Electronic Journal of Biotechnology ISSN: 0717-3458
http://www.ejbiotechnology.info
DOI: 10.2225/vol15-issue1-fulltext-8 *RESEARCH ARTICLE*

# Prediction of xylanase optimal temperature by support vector regression

**Guangya Zhang**[1] 🖂 · **Huihua Ge**[1]

1 Huaqiao University, College of Chemical Engineering, Xiamen, Fujian, PR China

🖂 Corresponding author: zhgyghh@hqu.edu.cn

Received July 4, 2011 / Accepted December 12, 2011
Published online: January 15, 2012

## Abstract

**Background:** Support vector machine (SVM), a novel powerful machine learning technology, was used to develop the non-linear quantitative structure-property relationship (QSPR) model of the G/11 xylanase based on the amino acid composition. The uniform design (UD) method was applied to optimize the running parameters of SVM for the first time. **Results:** Results showed that the predicted optimum temperature of leave-one-out (LOO) cross-validation fitted the experimental optimum temperature very well, when the running parameter $C$, Ɛ, and γ was 50, 0.001 and 1.5, respectively. The average root-mean-square errors (RMSE) of the LOO cross-validation were 9.53ºC, while the RMSE of the back propagation neural network (BPNN), was 11.55ºC. The predictive ability of SVM is a minor improvement over BPNN, but it is superior to the reported method based on stepwise regression. Two experimental examples proved the validation of the model for predicting the optimal temperature of xylanase. **Conclusion:** The results indicated that UD might be an effective method to optimize the parameters of SVM, which could be used as an alternative powerful modeling tool for QSPR studies of xylanase.

**Keywords:** amino acid composition, optimum temperature, support vector machine, uniform design, xylanase

## INTRODUCTION

Xylanase has a wide range of potential biotechnological applications. Recently the interest in xylanase has markedly increased due to the potential industrial uses, particularly in pulping and bleaching processes (Beg et al. 2001; Diaz et al. 2004; Oliveira et al. 2006). The thermo-alkalophilic conditions of xylanase-aided bleaching (60-80ºC, pH 8-10) combined with a high level of activity, demand a set of characteristics of xylanases, not usually found in native enzymes. An alternative for obtaining new thermostable enzymes is the modification of presently used xylanases to be more stable in extreme conditions. During the last twenty years, rational site-directed mutation (Moreau et al. 1994) and irrational directed evolution (Fenel et al. 2006) have become a routine approach for engineering xylanases to achieve this goal. Although, the so-called ‘semi-rational’ approach, which used computational techniques to perform *in silico* screening of protein sequences or to enhance the efficiency of directed evolution, has become an emerging area in protein engineering. However, it has not been employed in xylanase engineering. This approach has been applied with some success (Hayes et al. 2002; Mildvan, 2004), and researchers think that it may pave the way to exciting areas of enzyme research including efficient engineering of existing biocatalysts (Chica et al. 2005). Protein design algorithms (mathematical models) that provide quantitative structure-property relationship (QSPR) of proteins are the core part of the ‘semi-rational’ approach.

The support vector machine (SVM) is a new and very promising classification and regression method developed by Vapnik (1998). It has been shown that SVM has two distinct features. Firstly, it has high generalization ability. Secondly, it requires only small size of training samples. According to some literatures, SVM has shown promising results on several biological problems and is becoming established as a standard tool in bioinformatics (Ward et al. 2003; Cai et al. 2004; Chen et al. 2006).

In the present investigation, SVM, as a novel machine learning technique, was used to establish a model for predicting the optimum temperature of xylanase in G/11 family. During the process, the uniform-design method was applied to optimize the running parameters of SVM. The aim was to establish a new QSPR model and to confirm the possibility of predicting the optimum temperature of xylanases. The performances of SVM were better than that of back propagation neural network (BPNN) and the reported models, and may be useful for computer virtual screening in engineering for more thermostable new xylanases.

## MATERIALS AND METHODS

### Dataset construction

To reduce the redundancy, we downloaded the sequences of xylanases from UniProt, for it contains records with full manual annotation or computer-assisted, manually-verified annotation performed by biologists and based on published literature and sequence analysis (Bairoch et al. 2005). The optimum temperatures of xylanases obtained from Liu's work (Liu et al. 2006) have been shown in Table 1. Altogether, 25 xylanase sequences and their corresponding optimum temperatures were obtained.

