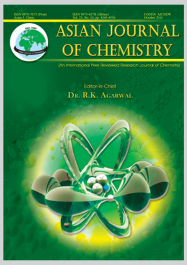
*Asian Journal of Chemistry; Vol. 24, No. 4 (2012), 1575-1578*

ASIAN JOURNAL OF CHEMISTRY

www.asianjournalofchemistry.co.in



# Establishing an Interpretability System for Support Vector Regression and Its Application in QSAR of Organophosphorus Insecticide

Li-Feng Wang[1], Xian-Sheng Tan[2], Lian-Yang Bai[2] and Zhe-Ming Yuan*

[1]Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Changsha, P.R. China
[2]Hunan Institute of Humanities, Science and Technology, Loudi, P.R. China

*Corresponding author: Fax: +86 731 84673775; Tel: +86 731 84635217, E-mail: zhmyuan@ sina.com

(*Received*: 15 March 2011; *Accepted*: 23 November 2011) AJC-10718

Aiming at the poor interpretability of support vector regression (SVR), a complete set of interpretability system for support vector regression was established based on F-test. The novel interpretability system includes the significance tests of model and single-factor importance, the single-factor effects and sensitivity analysis, the significance test of two factor interactions and so on. The analysis results of ternary dissymmetric organic phosphate insecticide preliminarily indicate that this new interpretability system is reasonable. Meanwhile, the quantitative structure-activity relationship (QSAR) model of insecticide based on support vector regression are superior to reference model in both fit and prediction, which further confirmed the outstanding regression performance of support vector regression.

**Key Words: Quantitative structure-activity relationship, F-test, Interpretability, Insecticide, Support vector regression.**

## INTRODUCTION

Insecticide plays an important role in insect pest management and the synthesis of new dosage form with high efficiency and low toxicity has been a research hotspot. Quantitative structure-activity relationship (QSAR) is one of the most widely used techniques in drug design and development. Establishing a quantitative structure-activity relationship model with high accuracy and interpretability is a key to molecular design[1]. Traditional modeling methods, such as multiple linear regression (MLR), stepwise linear regression (SLR), partial least square regression (PLS), quadratic polynomial regression (QPR), have good interpretability, but all of them base on empirical risk minimization and have poor analysis ability in the problem of high dimension, non linear and small samples[2-4]. Some researchers reported that artificial neural networks (ANN) had good nonlinear approximation capability, however it easily fell into local minimum and occurs overtraining or less training and its model structure was also difficult to determine[5,6].

Support vector machine (SVM) is a novel machine learning technique first presented by Vapnik 7 in 1995, which has drawn much attention in the fields of pattern classification and regression forecasting. Support vector machine bases on structural risk minimization instead of empirical risk minimization and it has the advantages of strong generalization ability, non-linear characteristics, avoiding over-fitting and dimension disaster, *etc*. Support vector machine contains support vector classification (SVC) and support vector regression (SVR), which has been wildely used in quantitative structure-activity relationship model. The high accuracy of support vector regression have been proved by some researchers, but the poor interpretability of support vector regression has not been resolved yet[7-9].

In order to improve the interpretability of support vector regression, this paper established a complete set of model detection and factor analysis system based on F test by referencing interpretability system of quadratic polynomial regression 4 model and then applied it for quantitative structure-activity relationship modeling of organophosphorus insecticide.

## EXPERIMENTAL

**Data set:** The data set used in this study was taken from the work of Jin *et al.*[10] and Zhou *et al.*[11] and shown in Table-1. This data set contains 7 factors of 22 organophosphorus insecticides, which include hydrophobic parameters $\pi$ and $\pi^2$, electric parameter $\sigma$, stereoscopic effect parameters L and $B_5$, zero-order and first-order connectivity indexes ${}^0X_v$ and ${}^1X_v$. Since the independent variables D were of different range, it was adjusted to comparable scale by standardization using the following equation:

$$D = \log \frac{a}{100 - a} + \log M \quad (1)$$

where, a represents insecticidal percentage when concentration of compounds is $10^{-4}$ and M denotes molecular weight.

