Research Article   **Open Access**

Prediction of Michaelis-Menten Constant in Beta-Cellobiosidase's Reaction with Lactoside as Substrate

Shaomin Yan[1] and Guang Wu[1,2]*

[1]State Key Laboratory of Non-food Biomass Enzyme Technology, National Engineering Research Center for Non-food Biorefinery, Guangxi Key Laboratory of Biorefinery, Guangxi Academy of Sciences, 98 Daling Road, Nanning, Guangxi, 530007, China
[2]DreamSciTech Consulting, 301, Building 12, Nanyou A-zone, Jiannan Road, Shenzhen, Guangdong, 518054, China

Abstract

The Michaelis-Menten constant, Km, is important to understand the characteristics of enzyme and its relationship with substrates and numerous conditions in biochemical reactions. Although the fast development is evidenced in enzymatic research, the Km value in each enzyme under various conditions still needs to be measured individually. On the other hand, the modern computational techniques and bioinformatics provide the opportunity to theoretically predict Km in enzyme with different substrates under various conditions. Cellulose 1,4-beta-cellobiosidase is an enzyme used in cellulose hydrolysis for bio-fuel industry, and huge efforts are made to enhance its efficiency through searching for new strains of beta-cellobiosidase as well as enzymatic engineering. Therefore it is considered important to develop methods to predict the Km value in beta-cellobiosidase's reaction. In this study, the information of amino acid properties in beta-cellobiosidase, pH and temperature in reaction, and lactoside as substrate were chosen as predictors to predict the Km values by feedforward backpropagation neural networks, and the delete-1 jackknife was used to validate the predictive model. The results show that 11 of 25 scanned amino acid properties could act as predictors, and that the amino-acid distribution probability appeared the best predictor. The two-layer structure of neural network configuration was sufficient for initial scanning. In consistent with previous studies, the Km value of enzymatic reactions was predictable using enzyme sequence information and reaction conditions with neural network models.

Keywords: Beta-cellobiosidase; Lactoside; Km value; Prediction

Introduction

In biochemical reactions, an important measure related to enzymatic function is the Michaelis-Menten constant, Km, because other measures such as pH, temperature, substrate concentrations, are mainly related to reactive conditions. Thus, the Km value is important to understand the characteristics of enzyme and its relationship with substrates and numerous conditions in biochemical reactions [1,2].

Actually, Km is the only parameter, from which the enzymatic kinetics as well as stimulators and inhibitors to enzymatic reactions can be formulated [3,4]. Therefore, it is important not only from a practical viewpoint but also from a theoretical viewpoint. However, the measurement of Km is performed case by case [5], and the measured value is difficult to extrapolate to the enzymes under same category. Yet, Km is also important to describe the absorption process, i.e. active absorption process [6], for which the Km value is measured individually too. Technically, Km is directly related to the affinity of enzyme to a certain substrate.

Without costly and time-consuming measurement, no Km values are available for newly found and designed enzymes. Therefore, it is necessary to develop methods to predict Km based on simple information for each enzyme before conducting costly experiments. Along this research line, several studies have been carried out very recently [7-11]; however, more studies are needed in order to systematically approach this issue from various angles.

Cellulose 1,4-β-cellobiosidase (EC 3.2.1.91) hydrolyzes 1,4-β-D-glucosidic linkages in cellulose and cellotetraose, and then releases cellobiose from the non-reducing ends of the chains. Recently, a new interest was directed to cellobiosidase because of its potential role in bio-fuel industry, meanwhile lactoside is a major substrate in biochemical reaction of β-cellobiosidase. In this study, the information of amino acid properties in beta-cellobiosidase, pH and temperature in reaction, and lactoside as substrate were chosen as predictors to predict the Km values by means of neural network in order to develop the predictive model.

Materials and Methods

Data

The Km values related to cellulose 1,4-β-cellobiosidases (EC 3.2.1.91) with lactoside as substrate were found in the Comprehensive Enzyme Information System BRENDA [13]. Up to May 2011, 5 β-cellobiosidases had their sequence information under the category of Km value as functional parameter, of which β-cellobiosidases P62694 and Q8J0K6 were documented with their mutants [14-18]. Still, each cellulose 1,4-β-cellobiosidase could have different Km values regarding different catalytic conditions, such as pH, temperature and substrate [7-12]. In total, this databank provided 38 matched sequences and Km values of β-cellobiosidases (Supplementary Data). The amino-acid sequences of β-cellobiosidases were obtained from the UniProt [19].