**Table 1. Observed and calculated temperature values of xylanase in family G/11.**

| ID | $T_{opt}$ | $T_{cal1}$ | $T_{cal2}$ | $T_{RSVM}$ | $T_{LSVM}$ | $T_{BPNN}$ |
|---|---|---|---|---|---|---|
| P29127 | 40.00 | 52.27 | 40.83 | 40.04 | 40.24 | 40.07 |
| P36218 | 40.00 | 36.67 | 40.73 | 40.05 | 40.20 | 40.77 |
| P17137 | 43.00 | 52.65 | 42.42 | 43.03 | 43.18 | 44.15 |
| Q06562 | 45.00 | 40.52 | 45.01 | 45.05 | 45.24 | 45.48 |
| P36217 | 45.00 | 50.02 | 47.04 | 45.06 | 44.80 | 43.88 |
| P29126 | 50.00 | 45.17 | 49.22 | 50.05 | 49.80 | 49.28 |
| P18429 | 50.00 | 52.13 | 53.93 | 50.00 | 49.75 | 51.61 |
| P00694 | 50.00 | 47.12 | 49.32 | 50.04 | 49.80 | 50.14 |
| P48793 | 50.00 | 46.63 | 47.04 | 50.08 | 50.25 | 50.52 |
| P33557 | 50.00 | 54.99 | 51.26 | 50.05 | 50.19 | 51.12 |
| P55328 | 50.00 | 54.99 | 54.22 | 50.06 | 51.80 | 54.72 |
| P55333 | 54.00 | 54.82 | 54.18 | 54.04 | 54.25 | 53.51 |
| P35811 | 55.00 | 49.47 | 55.31 | 54.95 | 54.77 | 55.57 |
| P45796 | 55.00 | 58.23 | 56.50 | 55.03 | 55.20 | 54.82 |
| P26220 | 55.00 | 52.38 | 55.04 | 55.04 | 54.78 | 54.99 |
| P26515 | 55.00 | 51.10 | 54.66 | 54.97 | 56.21 | 55.52 |
| P45705 | 55.00 | 55.56 | 53.83 | 54.99 | 54.81 | 55.43 |
| P09850 | 55.00 | 56.60 | 53.93 | 54.98 | 50.29 | 53.12 |
| P48824 | 55.00 | 57.96 | 54.08 | 54.95 | 55.19 | 54.84 |
| P55332 | 58.00 | 56.58 | 57.76 | 57.96 | 57.75 | 58.02 |
| P55334 | 60.00 | 63.10 | 59.99 | 59.96 | 60.22 | 60.12 |
| P55329 | 60.00 | 54.99 | 54.22 | 59.96 | 51.51 | 54.42 |
| P35809 | 65.00 | 63.24 | 64.62 | 64.98 | 64.79 | 64.30 |
| P33558 | 70.00 | 64.73 | 69.61 | 69.98 | 61.72 | 69.74 |
| O43097 | 82.00 | 72.37 | 82.39 | 81.98 | 76.84 | 82.25 |
| MAE | | 4.21 | 1.23 | 0.04 | 1.35 | 0.94 |
| RMSE | | 5.03 | 1.9 | 0.04 | 2.79 | 1.63 |

ID: the accession number of xylanase in Swiss-Prot; $T_{opt}$: the optimum temperature found in the literature at which the relative xylanase has the maximum activity; $T_{cal1}$ and $T_{cal2}$: the training temperature according to reference 12; $T_{RSVM,}$ $T_{LSVM}$ and $T_{BPNN}$: the training temperature based on RBF support vector machines, linear support vector machines and back propagation neural network in our models; MAE: mean absolute error; RMSE: root-mean-square error.

**Support vector regression** (Cortes and Vapnik, 1995; Vapnik, 1998)

SVM can be applied in both classification and regression; here we used support vector regression (SVR). In SVR, the basic idea is to map the data *X* into a higher-dimensional feature space *F via* a nonlinear mapping Φ and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i, d_i)\}_i^n$ ($x_i$ is the input vector, $d_i$ is the desired value, and *n* is the total number of data patterns), and SVM approximate the function using Equation 1:

$$y = f(x) = w\Phi(x) + b$$

**[Equation 1]**

Where Φ*(x)* is the high-dimensional feature space which is nonlinearly mapped from the input space *x*. The coefficients *w* and *b* are estimated by minimizing

$$R_{SVMs}(c) = C\frac{1}{n}\sum_{i=1}^{n} L_{\in}(d_i, y_i) + \frac{1}{2}\|w\|^2$$

**[Equation 2]**

$$L_{\in}(d, y) = \begin{cases} 0\ ,\ if\ |y - d| \le \in \\ |d - y| - \in\ ,\ otherwise \end{cases}$$

**[Equation 3]**

The first term $C\frac{1}{n}\sum_{i=1}^{n} L_{\in}(d_i, y_i)$ in Equation 2 is the empirical error (risk). They are measured by the Ɛ-insensitive loss function given by Equation 3. On the other hand, the second term $\frac{1}{2}\|w\|^2$ in Equation 2 is the regularization term. *C* is the regularized constant and it determines the trade off between the empirical risk and the regularization term. Ɛ is called the tube size and it is equivalent to the approximation accuracy placed on the training data points. Both *C* and Ɛ are user -prescribed parameters.

To obtain the estimations of *w* and *b*, Equation 2 is transformed to the primal function given by Equation 4 by introducing the positive slack variables $\xi_i$ and $\xi_i^*$ as following:

$$Minimize\ \ R_{SVMs}(w, \xi^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$Subject\ \ to\ \ \begin{cases} d_i - w\Phi(x_i) - b \le \in + \xi_i \\ w\Phi(x_i) + b - d_i \le \in + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, \dots, n \end{cases}$$

