Prediction of Temperature Optimum in Enzymatic Reaction of Beta-Cellobiosidases with Exhausted Jackknife Validation

Shaomin Yan, Guang Wu*

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
* hongguanglishibahao@yahoo.com

Abstract: Beta-cellobiosidases are enzymes playing an important role in modern industry, but many parameters related to their reactions are poorly documented. With increased interests in beta-cellobiosidases in bio-fuel industry, the prediction of parameters in enzymatic reactions has been listed on agenda. During the development of predictive model, the data were usually divided into two datasets, one was for model development and the other for model validation. The widely used validation method was the delete-1 jackknife validation. However, no systematical studies were conducted to determine whether jackknife validation with other deletion works better, because the numbers of validations with different deletions are increasing in a factorial fashion. Therefore, only small dataset can be used for such an exhausted jackknife validation. In this study, two aims were defined: (i) which amino acid property works better to predict temperature optimum of beta-cellobiosidases and (ii) with which deletion jackknife validation works better. The results show that the amino acid distribution probability works better in predicting the temperature optimum and the delete-1 jackknife validation works better.
[Shaomin Yan, Guang Wu. **Prediction of Temperature Optimum in Enzymatic Reaction of Beta-Cellobiosidases with Exhausted Jackknife Validation**. *Life Sci J* 2013;10(3):2180-2189]. (ISSN: 1097-8135). http://www.lifesciencesite.com 321

Keywords: beta-cellobiosidase; enzyme; jackknife validation; prediction; temperature optimum

1. Introduction

With the development, environmental friendly bio-fuel has become a common consensus for sustainable development, thus enzymes play very important roles in bio-fuel production through enzymatic reactions. Nevertheless, optimization of enzymatic reaction is a key factor to manufacture bio-fuel's efficiencies and industrial scale's effectiveness. Technically, optimization is available via monitoring and controlling a number of parameters, which represent in different aspects of enzymatic reactions. So far many experiments have been conducted to determine these parameters at optimum in enzymatic reactions. No doubt that the measured values of these parameters documented in literatures and patents can significantly reduce the labor, which will be spent on enzymatic reactions with new available enzymes. However, the gap between the number of measured values of enzymatic parameters and the number of available enzymes is widening because modern sequencing and engineering techniques discovered many new enzymes and modified many inefficient enzymes. In this context, the development of workably predictive model for predicting parameters in enzymatic reactions draws our attention [1-8].

Cellulose 1,4-β-cellobiosidases (EC 3.2.1.91) hydrolyze 1,4-β-D-glucosidic link in cellulose and cellotetraose, and then release cellobiose from the non-reducing ends of the chains. Recently, cellobiosidases has caused a new interest because of their potential role in bio-fuel industry [9, 10]. Heterologous expression of a beta-glucosidase gene has been conducted to improve the efficiency of enzyme's ability of decomposing cellulose [11-13]. Of various parameters in enzymatic reactions, we especially interested in the temperature optimum, because temperature can directly affect enzyme's mobility, fold, and compactness, and resulting in unfolding of Th Cel7A just above the optimum temperature limit [14]. Also, we had already conducted two studies on predicting Michaelis-Menten constant and turning over number of cellulose 1, 4-β-cellobiosidases [7, 8].

On the other hand, the development of predictive models generally goes through a two-step process, (i) to find out which model is suited to the relationship between predictors and predicted values, and (ii) to validate the developed model. Technically, the data included both measured values of predictors and predicted ones. They were not divided for the first step of model development, whereas the data would be divided into two groups for the second step of model development, i.e., one is used to develop predictive models and another to validate the models. However, the issue of how to divide the database into two groups has resulted in several different validation methods, among which the jackknife validation is

considered powerful [15].

A general practice is to use the delete-1 jackknife validation, which divides the data into two groups: one contains a single datum for validation while the other contains the rest of data for model development. In fact, the delete-1 jackknife validation appears to be the only one was used in real-life although theoretically it can be other jackknife validations, such as delete-2 jackknife, delete-3 jackknife, etc. This raises a question that which jackknife validation would work the best, however, few studies were done about this [3], because of the fact that the number of validations would increase with factorial factor [16]. Consequently, exhausted jackknife validations can only be applied to a very small data, which actually is the case for enzymes that have measured and documented parameters in enzymatic reactions.

Hence, this study has two aims: (i) to find out which property of amino acids is more useful for predicting the temperature optimum of beta-cellobiosidases, and (ii) to conduct exhausted jackknife validations in order to determine which one works the best.