Predictors

The information of amino acid properties was mainly obtained from AAIndex [20], which contains 540-plus amino-acid properties

with redundancy [21,22], so not all documented amino acid properties were used in this study, but the ones screened in the previous studies [7-12]. Those amino acid properties included amino acid charge, hydrophilicity or hydrophobicity, size and functional groups [23], such as the spatial properties [24,25], the hydrophobic properties [26-28], the electronic properties [29], and the secondary structure predictions [30]. Each of those amino acid properties had numerically constant value for a type of amino acid Supplementary Data; therefore they were not sensible to amino aid composition, location in enzyme, etc.

Amino Acid	Field effect index		Amino-acid number		Distribution probability	
	A7WNT9	A7WNU1	A7WNT9	A7WNU1	A7WNT9	A7WNU1
A	0.05	0.05	34	44	0.0028	0.0031
R	0.27	0.27	10	13	0.1143	0.0441
N	-0.56	-0.56	30	35	0.0269	0.0291
D	-1.77	-1.77	33	31	0.0053	0.0014
C	0.06	0.06	24	26	0.0091	0.0198
E	-1.14	-1.14	13	14	0.0617	0.0687
Q	-0.35	-0.35	23	21	0.0173	0.0243
G	0	0	61	60	0.0067	0.0154
H	-0.58	-0.58	6	6	0.3472	0.2315
I	0.04	0.04	16	14	0.0795	0.0011
L	-0.03	-0.03	26	22	0.0247	0.0033
K	0.51	0.51	21	23	0.0270	0.0101
M	-0.3	-0.3	10	15	0.1905	0.0011
F	-0.45	-0.45	16	17	0.0341	0.1280
P	0.02	0.02	26	25	0.0115	0.0053
S	-0.38	-0.38	50	47	0.0008	0.0013
T	-0.44	-0.44	64	55	0.0003	0.0022
W	-0.24	-0.24	11	10	0.0135	0.1905
Y	-0.42	-0.42	23	22	0.0222	0.0559
V	-0.04	-0.04	26	32	0.0073	0.0226

The amino-acid distribution probability, is computed according to the equation, $r!/(q_0! \times q_1! \times ... \times q_n!) \times r!/(r_1! \times r_2! \times ... \times r_n!) \times n^{-r}$, where "!" is the factorial function, r is the number of a type of amino acid, q is the number of partitions with the same number of amino acids and n is the number of partitions in the protein for a type of amino acid [31]. The computation can be found in the web site http://www.dream-scitech.com/Web-Based-Computation/ADP.htm, 2011.

Table 1: Field effect index, amino-acid number and distribution probability in β-cellobiosidase A7WNT9 and A7WNU1.
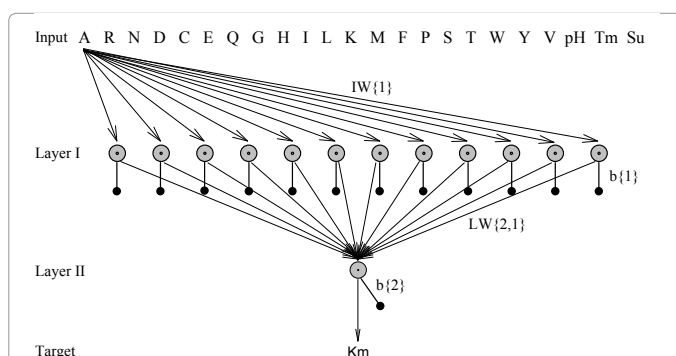