**[Equation 4]**

Finally, the regression function given by Equation 1 has the following explicit form

$$f(x,a_i,a_i^*) = \sum_{i=1}^{n_{SV}} (a_i - a_i^*) k(x,x_i) + b$$

**[Equation 5]**

Where $n_{SV}$ is the number of Support Vectors (SVs), $a_i$, $a_i$* are the introduced Lagrange multipliers and they satisfy the equality $a_i \cdot a_i^* = 0, a_i \geq 0, a_i^* \geq 0$, and the kernel function *K* corresponding to

$$K(x,x_i) = \phi(x)^T \phi(x_i)$$

**[Equation 6]**

Linear and radial basis function (RBF, Gaussian function) kernels are two commonly used kernels in SVR (Smola and Schölkopf, 1998) and are given by Linear kernel

$$K(x_i,x_j) = (x_i \cdot x_j)$$

**[Equation 7]**

RBF kernel $K(x_i,x_j) = exp(-\gamma \left\| x_i - x_j \right\|^2$

**[Equation 8]**

Where γ is a constant, the parameter of the kernel, it controls the width of the Guassian kernel (although itself is not the width) and therefore, controls the generalization ability of SVM. The generalization performance of SVR depends on a good setting of parameters: *C*, Ɛ and the kernel type and corresponding kernel parameters. Here, uniform design is employed in optimizing the running parameters.

**Uniform design**

Uniform design (UD) was first proposed by Fang (1980), based on theoretic accomplishments in number-theoretic method. Generally speaking, UD is a form of 'space filling' design. Suppose that the experimental domain consists of *s* factors and *h(x)* is a response of the experiment. In many cases, we can assume this domain to be the unit cube $C^s$. The expectation value of *h(x)* over the experimental domain *Eh(x)* can be estimated by the sample mean, $\hat{h} = \frac{1}{n} \sum_{x\rho} h(x)$ where *ρ* is a set of *n* experimental points over the domain. The famous Koksma–Hlawka inequality gives the upper error bound of the estimate of *Eh(x)*,

$$\left| Eh(x) - \hat{h} \right| \leq D(\rho)V(h)$$

**[Equation 9]**

Where *V(h)* is a measure of the variation of *h* and D(*ρ*) the discrepancy of set ρ. The inequality indicates that the more uniform set in ρ is over the experimental region, the more accurate the estimate *Eh(x)* provides to $\hat{h}$ ( Zhang et al. 1998). Thus, obtaining experimental points that are most uniformly scattered in the domain is the key step in uniform design. Uniform design has its own features, such as its functional agility of arranging experiment runs and its robustness against model uncertainty. For more detail information, reference is Liang and Fang's works (Fang and Yang, 2000; Liang et al. 2001).

The overall performances of SVM and BPNN were evaluated in terms of the root-mean-square error (RMSE) and mean absolute error (MAE), which was defined as below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

**[Equation 10]**

$$MAE = \frac{1}{n}\sum_{i-1}^{n}\left|y_i - \hat{y}_i\right|$$

**[Equation 11]**

Where $y^i$ and $\hat{y}_i$ stand for the actual value and training value (or predicted value), respectively.

**Cross-validation**

The performance and robustness of the models was evaluated by cross-validation. The jackknife test (leave-one-out, LOO) was used; it was deemed the most rigorous and objective with the least arbitrariness, as demonstrated by an incisive analysis in a recent review (Chou and Shen, 2007). We used 24 data points to train the models and tested it with the left one. This was repeated 25 times, leaving in turn a different data point out of the training set and using it to validate the resulting models.

**Software and computation environment**

To analyze the 20 amino acid compositions of xylanases, Bioedit software was used (version 5.0.9), and then each xylanase in the data set was characterized by a vector $x_i$(i = 1,…, N). The input vector $x_i$ has 20 coordinates for the amino acid composition (in the percentage). The SVR and BPNN were performed by the software of *WEKA*, which is a java package providing an environment for implementation of a large number of machine learning and statistical algorithms (Frank et al. 2004). All the computations were carried out on a Pentium IV computer with a 2.7 GHz processor and 512M RAM.

## RESULTS

### Optimizing the parameters of linear kernel SVM based on UD

Similar to other multivariate statistical models, the performances of SVM for regression depend on the combination of several parameters. They are penalty value *C*, ε of ε-insensitive loss function, the kernel type *K*, and its corresponding parameters. To get the best generalization ability, some strategies are needed for optimizing these factors. There are four possible choices of kernel functions, *i.e.* linear, polynomial, radial basis function (RBF), and sigmoid function. For regression tasks, we selected linear and RBF kernel.

For linear kernel, there are only two parameters, *C* and ε. Here the UD method was employed in optimizing the combination of the parameters based both on LOO cross-validation and training. The UD table for 2 factors with 16 levels was used and the results were shown in Table 2. For linear kernel, we found that when the regularization parameter (*C*), and ε -insensitive loss function (ε) was 1 and 0.005 respectively, the MAEs of LOO cross-validation and training were 7.56ºC and 1.35ºC, respectively, with an average MAE of 4.46ºC. The LOO cross-validation results of the 9th run (R9) were slightly better, but the training results of it were much worse. On the other hand, the training results of the 6th run (R6) were slightly better than our chosen run (R12), but the LOO cross-validation results of R6 were much worse (about 5.64ºC). From the results of R7, one can see that parameters should not be chosen only based on training error, this could easily lead to over-fitting. So the optimal *C* and ε for linear SVM were finally chosen as 1 and 0.005. The training and LOO cross-validation results of linear SVM were shown in Table 1 and Figure 1.