TABLE-1
STRUCTURE PARAMETERS AND ACTIVITY INDEX OF COMPOUNDS

| No. | $\pi$ | $\pi^2$ | $\sigma$ | L | $B_5$ | $^0X_v$ | $^1X_v$ | D |
|---|---|---|---|---|---|---|---|---|
| 1 | -0.19 | 0.04 | 0.00 | 4.83 | 4.13 | 8.61 | 8.51 | 0.78 |
| 2 | 0.64 | 0.41 | 1.74 | 8.24 | 4.50 | 12.12 | 9.68 | 1.55 |
| 3 | 0.57 | 0.32 | 1.36 | 6.88 | 4.87 | 11.15 | 9.13 | 1.70 |
| 4 | 0.98 | 0.96 | 0.38 | 4.90 | 6.53 | 12.34 | 9.64 | 1.27 |
| 5 | 1.38 | 1.90 | 0.87 | 5.20 | 7.37 | 12.47 | 9.70 | 1.53 |
| 6 | 1.45 | 2.10 | 0.77 | 4.68 | 6.71 | 12.47 | 9.70 | 1.49 |
| 7 | 0.45 | 0.20 | 0.65 | 4.53 | 5.95 | 11.42 | 9.23 | 1.24 |
| 8 | 0.96 | 0.92 | 0.60 | 5.30 | 7.03 | 12.34 | 9.64 | 1.36 |
| 9 | 1.60 | 2.56 | 0.71 | 5.40 | 7.66 | 13.30 | 10.12 | 1.57 |
| 10 | 0.60 | 0.36 | 0.38 | 5.54 | 5.25 | 11.43 | 9.67 | 1.18 |
| 11 | 1.04 | 1.08 | -0.16 | 6.67 | 5.79 | 12.14 | 10.17 | 1.32 |
| 12 | 2.63 | 6.92 | 0.82 | 10.31 | 9.62 | 16.98 | 11.28 | 1.19 |
| 13 | -0.68 | 0.46 | 3.15 | 5.05 | 4.30 | 10.02 | 8.42 | 1.64 |
| 14 | 0.33 | 0.11 | 0.82 | 6.09 | 7.32 | 12.75 | 9.75 | 1.70 |
| 15 | 2.80 | 7.84 | 0.82 | 10.96 | 8.41 | 16.19 | 10.60 | 1.43 |
| 16 | -1.02 | 1.04 | 2.55 | 3.53 | 3.08 | 9.03 | 7.56 | 1.70 |
| 17 | -0.58 | 0.34 | 1.19 | 4.83 | 3.42 | 9.74 | 8.13 | 1.49 |
| 18 | -1.77 | 3.13 | 3.74 | 2.78 | 1.97 | 8.11 | 7.10 | 1.93 |
| 19 | -2.48 | 6.15 | 1.79 | 6.80 | 6.06 | 10.76 | 9.18 | 1.73 |
| 20 | -0.05 | 0.00 | -0.11 | 4.83 | 4.13 | 11.53 | 8.81 | 1.78 |
| 21 | -1.76 | 3.10 | 2.12 | 5.93 | 4.68 | 9.89 | 8.23 | 1.97 |
| 22 | 0.50 | 0.25 | -0.27 | 4.53 | 5.95 | 12.75 | 9.75 | 1.72 |

## Establishing interpretability system for support vector regression

**Principle and software of support vector regression:** The basic idea of support vector regression is to map the data x into a higher-dimensional feature space by using kernel function and then make linear regression in this space. A detailed description about the theory of support vector regression can be found in references[7,12-14]. LIBSVM2.86 originated from http://www.csie.ntu.edu.tw/-cjlin/libsvm/index.html was adopted to implement support vector regression model in this work. All primitive variables were normalized into -1, +1 and the optimal kernel parameters c, g, p were searched automatically with gridregression.py. The process was completed by LIBSVM2.86 with C++ compiled by us and this self-compiling program was tested and verified by successive verification.

**Non-linear screen of factors based on support vector regression:** The screen of factors is important because not all of factors have remarkable effects on prediction. For non-linear relationship often existing among factors, stepwise linear regression linear screen cannot guarantee good effect[15]. Thus, building a nonlinear screen method to select factors based on support vector regression is necessary. Assume there are n samples, p factors, indistinctive factors are successively swept from support vector regression model containing all descriptors with last-elimination method.