2. Materials and Methods

2.1. Data

The information related to enzymatic reaction of beta-cellobiosidase (EC 3.2.1.91) was found in the Comprehensive Enzyme Information System BRENDA [6] and corresponding amino acid sequences were obtained from the Universal Protein Resource [17]. Under the functional parameter of temperature optimum, only 20 beta-cellobiosidases have their sequence information. Such a lack of information on beta-cellobiosidases in enzyme databank gives us a strong motivation to develop methods to predict the parameters in enzymatic reactions.

2.2. Predictors

The AAindex contains more than 540 amino acid properties [18], however, of which many have been considered abundant [19]. Therefore, we choose 23 properties that related to amino acid charge, hydrophilicity or hydrophobicity, size and functional groups, which are important indicators for protein structure and protein–protein interactions [20], including the spatial properties [21, 22], hydrophobic properties [23-25], electronic properties [26], and the secondary structure predictions [27]. Actually, all 540-plus amino acid properties [18] reflect the characteristic of individual amino acid, and thus they are constants, i.e., each type of amino acid has a certain value for a given property (The 4th and 5th columns in Table 1). Clearly these properties lack information for a whole protein, so we use the amino

Table 1. Difference between $H_M\Delta PH$ and amino acid distribution probability with respect to β-cellobiosidases P62694 and Q5S1P9

Amino Acid	Number		$H_M\Delta PH$		$H_M\Delta PH \times$ Number		Distribution probability	
	P62694	Q5S1P9	P62694	Q5S1P9	P62694	Q5S1P9	P62694	Q5S1P9
A	34	45	0.05	0.05	1.70	2.25	0.0020	0.0048
R	12	12	-0.75	-0.75	-9.00	-9.00	0.1862	0.1163
N	33	27	-0.2	-0.2	-6.60	-5.40	0.0274	0.0278
D	24	36	1.8	1.8	43.20	64.80	0.0285	0.0279
C	24	22	-0.01	-0.01	-0.24	-0.22	0.0151	0.0325
E	19	13	1.25	1.25	23.75	16.25	0.0559	0.0463
Q	23	22	-0.07	-0.07	-1.61	-1.54	0.0396	0.0071
G	63	58	0	0	0.00	0.00	0.0012	0.0066
H	5	7	0.21	0.21	1.05	1.47	0.2880	0.3213
I	12	14	0.08	0.08	0.96	1.12	0.0310	0.1649
L	28	26	0.07	0.07	1.96	1.82	0.0375	0.0008
K	14	19	-1.11	-1.11	-15.54	-21.09	0.0618	0.1118
M	7	12	-0.04	-0.04	-0.28	-0.48	0.1071	0.0013
F	16	18	0.06	0.06	0.96	1.08	0.0213	0.0831
P	26	21	0.1	0.1	2.60	2.10	0.0330	0.0707
S	57	53	-0.05	-0.05	-2.85	-2.65	0.0079	0.0084
T	58	63	-0.03	-0.03	-1.74	-1.89	0.0062	0.0000
W	9	9	0.15	0.15	1.35	1.35	0.1475	0.1967
Y	25	24	0.02	0.02	0.50	0.48	0.0446	0.0714
V	24	28	0.09	0.09	2.16	2.52	0.0107	0.0125

$H_M\Delta PH$ is normalized Mulliken population data for the amino acid side chains in the context of phenol [24]. The amino acid distribution probability is computed according to the following equation: $n!/(q_0!\times q_1!\times\ldots\times q_n!)\times r!/(r_1!\times r_2!\times\ldots\times r_n!)\times n^{-r}$, 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, n is the number of partitions in the protein for a type of amino acid, and its web calculation can be available at http://www.nerc-nfb.ac.cn/calculation/dp.htm

acid composition of each beta-cellobiosidase to time amino acid properties in order that these properties can include some information on whole protein (The 6th and 7th columns in Table 1).

A property, which combines both individual amino acid characteristic and whole protein characteristic, is the amino acid distribution probability (see reviews and textbook [28-32]). This property can be referred to the statistical mechanics, where the distribution of elementary particles in energy states can be classified according to three assumptions with respect to whether or not to distinguish each particle

Table 2. Amino acids properties to be scanned as possible predictors for predicting temperature optimum of β-cellobiosidases