Figure 1: 12-1 feed forward back propagation neural network to model 23 predictors and *Km*. Each gray circle presents a neuron. A, alanine; R, arginine; N, asparagine; D, aspartic acid; C, cysteine; E, glutamic acid; Q, glutamine; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; M, methionine; F, phenyl-alanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine; pH, pH value; Tm, temperature; Su, substrate; IW{1}, the input weights; LW{2,1}, the layer weights to the second layer from the first layer; b{1} and b{2}, the biases related to each neuron at the first and second layers.
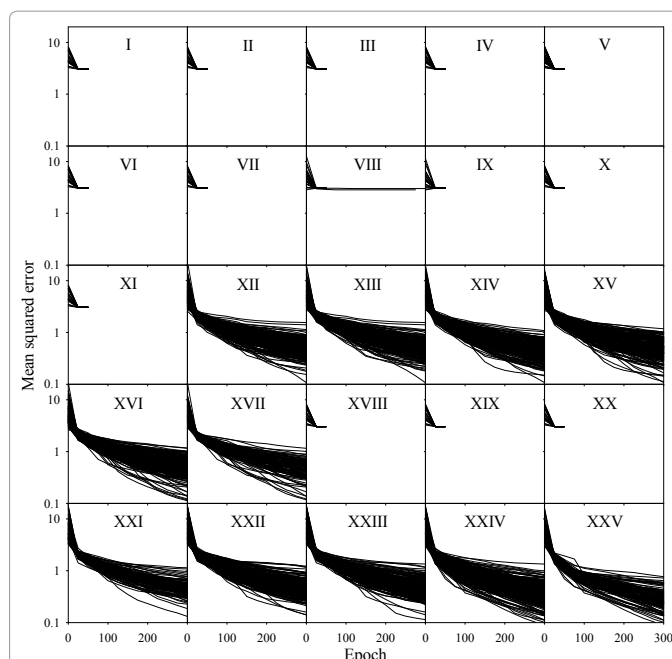

Figure 2: Convergence of mean squared error performance function with 100 different initial weights and biases generated by random initialization function in 12-1 neural network with different amino acid properties.

On the other hand, an amino acid property, which outperformed those amino acid properties from AAIndex in previous studies [7-12], was the amino-acid distribution probability [31,32]. This property does not have a constant value for each type of amino acids, but is subject to the length of enzyme and position of each amino acid. Table 1 showed the difference between the field effect index and the amino-acid distribution probability for two β-cellobiosidases.

The predictors of pH and temperature were measured values in database [13], and the predictor of substrate, lactoside, was different with respect to its substituent groups. So there were 23 predictors for predicting the *Km* values.

Predictive Model

The previous studies showed that the neural network could be the best model for the prediction [7-12], because the relationship between predictors and *Km* value is not readily known, while the neural network can theoretically model either cause-consequence relationship or phenomenological relationship. However, the neural network model used in previous studies appeared very complicated, thus the application of simple neural network model without compromising its predictive ability was the main task, which meant to determine how many layers and how many neurons work better.

The developed predictive model was validated using the delete-1 jackknife validation as used in previous studies [7-12] because it was considered very powerful for this type of studies [33].

Results and Discussion

Using a 12-1 feedforward backpropagation neural network as an example, Figure 1 showed the predictive model for predicting the *Km* value. This predictive model was different from current models in bioinformatic studies that included as many predictors as possible.

Theoretically, a predictive model would be chosen with as fewer predictors as possible [34,35], which was the approach used in this study.

Figure 2 showed the convergence with each predictor. This was an important step during model development, because Figure 2 gave a clear picture of which predictor could not converge, which were predictors I, II, III, IV, V, VI, VII, VIII, IX, X, XI, XVIII, XIX, and XX (Supplementary Data). Nevertheless, the predictors that could not converge would not be useful for the prediction. On the other hand, whether a predictor converged guaranteed whether correct model parameters could be obtained [36,37]. Therefore Figure 2 played a role to select workable predictors.