For the first-round selection,denoted by $F_j = (Q_j\text{-}Q)/(Q/(n\text{-}m\text{-}1))$, j = 1, 2, …, m, (2) its degree of freedom is (1, n-m-1) and where $Q = \sum_{i=1}^{n}(y_i - \hat{y}_t)^2$ (3) is residual sum of squares of m factors and $Q_j$ is residual sum of squares with the jth factor deleted. If min Fj > F(a,1,n-m-1), it indicates that there is no descriptor to reject and the elimination is over. On the contrary, the next-round selection is carried out after rejecting the jth factor (change the m in the formulae into m-1 at this time) until no factor can be rejected.

Under the assumption that there are  factors reserved, on the basis of support vector regression, n samples and  m' factors, kernel parameters c, g, p are automatic optimized with leave-one-out and support vector regression model is ultimately established after training. The radial basis function (RBF) kernel function is selected as optimal kernel function with success experiences in this work.

**Leave-one-out method:** The leave-one-out (LOO) method is the extreme case of cross-validation. It is one of the most important and efficient methods to evaluate the stability of model and the predict ability of external samples. To perform a leave-one-out test, one single sample is left out as a test data and the rest samples are used as the training dataset and then the next sample is left out, the process repeats until all the samples have been tested once. Leave-one-out method needs large amount of calculation, so it is suitable for small samples[16].

**Testing significance of regression model:** In previous studies[17,18], mean squared error (MSE) was often used as an evaluation index to assess the models established by support vector regression. However, it does not have comparability among different datasets and cannot give out qualitative judgement whether the model is available. In order to test whether the regression of support vector regression model is significant, we adopted statistics F and defined it as F = U/m'/(Q/(n-m'-1) (4), its degree of freedom is (m', n-m'-1) and where U defined as regression sum of squares $U = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ (5), which reflects fluctuation of dependent variable caused by variation of factors and Q reflects error-sum caused by experimental error and other reasons, $\hat{y}_i$ represents fit value of the ith sample treated through back substitution in support vector regression model, $y_i$ is measured value and $\hat{y}_i$ denotes average value of $y_i$. If F > $F_a$ (m',n-m'-1), we can assert the model has significant nonlinear regression at level a.

**Analyzing improtance of factors:** If factor $x_j$ has an important influence on dependent variable y, predict value $\hat{y}$ will vary obviously with the change of $x_j$. We first fix factor $x_j$ as $\bar{x}_j$ (regarded as zero-level of $x_j$) and import it into support vector regression model, then regression sum of squares $U_j$ and residual sum of squares $Q_j$ could be obtained from predict value. U-$U_j$ represents the contribution of regression sum of squares from factor $x_j$. In multiple linear regression or quadratic polynomial regression model, sum of squares is expressed as SSy, where $SS_y = \sum_{i=1}^{n}(y_i - \bar{y}_j)^2$ (6) and $SS_y = Q + U$. But in support vector regression model, $SS_y \neq Q + U$, $SS_y \neq Q_j + U_j$. In order to make importance comparison among factors, we noticed that the numerical magnitude of $X_j$,$U_j$ and $Q_j$ of the same factor just has relative value, so we make $U_j$ and $Q_j$ normalized into $SS_y = Q_j' + U_j'$ through formulas: $Q_j' = Q_j/(Q_j + U_j) \times SS_y$ (7) and $U_j' = U_j/(Q_j + U_j) \times SS_y$ (8). By the same principle, Q and U normalized into $SS_y$ = Q' + U' through formulas: Q' = Q/(Q+U) × $SS_y$ (9) and U' = U/(Q + U) × $SS_y$ (10). Let $V_j$ be formulae as $V_j$ = U'-$U_j$' = $Q_j$'-Q' (11), then statistic $F_j$ obtained as $F_j = V_j$/(Q'/(n-m'-1) (12) could be used for significant test to distinguish relative importance of each factor and its degree of freedom is (1,n-m'-1) .

**Analyzing single-factor effect and sensitivity:** In order to get the optimal value range of $x_j$, we need to know the influence trend of dependent variable y by the single factor in application. When we analyze the single-factor effect of $x_j$, we can fix all factors except $x_j$ as their average value and take values for $x_j$ with certain step size in a given interval, then introduce them into support vector regression model to get the predict value $\hat{y}_j$ and map the $x_j - \hat{y}_j$ relation picture. While the other factors take average value, we can get the variation of dependent variable with $x_j$, especially the value of $x_j$ as dependent variable take extreme value. The sensitivity analysis of each factor references to Tang and Feng[18].