**Table 2. Uniform design of the two factors $U_{16}$ ($16^2$) to optimize the combination of the parameters of SVM with linear kernel based on LOO cross-validation and training.**

| Run no. | Linear | | MAE | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|
| | C | ε | CV | TR | AV | CV | TR | AV |
| R1 | 0.005 | 0.1 | 6.80 | 6.32 | 6.56 | 9.58 | 8.89 | 9.24 |
| R2 | 1000 | 0.15 | 7.95 | 4.49 | 6.22 | 10.94 | 4.90 | 7.92 |
| R3 | 5000 | 0.05 | 21.86 | 2.10 | 11.98 | 52.81 | 2.47 | 27.64 |
| R4 | 0.5 | 0.2 | 8.00 | 5.07 | 6.54 | 11.11 | 5.86 | 8.49 |
| R5 | 0.1 | 0.4 | 8.06 | 7.26 | 7.66 | 10.96 | 9.05 | 10.01 |
| R6 | 100 | 0.00001 | 13.20 | 0.79 | 7.00 | 24.51 | 1.91 | 13.21 |
| R7 | 25000 | 0.001 | 266.52 | 0.80 | 133.66 | 634.23 | 1.84 | 318.04 |
| R8 | 5 | 1 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| R9 | 0.05 | 0.0001 | 6.92 | 5.15 | 6.04 | 9.28 | 7.80 | 8.54 |
| R10 | 500 | 0.6 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| R11 | 0.01 | 0.7 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| **R12** | **1** | **0.005** | **7.56** | **1.35** | **4.46** | **9.92** | **2.79** | **6.36** |
| R13 | 50000 | 0.3 | 7.86 | 1.35 | 4.61 | 10.58 | 2.79 | 6.69 |
| R14 | 50 | 0.5 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| R15 | 10 | 0.01 | 8.80 | 1.00 | 4.90 | 15.40 | 1.77 | 8.59 |
| R16 | 10000 | 0.8 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |

CV: leave-one-out cross validation; TR: training; AV: average; MAE: mean absolute error; RMSE: root-mean-square error. The optimum values are highlighted.

### Optimizing the parameters of RBF kernel SVM based on UD

For the RBF kernel, there are three parameters *C*, ε and γ. Here the UD method was also employed in optimizing the combination of the parameters based both on LOO cross-validation and training. The UD table for 3 factors with 16 levels was used and the results were shown in Table 3.

From the results of the Table 3, we can see that different combination of the three parameters might result in different MAE and RMSE values. When *C*, ε and γ, were 50, 0.001 and 1.5 respectively, the MAEs of LOO cross-validation and training were 6.88ºC and 0.04ºC, respectively, with an average MAE of 3.46ºC. Although the LOO cross-validation results of the 5th run (R5) were slightly better (about 0.28ºC), the training results of it were much worse (about 5.5ºC). From the results of R5, one can see that parameters should not be chosen only based on LOO cross-validation error. So the optimal *C*, ε and γ for RBF SVM were finally chosen as 50, 0.001 and 1.5. The training and LOO cross-validation results of RBF SVM were also shown in Table 1 and Figure 1.

According to Table 2 and Table 3, one can observe that many different combinations of parameters resulted in the same LOO cross-validation and training errors, which means that SVMs are not so sensitive to parameters. Meanwhile, the RBF kernel is superior to linear kernel, which was in accordance with some former researches for support vector regression tasks (Xue et al. 2004; Liu et al. 2005).

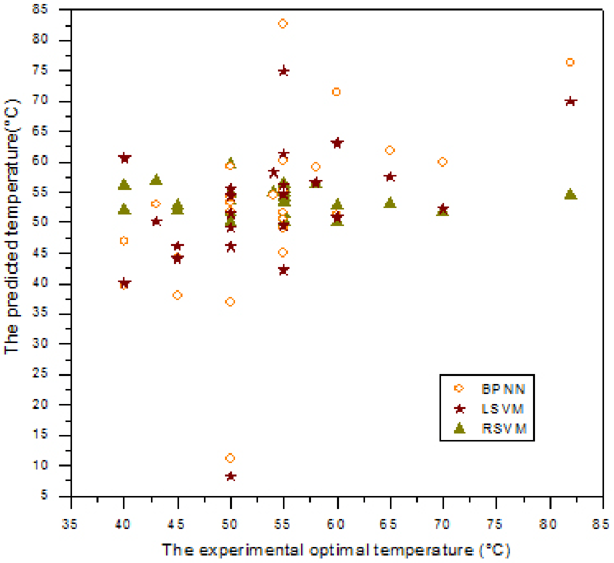

**Fig. 1 The predicted temperatures *vs* the experimental optimal temperatures of BPNN and SVM.** BPNN: back-propagation neural network; LSVM: support vector machine with linear kernel; RSVM: support vector machine with RBF kernel. It is the prediction results of the LOO cross validation.

### Compared with back propagation neural network (BPNN)

Recently, a few studies have shown that SVM yielded better results than alternative machine learning techniques such as BPNN. In this study, we have compared the performance of SVM and BPNN with the same datasets. The architecture of BPNN was also optimized by UD and the results were shown in Table 4. During the process, the maximum iterations were appointed as 1000. According to Table 4, the learning rate ($\eta$), momentum parameter and the neuron number of the hidden layer was chosen as 0.04, 0.6 and 11, respectively. The MAEs of LOO cross-validation and training were 7.73ºC and 0.97ºC, respectively, with an average MAE of 4.35ºC. The training and LOO cross-validation results of BPNN were shown in Table 1 and Figure 1.

