsu, GA strategy for variable selection in QSAR studies: GAPLS and D-optimal designs for predictive QSAR model, J. Mol. Struct. (Theochem.) 425 (1998) 255–262.
[38] A. Golbraikh, A. Tropsha, Beware of q2! J. Mol. Graphics Mod. 20 (2002) 269–276.
[39] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.
[40] P.F. de Aguiar, B. Bourguignon, M.S. Khots, D.L. Massart, R. Phan-Than-Luu, Tutorial D-optimal Designs, Chemom. Intell. Lab. Syst. 30 (1995) 199–210.
[41] P. Gramatica, P. Pilutti, E. Papa, Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling, J. Chem. Inf. Comput. Sci. 44 (2004) 1794–1802.
[42] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[43] S. Wold, J. Trygg, A. Berglund, H. Antti, Some recent developments in PLS Modeling, Chemom. Intell. Lab. Syst. 58 (2001) 131–150.
[44] I.S. Helland, Some theoretical aspects of partial least squares regression, Chemom. Intell. Lab. Syst. 58 (2001) 97–107.

[45] S. Hua, Z. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, J. Mol. Biol. 308 (2001) 397–407.
[46] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Comput. Chem. 26 (2001) 5–14.
[47] S.J. Hong, S.M. Weiss, Advances in predictive models for data mining, Pattern Recogn. Lett. 22 (2001) 55–61.
[48] A.I. Belousov, S.A. Verzakov, J. von Frese, A flexible classification approach with optimal generalization performance: support vector machines, Chemom. Intell. Lab. Syst. 64 (2002) 15–25.
[49] Y.D. Cai, K.Y. Feng, Y.X. Li, K.C. Chou, Support vector machine for predicting α-turn types, Peptides 24 (2003) 629–630.
[50] J.B. Gaoa, S.R. Gunnb, C.J. Harrisb, SVM regression through variational methods and its sequential implementation, Neurocomputing 55 (2003) 151–167.
[51] R.H. Myers, D.C. Montgomery, Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Wiley, New York, 1995.
[52] 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.
[53] B. Scholkopf, J. Burges, A. Smola, Advances in Kernel Methods: Support Vector Machine, MIT Press, Cambridge, MA, 1999.
[54] V. Cherkassky, F. Mulier, Learning From Data: Concepts, Theory, and Methods, Wiley, New York, 1998.
[55] A. Nyström, P.M. Andersson, T. Lundstedt, Multivariate data analysis of topographically modified α-melanotropin analogues using auto and cross auto covariances (ACC), Quant. Struct. -Act. Relat. 19 (2000) 264–269.
[56] P.M. Andersson, M. Sjöstrom, T. Lundstedt, Preprocessing peptide sequences for multivariate sequence-property analysis, Chemom. Intell. Lab. Syst. 42 (1998) 41–50.