**Analyzing the interaction between factors:** Analysis of the interaction between factors could reference to variance analysis of two-way classification data[19]. For m' retained factors, the method to analyze interaction between factor A and B is that: fix all factors except A and B into their average value, take values for factor A with a levels at same spacing in its original taking value interval and correspondingly take b levels for factor B to compose a*b tested samples (generally, a = b = 5), then imput tested samples into support vector regression model to get a*b predict value $\hat{y}_{ij}$ . As $\hat{y}_{ij}$ is the predict value obtained by model, its experiment error can be regarded as 0, so interacting sum of squares of A*B takes the formal $SS_{AB} = SS_y - SS_A - SS_B$ (13), where $SS_A$ is deviation square of factor A and represents as $SS_A = b\sum_{i=1}^{a}(\bar{\hat{y}}_i - \bar{\hat{y}})^2$ (14), where $\bar{\hat{y}}_i$ represents the average of b predict value of factor A in the ith level and $\bar{\hat{y}}$ is the average of all forecast value $\hat{y}_{ij}$ , $SS_B$ is similar with $SS_A$. The statistic $F_{AB}$ obtained as $F_{AB}$ = $SS_{AB}$/(a-1) × (b-1)/(Q/(n-m'-1) (15), whose degree of freedom is (a-1) × (b-1), n-m'-1), is used to test interaction significant of A*B and make importance ranking for several two-factor interactions.

## RESULTS AND DISCUSSION

**Screen of factors:** For all 22 samples, factors were screened nonlinearly with leave-one-out method based on support vector regression, the results shown that the second and fifth factors were deleted, which respectively correspond to hydrophobic parameters $\pi^2$ and stereoscopic effect parameters $B_5$. It indicated that those two factors have limited effects of toxicity of ternary dissymmetric organic phosphate insecticide. The result coincided with the description that factor π has good correlation with $\pi^2$ and $B_5$ has high relativity with $^0X_v$ and $^1X_v$[10]. The fit and predict model established with remained 5 factors was much better than corresponding models based on all 7 factors. The multiple correlation coefficients of fit was F = 30.58 referencing to F = 14.91 (F 0.01 (5, 16) = 4.44), while the predict model was that F = 11.3 and F = 4.27, respectively.

**Model comparison in fit and predict:** The fit and leave-one-out predict for activity of organophosphorus insecticide were carried out on the basis of support vector regression with 5 retained factors and stepwise linear regression adopted by Jin *et al.*[10] with all factors, respectively. Support vector regression model shows great advantage by comparing corresponding result with stepwise linear regression, the detailed result is shown in Table-2.

TABLE-2
FIT AND PREDICT OF SUPPORT VECTOR REGRESSION (SVR) AND STEPWISE LINEAR REGRESSION (SLR)

| Model | **Fit** | | **Predict with LOO** | |
|---|---|---|---|---|
| | SVR | SLR | SVR | SLR |
| MSE | 0.0066 | 0.0247 | 0.02 | 0.0604 |
| MAPE %) | 4.0873 | 8.1959 | 9.5553 | 15.4162 |
| F | 30.5802 | 3.9777 | 11.3037 | 1.5782 |
| R | 0.955 | 0.8157 | 0.8621 | 0.5153 |

**Importance analysis of factors:** Jin *et al.*[10] adopted stepwise linear regression equation to describe the relationship between different factors and insecticidal activity, the importance rank of all factors based on linear terms is $^0X_v$ ** > π** > $^1X_v$ > $\pi^2$ > s, the other two factors $B_5$ and L were deleted in path analysis (superscript character ** represents extremely significant level of 0.01, superscript character * represents significant level of 0.05 and non-superscript represents no significant level of 0.05). The importance analysis results of factors based on support vector regression are shown in Table-3, the importance of 5 retained factors all reach extremely significant level (F 0.01(1,16) = 8.53), the only order is $^1X_v$ ** > π** > $^0X_v$ ** > L ** > σ**. Because of high correlation between $^0X_v$ and $^1X_v$, our interpretability system and traditional model all explained that hydrophobic parameters and connectivity index are the most influence factors for activity of organophosphorus insecticide.