According to Table 4, one can observe that different combinations of parameters resulted in different LOO cross-validation and training MAEs and RMSEs. This means that BPNN may be more sensitive to its running parameters when compared with SVMs, especially RBF SVM. And also, the LOO cross-validation results of BPNN were widely different when different sets of training and LOO cross-validation were employed. The maximum and minimum LOO cross-validation MAEs were 38.89ºC and 0.2ºC, respectively, while the corresponding MAEs of RBF SVM were 27.44ºC and 0.03ºC. The predicted errors of all the 25 runs of BPNN and SVM are shown in Figure 2. For linear SVM, 13 samples had small differences to their experimental optimal temperatures ($|ERROR|$<5ºC), and RBF SVM also had 13 samples, while 11 samples had small differences to their experimental optimal temperatures in BPNN. The predicted RMSEs of linear SVM, RBF SVM and BPNN were 9.92ºC, 9.55ºC and 11.52ºC, respectively. As analyzed above, it can be seen that the SVM based models showed minor robustness than BPNN. It was consistent with the inherent advantages over BPNN, which did not show robustness especially in the condition of only small amount of training samples were available.

**Table 3. Uniform design of the three factors $U_{15}$ ($16^3$) to optimize the combination of the parameters of SVM with RBF kernel based on LOO cross-validation and training.**

| Run no. | RBF | | | MAE | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|
| | C | ε | Gamma | CV | TR | AV | CV | TR | AV |
| R1 | 1 | 0.9 | 1 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| R2 | 0.01 | 0.005 | 0.07 | 6.48 | 6.28 | 6.38 | 9.13 | 9 | 9.07 |
| R3 | 10000 | 0.2 | 0.005 | 7.17 | 5.05 | 6.11 | 9.7 | 5.84 | 7.77 |
| R4 | 500 | 0.1 | 0.3 | 7.05 | 3.34 | 5.20 | 9.3 | 3.59 | 6.45 |
| R5 | 5 | 0.05 | 0.001 | 6.6 | 5.54 | 6.07 | 9.1 | 7.94 | 8.52 |
| R6 | 0.1 | 0.15 | 5 | 6.83 | 5.4 | 6.12 | 9.51 | 7.29 | 8.40 |
| R7 | 5000 | 0.6 | 2.5 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| R8 | 1000 | 0.0001 | 0.03 | 16.22 | 0.38 | 8.30 | 35.62 | 1.59 | 18.61 |
| R9 | 0.05 | 0.7 | 0.01 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| R10 | 25000 | 0.01 | 0.9 | 6.96 | 0.43 | 3.70 | 9.51 | 0.43 | 4.97 |
| R11 | 10 | 0.4 | 0.09 | 8.56 | 7.54 | 8.05 | 11.12 | 9.08 | 10.10 |
| R12 | 50000 | 0.8 | 0.1 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| R13 | 0.005 | 0.5 | 0.5 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| R14 | 100 | 1 | 0.05 | 10.08 | 9.84 | 9.96 | 11.97 | 11.48 | 11.73 |
| **R15** | **50** | **0.001** | **1.5** | **6.88** | **0.04** | **3.46** | **9.55** | **0.04** | **4.80** |
| R16 | 0.5 | 0.00001 | 0.1 | 6.98 | 4.47 | 5.73 | 9.45 | 7.11 | 8.28 |

CV: leave-one-out cross validation; TR: training; AV: average; MAE: mean absolute error; RMSE: root-mean-square error. The optimum values are highlighted.

**Table 4. Uniform design of the three factors $U_{15}$ ($15^3$) to optimize the combination of the parameters of BPNN based on LOO cross-validation and training.**

| Run no. | Three factors | | | MAE | | | RMSE | | |
|---|---|---|---|---|---|---|---|---|---|
| | LR (η) | MP | NHL | CV | TR | AV | CV | TR | AV |
| R1 | 0.15 | 0.65 | 19 | 10.03 | 1.68 | 5.86 | 17.23 | 2.07 | 9.65 |
| R2 | 0.07 | 0.8 | 17 | 9.95 | 1.36 | 5.66 | 12.46 | 1.82 | 7.14 |
| **R3** | **0.04** | **0.6** | **11** | **7.73** | **0.97** | **4.35** | **11.52** | **1.64** | **6.58** |
| R4 | 0.09 | 0.45 | 14 | 8.61 | 1.05 | 4.83 | 11.98 | 1.64 | 6.81 |
| R5 | 0.5 | 0.55 | 15 | 19.49 | 5.02 | 12.26 | 34.13 | 5.33 | 19.73 |
| R6 | 0.05 | 0.35 | 16 | 7.83 | 0.94 | 4.39 | 12.44 | 1.63 | 7.04 |
| R7 | 0.25 | 0.1 | 13 | 9.86 | 1.14 | 5.50 | 15.65 | 1.79 | 8.72 |
| R8 | 0.45 | 0.4 | 7 | 10.53 | 2.39 | 6.46 | 14.54 | 2.65 | 8.60 |
| R9 | 0.2 | 0.5 | 3 | 8.79 | 1.44 | 5.12 | 11.66 | 1.84 | 6.75 |
| R10 | 0.08 | 0.7 | 5 | 8.53 | 1.68 | 5.11 | 11.39 | 2.1 | 6.75 |
| R11 | 0.06 | 0.2 | 8 | 7.77 | 0.94 | 4.36 | 11.73 | 1.63 | 6.68 |
| R12 | 0.1 | 0.3 | 10 | 8.67 | 0.94 | 4.81 | 12.41 | 1.6 | 7.01 |
| R13 | 0.3 | 0.85 | 9 | 18.9 | 11.55 | 15.23 | 29.02 | 13.31 | 21.17 |
| R14 | 0.4 | 0.75 | 12 | 14.07 | 9.41 | 11.74 | 23.49 | 11.17 | 17.33 |
| R15 | 0.35 | 0.25 | 18 | 9.92 | 1.48 | 5.70 | 15.34 | 1.95 | 8.65 |