Figure 3 showed the performance of predictions obtained by using the selected predictors from Figure 2 in terms of P value, which was the statistical difference between predicted and measured Km values, and $R2$ value, which was the squared correlation coefficient between predicted and measured Km values. Actually, Figure 3 played role to determine number of neurons ranged from 1-1 to 20-1 for the first layer in two-layer neural networks. In general, the P value and $R2$ increased in predictions using 1-1 to 3-1 neurons, and some predictors got the
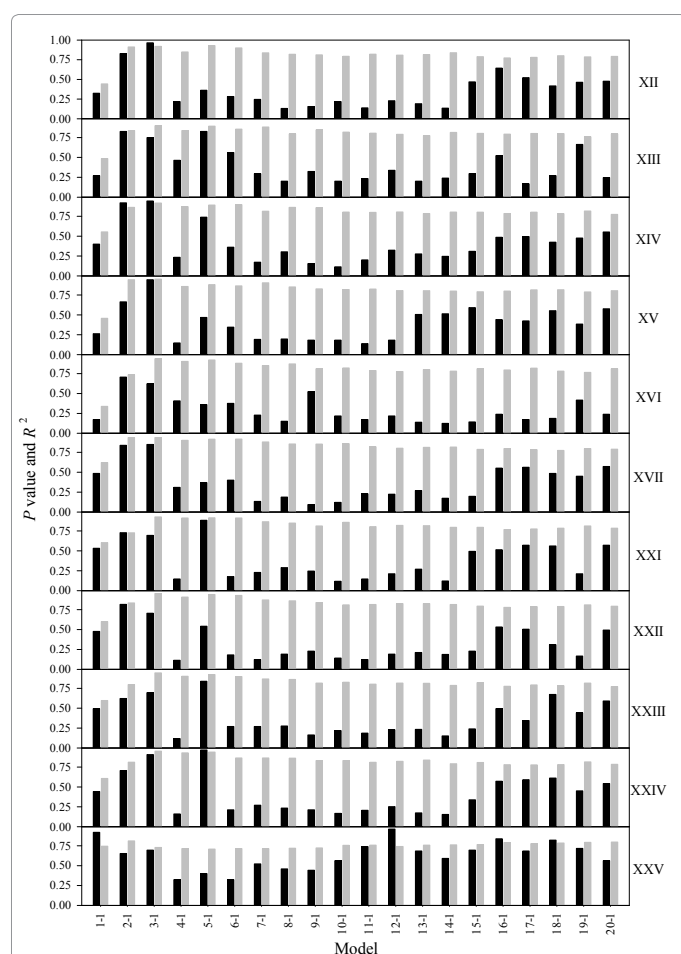
Figure 4: Comparison between recorded and mean predicted Km in different 3-1 models for fitting and delete-1 jackknife validation. Black bars, P value obtained from Wilcoxon signed rank test; gray bars, squared correlation coefficient in regression.

Figure 3: Comparison between recorded and mean predicted Km in different fitting models. Black bars, P value obtained from Wilcoxon signed rank test; gray bars, squared correlation coefficient in regression.
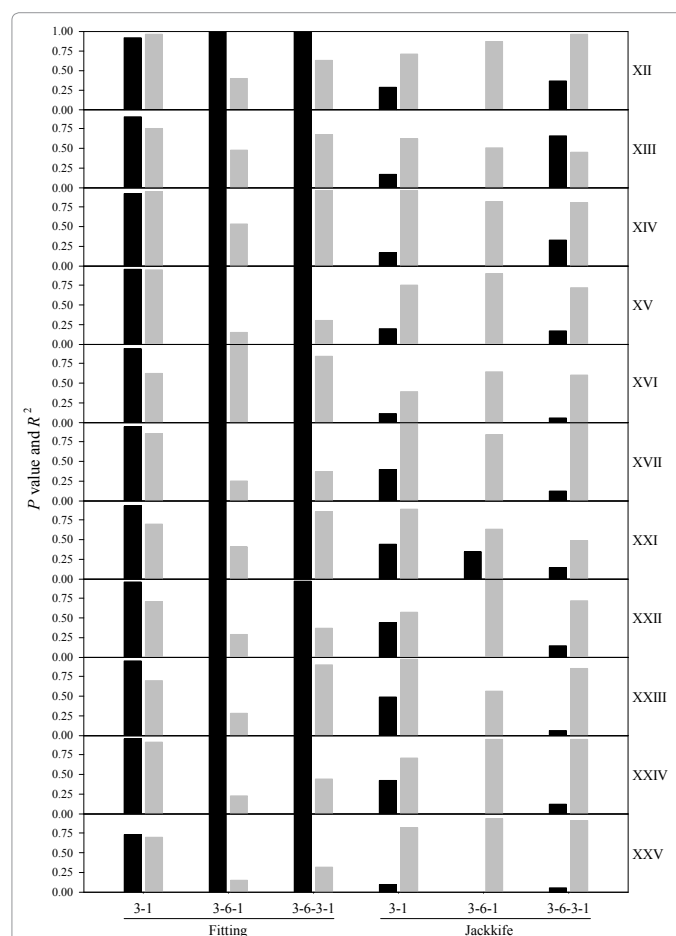
highest values in predictions using 5-1 neurons. Thereafter, the $R2$ values were stable while the P values changed as the neuron increased. Among 11 predictors, the last one (XXV) that was the amino-acid distribution probability provided better predicting results than others.