TABLE 3
DETAILS OF IMPORTANCE ANALYSIS OF FACTORS

| | π | σ | L | $^0$Xv | $^1$Xv |
|---|---|---|---|---|---|
| Q | 0.8786 | 0.6606 | 1.7184 | 7.4974 | 10.2402 |
| U | 1.1226 | 1.1938 | 2.7025 | 9.6184 | 8.012 |
| Q' | 0.7133 | 0.5787 | 0.63149 | 0.7116 | 0.9115 |
| U' | 0.9113 | 1.0458 | 0.9931 | 0.9129 | 0.7131 |
| F | 58.1569 | 44.1694 | 49.6531 | 57.9855 | 78.7604 |

**Analysis of single-factor effect and sensitivity:** The results of single-factor effect and sensitivity are shown in Figs. 1 and 2, in which each factor got normalized x-coordinate value by formal $x_j$' = ($x_j$ - min x)/(max x-min x) (16), where and corresponded to upper and lower limit of $x_j$ in each column, respectively. In Fig. 1, we can draw that insecticidal activity increases with increasing factor $^0X_v$ and L but decreases with increasing $^1X_v$ and π and it is almost stable with increasing σ, where $^0X_v$ and $^1X_v$ have significant effect. In Fig. 2, with increasing insecticidal activity, factor π, σ and L keep invariant, factor $^0X_v$ decreases sharply at first and then changes to be steady, while $^1X_v$ is just in contrary to $^0X_v$.

**Analysis of the interaction between factors:** The results of interaction between two factors based novel system are

TABLE-4
SIGNIFICANCE TEST OF THE INTERACTION BETWEEN FACTORS

| | x1x3 | x1x4 | x1x6 | x1x7 | X3x4 | x3x6 | x3x7 | x4x6 | x4x7 | x6x7 |
|---|---|---|---|---|---|---|---|---|---|---|
| SST | 0.74 | 2.44 | 28.81 | 21.97 | 1.27 | 26.17 | 23.52 | 26.02 | 24.14 | 49.10 |
| SSx1 | 0.71 | 0.96 | 0.76 | 0.70 | 0.03 | 0.07 | 0.02 | 0.72 | 1.26 | 27.82 |
| SSx2 | 0.02 | 1.31 | 28.03 | 21.24 | 1.24 | 25.83 | 23.24 | 24.32 | 22.69 | 21.26 |
| SSx1x2 | 0.003 | 0.17 | 0.03 | 0.02 | 0.001 | 0.27 | 0.26 | 0.98 | 0.19 | 0.01 |
| F | 0.02 | 1.19 | 0.21 | 0.15 | 0.01 | 1.87 | 1.77 | 6.72 | 1.29 | 0.07 |

shown in Table-4, from which we can get the order x4x6** > x3x6 > x3x7 > x4x7 > x1x4 > x1x6 > x1x7 > x6x7 > x1x3 > x3x4, where the interaction of x4x6 reaches extremely significant level, whose F value equals to 6.72 and F 0.01(16,15) = 3.49. And other interaction between different factors does not reach significant level. Factor x2 and x5 no longer participate in interaction analysis because they have already been rejected in support vector regression model.