LR(η): the learning rate; MP: the momentum parameter; NHL: the number of hidden layer nodes; CV: leave-one-out cross validation; TR: training; AV: average; MAE: mean absolute error; RMSE: root-mean-square error. The optimum values are highlighted.

To validate the prediction models, we showed two examples. Firstly, we cloned the xylanase gene of *Bacillus pumilus*, sequenced it and expressed it in *Escherichia coli*. The accession number of the gene in NCBI is EF090270 and the protein ID is ABM54186.1 (http://www.ncbi.nlm.nih.gov/nuccore/EF090270). The optimal temperature of the purified xylanase was 50ºC, which was shown in Figure 3. We calculated the amino acid composition of the xylanase and used the model to predict the optimal temperature. For linear kernel SVM, the predicted optimal temperature was 49.89ºC, for RBF kernel SVM was 50.02ºC, and for BPNN was 49.94ºC, the MAEs were 0.11ºC, 0.02ºC and 0.06ºC, respectively. Secondly, we designed a new thermophilic xylanase, synthesized its coding gene *de nove* and expressed it in *Escherichia coli*. The optimal temperature of it was 60ºC and it can keep over 50% activity at 70ºC for one hour (Fu et al. 2012). The predicted results of linear kernel SVM, RBF kernel SVM and BPNN were 54.77ºC, 55.25ºC, and 56.05ºC, and the MAEs were 5.23ºC, 4.75ºC, and 3.95ºC, respectively. From the two experimental examples, we could conclude that the model we proposed might work as a useful tool for QSPR studies of xylanase and facilitate the engineering for new one.

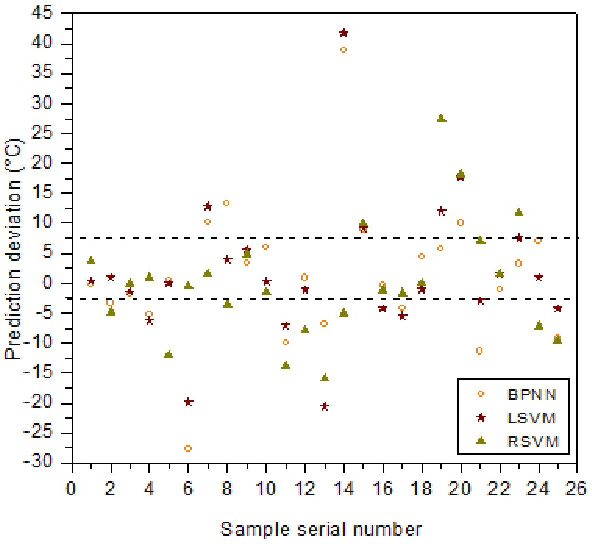

**Fig. 2 The prediction errors of BPNN and SVM.** BPNN: back-propagation neural network; LSVM: support vector machine with linear kernel; RSVM: support vector machine with RBF kernel; the prediction error is predicted temperature plus experimental temperature. It is the prediction results of the LOO cross validation.

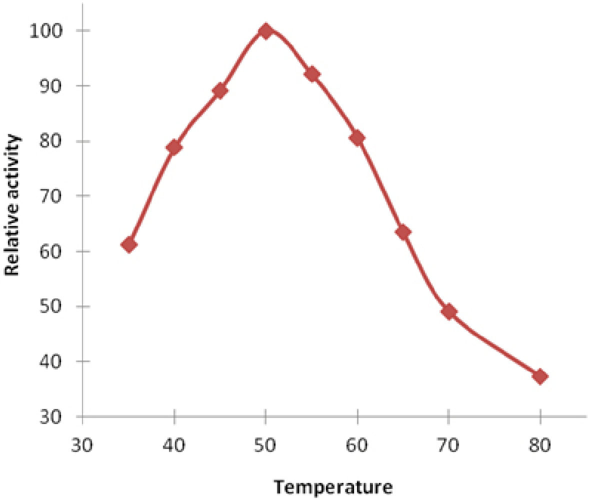

**Fig. 3 The optimal temperature of xylanase.**

## DISCUSSION

Some important parameters (*C*, Ɛ, and γ) had to be optimized during SVM training and testing in order to gain a good predictive performance of SVR model. There was some studies deal with the optimization of running parameters (Xue et al. 2004; Yao et al. 2004; Liu et al. 2005), but all of them fixed two of the parameters and check the curve of RMSE versus the left one to find its optimal value. Often the fixed values of the parameters were selected based on human expertise or even experience. For example, researchers know that too small value of *C* will lead to insufficient stress placed on fitting the training data and too big value of *C* will lead to overfit the training data. But how big is not too small or too big? Different researchers may chose different values of *C* that they think it is not too small or too big. So using uniform design to optimize the parameters would have at least two advantages compared with their methods. Firstly, it allowed much larger searching space of the combination of the parameters and thus the chances of finding the optimal combination of the parameters would increase. Secondly, uniform design was much easier than their methods; it needed only 16 runs for 3 factors and 16 levels, while their methods need 48 runs. This was because of the inherent advantages of uniform design.