Figure 4 showed the performance of predictions obtained from 11 selected predictors in different 3-1 models, which reflected another aspect of model selection, i.e. how many layers were suitable for a predictive model. As Figure 2 chose predictors and Figure 3 chose the number of neurons in a neural network, so thus Figure 4 was a necessary step for model development. Also, the delete-1 jackknife validation was applied because more elaborations became possible with the narrowing of searching range for model and predictor selections. As can be seen, multiple layers did not reveal remarkable improvement for predictions.

The predictions of 11 selected predictors were conducted in different 12-1 models for fitting (Figure 5) and validation (Figure 6), to further evaluate the influence of multi-layer models on the prediction performance, and to answer whether a very sophisticated model could improve the predictions. Generally speaking, the P values were higher obtained from fitting and the $R2$ values higher obtained from validation, but no clear feature could be drawn as the model layers increased.
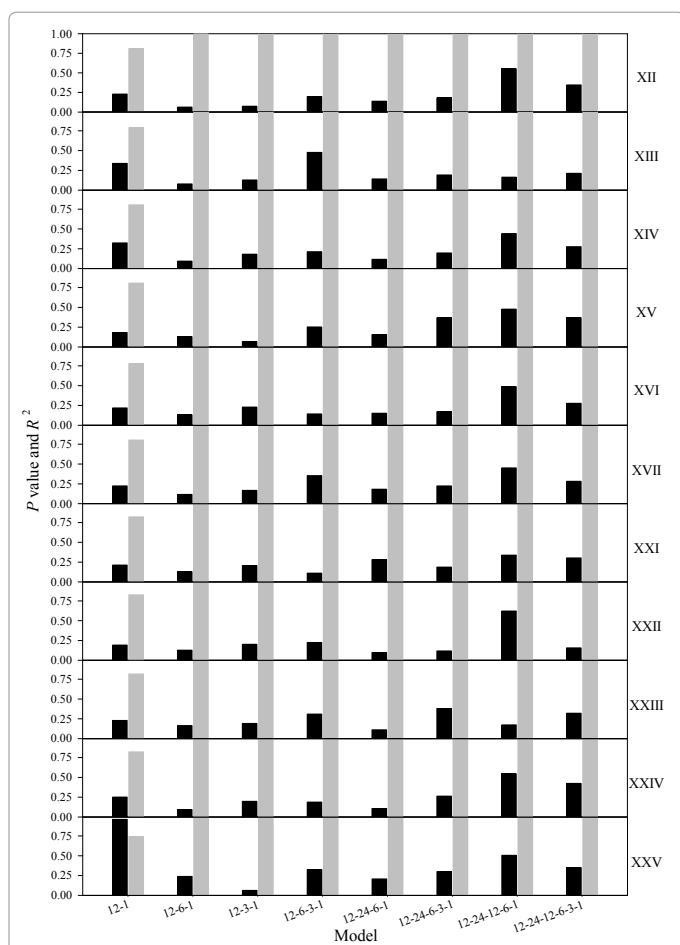
Figure 5: Comparison between recorded and mean predicted *Km* in different 12-1 models for fitting. Black bars, *P* value obtained from Wilcoxon signed rank test; gray bars, squared correlation coefficient in regression.
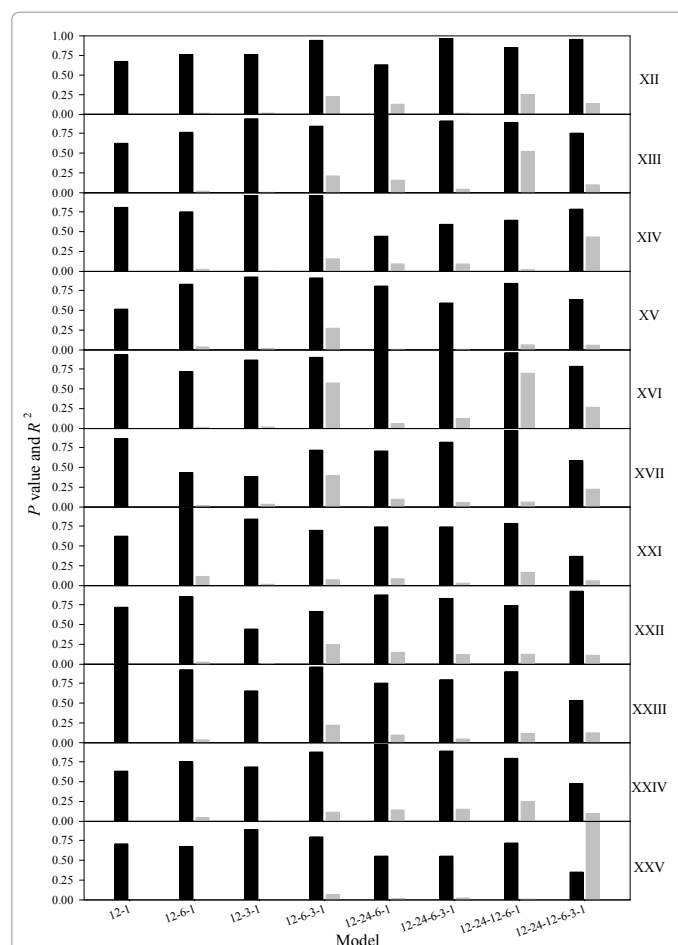