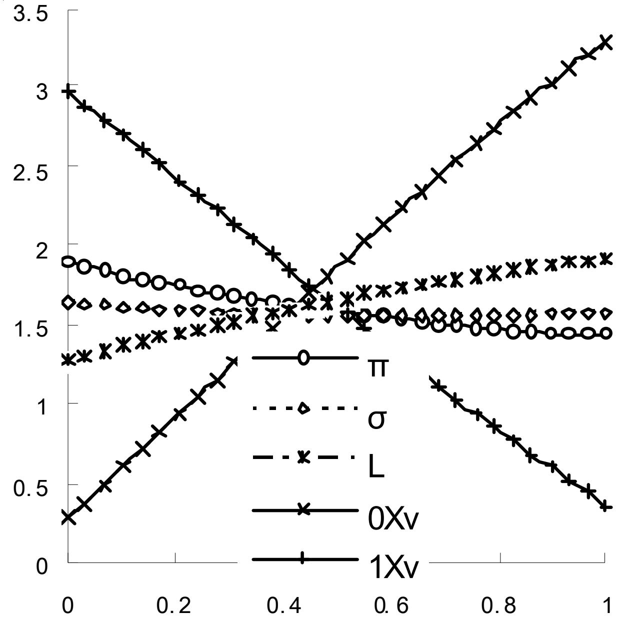


Fig. 1. X-Y line graph of single-factor effect

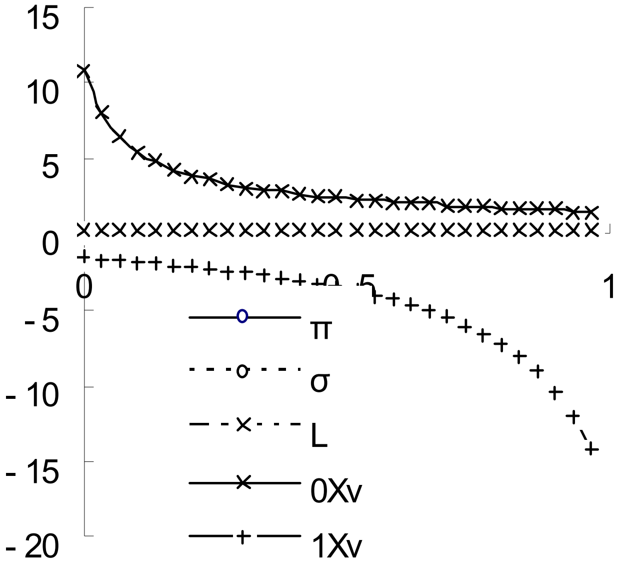


Fig. 2. X-Y line graph of sensitivity

## Conclusion

(1) Support vector regression model is much better than previous published model in quantitative structure-activity relationship modeling for fit and prediction of organophosphorus insecticide.

(2) The novel interpretability system for support vector regression can give out the comparability among different datasets and qualitative judgement whether the model is available.

(3) Compared with traditional model, the obtained explanatory results of organophosphorus insecticide preliminarily indicate that the new interpretability system is reasonable.

(4) The rationality and the precise difference with traditional models of this novel system still need much more verifiable experimental results to further support.

## ACKNOWLEDGEMENTS

This work was supported by a grant from the Science Foundation for Distinguished Young Scholars of Hunan Province, China (No. 10JJ1005), the Research Fund for the Doctoral Program of Higher Education of China (No. 200805370002).

## REFERENCES

1. M.H. Fatemi and S. Gharaghani , *Bioorg. Med. Chem.*, **15**, 24 (2007).
2. P.P. Roy and K. Roy, *QSAR Comb. Sci.*, **27**, 3 (2009).
3. X.S. Tan, Z.M. Yuan, T.J. Zhou, C.J. Wang and J.Y. Xiong, *Chem. J. Chin. Univ.*, **29**, 1 (2008).
4. A.R. Conn, K. Scheinberg and L.N. Vicente, *IMA J. Numer. Anal.*, **28**, 4 (2008).
5. I. Yilmaz and A.G. Yuksek, *Rock. Mech. Rock. Eng.*, **41**, 5 (2008).
6. P.J. Lisboa and A.F. Taktak, *Neural. Networks*, **19**, 4 (2006).
7. V.N. Vapnik, The Nature of Statistical Learning Theory, New York, pp. 385-421 (2004).
8. J.S. Alex and S.A. Brenhard, *Stat. Comput.*, **14**, 3 (2000).
9. P.A. Alexei and K. Mikhail, *Stoch. Env. Res. Risk. A.*, **22**, 5 (2008).
10. W. Jin, W.H. Huang and X.H. Lu, *J. Anal. Sci.*, **18**, 6 (2002).
11. Y. Zhou, Y.K. Huang and G.Y. Sui, *J. Zhejiang. Univ. Technol.*, **27**, 2 (1999).
12. J.C. Meng and L. Xia, *Int. J. Infrared Milli. Waves*, **28**, 7 (2007).
13. W.M. Zhong, D.Y. Pi and Y.X. Sun, *J. Cent. South Univ. Technol.*, **14**, 3 (2007).
14. Q. Wu, W.Y. Liu and Y.H. Yang, *J.J. Cent. South Univ. Technol.*, **14**, 442 (2007).
15. S. Fukuda, A.M. Gillinov and P.M. McCarthy, *Circulation*, 114 (2006).
16. Z.X. Yu, C.B. Zhou and J.P. Li, *Chin. J. Rock. Mech. Eng.*, **24**, 14 (2005).
17. R. Guha, *J. Comput. Aid. Mol. Des.*, **22**, 12 (2008).
18. Q.Y. Tang and M.G. Feng, DPS Data Processing System-Experiment Design, Statistical Analysis and Data Mining, Science Press, Beijing, pp. 295-304 (2007).
19. J.Y. Gai, Statistics Method for Experiments, Chinese Agriculture Press, Beijing, pp. 118-120 (2003).