Recently two linear models for both single residue and dipeptides and optimum temperature of xylanase in the G/11 family were established based on stepwise regression (Liu et al. 2006).The training RMSEs of their models were 5.03ºC and 1.91ºC, respectively, and they calculated the maximal and minimal optimum temperature of xylanase as 120.84ºC and 10.83ºC. From these results we can conclude that the model we established here was much more accurate. This indicated that the relationship between amino acid composition and the xylanase optimum temperature was very complicated and one might not gain the satisfactory results based on the simple linear models, while SVM is a more powerful tool for prediction of nonlinearities.

Using the crystal information of xylanase, one can pinpoint the residues that may suitable for mutations. Consequently, saturation mutagenesis (where all 20 native amino acids are tested at each pinpointed position) can be applied to generate large, virtual libraries of mutants. Then, our model, for predicting the xylanase optimal temperatures, can be used for pre-screening the virtual libraries. The optimal sequences were chosen based on their predicted optimal temperatures; the mutants were then generated experimentally by mutagenesis and recombination. Therefore, the model can decrease the sequence space, while maintaining broad diversity, to a number easily amenable to experimental screening.

As analyzed above, SVM only showed a minor improvement over BPNN in our study, the large variation (from 0.03ºC to 27.44ºC) in prediction indicated that it should be used with some cautions. At the same time, the MAE of LOO cross-validation was 6.88ºC, and the mean absolute percent error was 12.8%, one can see that is not good enough for directing xylanase engineering. Perhaps further improvement may be achieved by collecting more data sets of higher quality. It should be possible to increase the number of data entries and eliminate the noisy data entries from the updated databases. We think when the MAE of LOO cross-validation was within 5ºC, it may be good enough for directing xylanase engineering and our results were close to this object.

**Financial support:** This work was supported by the National Natural Science Foundation of China (No. 20806031), the Cultivation Project of Huaqiao University for the China National Funds for Distinguished Young Scientists (No.JB-GJ1006) and the Program for New Century Excellent Talents in Universities of Fujian Province (No. 07176C02).

## REFERENCES

BAIROCH, A.; APWEILER, R.; WU, C.H.; BARKER, W.C.; BOECKMANN, B.; FERRO, S.; GASTEIGER, E.; HUANG, H.; LOPEZ, R.; MAGRANE, M.; MARTIN, M.J.; NATALE, D.A.; O'DONOVAN, C.; REDASCHI, N. and YEH, L.S. (2005). The universal protein resource (UniProt). *Nucleic Acids Research*, vol. 33, no. 1, p. D154-D159. **[CrossRef]**

BEG, Q.A.; KAPOOR, M.; MAHAJAN, G. and HOONDAL, S. (2001). Microbial xylanases and their industrial applications: A review. *Applied Microbiology and Biotechnology*, vol. 56, no. 3-4, p. 326-338. **[CrossRef]**

CAI, C.Z.; HAN, L.Y.; JI, Z.L. and CHEN, Y.Z. (2004). Enzyme family classification by support vector machines. *Proteins: Structure, Function, and Bioinformatics*, vol. 55, no. 1, p. 66-76. **[CrossRef]**

CHEN, C.; TIAN, Y.X.; ZOU, X.Y.; CAI, P.X. and MO, J.Y. (2006). Using pseudo-amino acid composition and support vector machine to predict protein structural class. *Journal of Theoretical Biology*, vol. 243, no. 3, p. 444-448. **[CrossRef]**

CHICA, R.A.; DOUCET, N. and PELLETIER, J.N. (2005). Semi-rational approaches to engineering enzyme activity: Combining the benefits of directed evolution and rational design. *Current Opinion in Biotechnology,* vol. 16, no. 4, p. 378-384. **[CrossRef]**

CHOU, K.C. and SHEN, H.B. (2007). Recent progresses in protein subcellular location prediction. *Analytical Biochemistry*, vol. 370, no. 1, p. 1-16. **[CrossRef]**

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Machine Learning*, vol. 20, no. 3, p. 273-297. **[CrossRef]**

DIAZ, M.; RODRIGUEZ, S.; FERNÁNDEZ-ABALOS, J.M.; RIVAS, J.D.L.; RUIZ-ARRIBAS, A.; SHNYROV, V.L. and SANTAMARÍA, R.I. (2004). Single mutations of residues outside the active center of the xylanase Xys1Δ from *Streptomyces halstedii* JM8 affect its activity. *FEMS Microbiology Letters*, vol. 240, no. 2, p. 237-243. **[CrossRef]**

FANG, K.T. (1980). The uniform design: Application of number-theoretic methods in experimental design. *Acta Mathematicae Applicatae Sinica,* vol*.* 3, p. 363-372.