Figure 6: Comparison between recorded and mean predicted *Km* in different 12-1 models for delete-1 jackknife validation. Black bars, *P* value obtained from Wilcoxon signed rank test; gray bars, squared correlation coefficient in regression.

Currently, considerable data are available for various bioinformatic models, but unfortunately few data are available related to parameters of enzymatic reactions. Therefore a small dataset was used in this study although they were all the data available in literature. This is the pressing point that the methods for the prediction of enzyme function parameters should be developed.

This study advanced our knowledge on the prediction of *Km* values not only in view of the amino acid properties in enzymes as predictors, but also in view of pH, temperature and substrate in enzymatic reaction as predictors, whereas previous studies included only the amino acid properties as predictors [8,10].

In conclusion, the results demonstrated that the *Km* value of cellulose 1,4-beta-cellobiosidases could be predicted using the neural network models with their sequence information and reaction conditions. Eleven of 25 scanned amino acid properties could act as the predictors, among which the amino-acid distribution probability appeared the best predictor, and the two-layer structure of neural network configuration was sufficient for initial scanning.

Acknowledgment

References

1. Hutzler MJ, Linder CD, Melton RJ, Vincent J, Daniels SJ (2010) In vitro-in vivo correlation and translation to the clinical outcome for CJ-13,610, a novel inhibitor of 5-lipoxygenase. Drug Metab Dispos 38: 1113-1121.

2. Rokitta D, Fuhr U (2010) Comparison of enzyme kinetic parameters obtained in vitro for reactions mediated by human CYP2C enzymes including major CYP2C9 variants. Curr Drug Metab 11: 153-161.

3. Dong JQ, Salinger DH, Endres CJ, Gibbs JP, Hsu CP, et al. (2011) Quantitative prediction of human pharmacokinetics for monoclonal antibodies: retrospective analysis of monkey as a single species for first-in-human prediction. Clin Pharmacokinet 50: 131-142.

4. Liu L, Halladay JS, Shin Y, Wong S, Coraggio M, et al. (2011) Significant species difference in amide hydrolysis of GDC-0834, a novel potent and selective Bruton's tyrosine kinase inhibitor. Drug Metab Dispos 39: 1840-1849.

5. Jałowiecki P, Janasik B (2007) Physiologically-based toxicokinetic modeling of durene (1,2,3,5-tetramethylbenzene) and isodurene (1,2,4,5-tetramethylben-zene) in humans. Int J Occup Med Environ Health 20: 155-165.

6. Tachibana T, Kato M, Sugiyama Y (2011) Prediction of Nonlinear Intestinal Absorption of CYP3A4 and P-Glycoprotein Substrates from their In Vitro *Km* Values. Pharm Res.

7. Yan S, Wu G (2011) Searching of predictors to predict pH of cellulases. Appl Biochem Biotechnol 165: 856-869.

8. Yan SM, Wu G (2011) Prediction of Michaelis-Menten constant of beta-glucosidases using nitrophenyl-beta-D-glucopyranoside as substrate. Protein Pept Lett 18: 1053-1057.

9. Yan S, Shi D, Nong H, Wu G (2011) Simultaneously predicting pH and temperature optimum in catalytic reaction of beta-glucosidase. Guangxi Sci 18: 253-260.

10. Yan S, Shi D, Nong H, Wu G, Predicting *Km* values of beta-glucosidases using cellobiose as substrate. Interdiscip Sci: Comput Life Sci.

11. Yan S, Wu G. Prediction of turnover number of cellobiohydrolase. Protein Pept Lett (in press)

12. Yan S, Wu G (2011) Prediction of optimal pH and temperature of cellulases using neural network. Protein Pept Lett (Epub ahead of print)

13. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, et al (2011) BRENDA, the enzyme information system in 2011. Nucleic Acids Res 39: 670-676.

14. Becker D, Braet C, Brumer H 3rd, Claeyssens M, Divne C, et al. (2001) Engineering of a glycosidase Family 7 cellobiohydrolase to more alkaline pH optimum: the pH behaviour of Trichoderma reesei Cel7A and its E223S/A224H/L225V/T226A/D262G mutant. Biochem J 356: 19-30.

15. Von Ossowski I, Ståhlberg J, Koivula A, Piens K, Becker D, et al. (2003) Engineering the exo-loop of Trichoderma reesei cellobiohydrolase, Cel7A. A comparison with Phanerochaete chrysosporium Cel7D. J Mol Biol 333: 817-829.

16. Stahlberg J, Nerinckx W, Teeri TT, Claeyssens M (2004) Structure-reactivity studies of Trichoderma reesei cellobiohydrolase Cel7A. Am Chem Soc Symp Ser 889: 207-226.

17. Voutilainen SP, Puranen T, Siika-Aho M, Lappalainen A, Alapuranen M, et al. (2008) Cloning, expression, and characterization of novel thermostable family 7 cellobiohydrolases. Biotechnol Bioeng 101: 515-528.

18. Voutilainen SP, Boer H, Alapuranen M, Jänis J, Vehmaanperä J, et al. (2009) Improving the thermostability and activity of Melanocarpus albomyces cellobiohydrolase Cel7B. Appl Microbiol Biotechnol 83: 261-272.

19. The UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: 142-148.

20. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, et al. (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36: 202-205.

21. Yang XY, Shi XH, Meng X, Li XL, Lin K, et al. (2010) Classification of transcription factors using protein primary structure. Protein Pept Lett 17: 899-908.

22. Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci USA 102: 6395-6400.

23. Burlingame AL, Carr SA (1996) Mass Spectrometry in the Biological Sciences. Humana Press, Totowa, NJ.

24. Zamyatin AA (1972) Protein volume in solution. Prog Biophys Mol Biol 24: 107-123.

25. Darby NJ, Creighton TE (1993) Dissecting the disulphide-coupled folding pathway of bovine pancreatic trypsin inhibitor. Forming the first disulphide bonds in analogues of the reduced protein. J Mol Biol 232: 873-896.

26. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105-132.

27. Trinquier G, Sanejouand YH, Hausman RE (1998) Which effective property of amino acids is best preserved by the genetic code? Protein Eng 11: 153-169.

28. Cooper GM (2004) The cell: A Molecular Approach. ASM Press, Washington.

29. Dwyer DS (2005) Electronic properties of amino acid side chains: quantum mechanics calculation of substituent effects. BMC Chem Biol 5: 2

30. Chou PY, Fasman GD (1978) Prediction of secondary structure of proteins from amino acid sequence. Adv Enzymol Relat Subj Biochem 47: 45-148.

31. Feller W (1968) An Introduction to Probability Theory and Its Applications. (3rd edn), Wiley, New York.

32. Wu G, Yan S (2008) Lecture Notes on Computational Mutation. Nova Science Publishers, New York.

33. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273: 236-247.

34. Akaike H. (1974) A new look at the statistical model identification. IEEE Trans Automatic Control 19: 716-723.

35. Akaike information criterion

36. Demuth H, Beale M (2001) Neural network toolbox for use with MatLab. User's guide, version 4.

37. MathWorks Inc (2001) MatLab - The Language of Technical Computing.