FANG, K.T. and YANG, Z.H. (2000). On uniform design of experiments with restricted mixtures and generation of uniform distribution on some domains. *Statistics and Probability Letters*, vol. 46, no. 2, p. 113-120. **[CrossRef]**

FENEL, F.; ZITTING, A.J. and KANTELINEN, A. (2006). Increased alkali stability in *Trichoderma reesei* endo-1, 4-β-xylanase II by site directed mutagenesis. *Journal of Biotechnology,* vol. 121, no. 1, p. 102-107. **[CrossRef]**

FRANK, E.; HALL, M.; TRIGG, L.; HOLMES, G. and WITTEN, I.H. (2004). Data mining in bioinformatics using *Weka*. *Bioinformatics*, vol. 20, no. 15, p. 2479-2481. **[CrossRef]**

FU, X.P.; WANG, W.Y. and ZHANG, G.Y. (2012). Construction of an expression vector with elastin-like polypeptide tag to purify the xylanase with non-chromatographic method. *Acta Microbiologica Sinica*. In press.

HAYES, R.J.; BENTZIEN, J.; ARY, M.L.; HWANG, M.Y.; JACINTO, J.M.; VIELMETTER, J.; KUNDU, A. and DAHIYAT, B.I. (2002). Combining computational and experimental screening for rapid optimization of protein properties. *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, no. 25, p. 15926-15931. **[CrossRef]**

LIANG, Y.Z.; FANG, K.T. and XU, Q.S. (2001). Uniform design and its applications in chemistry and chemical engineering. *Chemometrics and Intelligent Laboratory Systems,* vol. 58, no. 1, p. 43-57. **[CrossRef]**

LIU, H.X.; YAO, X.J.; XUE, C.X.; ZHANG, R.S.; LIU, M.C.; HU, Z.D. and FAN, B.T. (2005). Study of quantitative structure-mobility relationship of the peptides based on the structural descriptors and support vector machines. *Analytica Chimica Acta*, vol. 542, no. 2, p. 249-259. **[CrossRef]**

LIU, L.; DONG, H.; WANG, S.; CHEN, H. and SHAO, W. (2006). Computational analysis of di-peptides correlated with the optimal temperature in G/11 xylanase. *Process Biochemistry*, vol. 41, no. 2, p. 305-311. **[CrossRef]**

MILDVAN, A.S. (2004). Inverse thinking about double mutants of enzymes. *Biochemistry*, vol. 43, no. 46, p. 14517-14520. **[CrossRef]**

MOREAU, A.; SHARECK, F.; KLUEPFEL, D. and MOROSOLI, R. (1994). Increase in catalytic activity and thermostability of the xylanase A of *Streptomyces lividans* 1326 by site-specific mutagenesis. *Enzyme and Microbial Technology,* vol. 16, no. 5, p. 420-424. **[CrossRef]**

OLIVEIRA, L.A.; PORTO, A.L.F. and TAMBOURGI, E.B. (2006). Production of xylanase and protease by *Penicillium janthinellum* CRC 87m-115 from different agricultural wastes. *Bioresource Technology*, vol. 97, no. 6, p. 862-867. **[CrossRef]**

SMOLA, A.J. and SCHÖLKOPF, B. (1998). *A Tutorial on Support Vector Regression*. NeuroCOLT Technical Report NC-TR-98-030. Royal Holloway College, University of London, UK.

VAPNIK, V. (1998). *Statistical learning theory*. John Wiley and Sons, New York. 740 p. ISBN 0-471-03003-1.

WARD, J.J.; MCGUFFIN, L.J.; BUXTON, B.F. and JONES, D.T. (2003). Secondary structure prediction with support vector machines. *Bioinformatics*, vol. 19, no. 13, p. 1650-1655. **[CrossRef]**

XUE, C.X.; ZHANG, R.S.; LIU, H.X.; LIU, M.C.; HU, Z.D. and FAN, B.T. (2004). Support vector machines-based quantitative structure-property relationship for the prediction of heat capacity. *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 4, p. 1267-1274. **[CrossRef]**

YAO, X.J.; PANAYE, A.; DOUCET, J.P.; ZHANG, R.S.; CHEN, H.F.; LIU, M.C.; HU, Z.D. and FAN, B.T. (2004). Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regressions. *Journal of Chemical Information and Modeling*, vol. 44, no. 4, p. 1257-1266. **[CrossRef]**

ZHANG, L.; LIANG, Y.Z.; JIANG, J.H.; YU, R.Q. and FANG, K.T. (1998). Uniform design applied to nonlinear multivariate calibration by ANN. *Analytica Chimica Acta*, vol. 370, no. 1, p. 65-77. **[CrossRef]**

## How to reference this article:

ZHANG, G. and GE, H. (2012). Prediction of xylanase optimal temperature by support vector regression. *Electronic Journal of Biotechnology*, vol. 15, no. 1. http://dx.doi.org/10.2225/vol15-issue1-fulltext-8

**Note:** Electronic Journal of Biotechnology is not responsible if on-line references cited on manuscripts are not available any more after the date of publication. Supported by UNESCO / MIRCEN network.