Amino acid	A	R	N	D	C	E	Q	G	H	I
Mass, Dalton	71.09	156.19	115.09	114.11	103.15	129.12	128.14	57.05	137.14	113.16
Surface Area, Å^2	115	225	150	160	135	190	180	75	195	175
Residue Volume, Å^3	88.6	173.4	114.1	111.1	108.5	138.4	143.8	60.1	153.2	166.7
van der Waals volume, Å^3	67	148	96	91	86	114	109	48	118	124
Residue Non-polar Surface Area, Å^2	47	86	135	155	164	124	48	137	39+155	37+199
Residue Burial, kcal/mol	1.18	2.15	3.38	3.88	4.1	3.1	1.2	3.43	3.46	4.11
Side Chain Burial, kcal/mol	0	1	2.2	2.7	2.9	1.9	0	2.3	2.3	2.9
Hydropathy index	1.8	4.5	-3.5	-3.5	2.5	-3.5	-3.5	-0.4	-3.2	4.5
Ranking of amino acid polarities	9	15	16	19	7	17	18	11	10	1
pK_a	9.69	9.04	8.8	9.6	10.28	9.67	9.13	9.6	9.17	9.68
σ_I	0.05	-0.26	-0.14	0.51	-0.01	0.68	-0.1	0	-0.01	0.06
$H_M\Delta PH$	0.05	-0.75	-0.2	1.8	-0.01	1.25	-0.07	0	0.21	0.08
σ_R	0	-0.49	-0.06	1.29	0.01	0.57	0.03	0	0.22	0.02
σ_α	-0.01	-0.08	-0.04	-0.03	-0.03	-0.04	-0.05	0	-0.06	-0.04
σ_F	0.05	0.27	-0.56	-1.77	0.06	-1.14	-0.35	0	-0.58	0.04
A_I	0.05	0.26	0.24	0.51	0.01	0.68	0.1	0	0.01	0.06
P(alpha)	142	98	67	101	70	151	111	57	100	108
P(beta)	83	93	89	54	119	37	110	75	87	160
P(turn)	66	95	156	146	119	74	98	156	95	47
f(i)	0.06	0.07	0.161	0.147	0.149	0.056	0.074	0.102	0.14	0.043
f(i+1)	0.076	0.106	0.083	0.11	0.05	0.06	0.098	0.085	0.047	0.034
f(i+2)	0.035	0.099	0.191	0.179	0.117	0.077	0.037	0.19	0.093	0.013
f(i+3)	0.058	0.085	0.091	0.081	0.128	0.064	0.098	0.152	0.054	0.056

Table 1 continued

Amino acid	L	K	M	F	P	S	T	W	Y	V
Mass, Dalton	113.16	128.17	131.19	147.18	97.12	87.08	101.11	186.12	163.18	99.14
Surface Area, Å^2	170	200	185	210	145	115	140	255	230	155
Residue Volume, Å^3	166.7	168.6	162.9	189.9	112.7	89	116.1	227.8	193.6	140
Van der Waals volume, Å^3	124	135	124	135	90	73	93	163	141	105
Residue Non-polar Surface Area, Å^2	38+116	43+86	90	56	66	42	69	45	122	89
Residue Burial, kcal/mol	2.81	2.45	2.25	1.4	1.65	1.05	1.73	1.13	3.05	2.23
Side Chain Burial, kcal/mol	1.6	1.3	1.1	0.2	0.5	-0.1	0.5	0.1	1.9	1.1
Hydropathy index	3.8	-3.9	1.9	2.8	-1.6	-0.8	-0.7	-0.9	-1.3	4.2
Ranking of amino acid polarities	3	20	5	2	13	14	12	6	8	4
pK_a	9.6	8.95	9.21	9.13	10.6	9.15	9.1	9.39	9.11	9.62
σ_I	0.02	-0.16	0.08	0.04	0	-0.03	-0.05	0.06	0.05	0.01
$H_M\Delta PH$	0.07	-1.11	-0.04	0.06	0.1	-0.05	-0.03	0.15	0.02	0.09
σ_R	0.05	-0.95	-0.12	0.02	0.1	-0.02	0.02	0.09	-0.03	0.08
σ_α	-0.04	-0.05	-0.05	-0.08	-0.04	-0.02	-0.03	-0.12	-0.09	-0.03
σ_F	-0.03	0.51	-0.3	-0.45	0.02	-0.38	-0.44	-0.24	-0.42	-0.04
A_I	0.02	0.16	0.08	0.04	0	0.03	0.05	0.06	0.05	0.01
P(alpha)	121	114	145	113	57	77	83	108	69	106
P(beta)	130	74	105	138	55	75	119	137	147	170
P(turn)	59	101	60	60	152	143	96	96	114	50
f(i)	0.061	0.055	0.068	0.059	0.102	0.12	0.086	0.077	0.082	0.062
f(i+1)	0.025	0.115	0.082	0.041	0.301	0.139	0.108	0.013	0.065	0.048
f(i+2)	0.036	0.072	0.014	0.065	0.034	0.125	0.065	0.064	0.114	0.028
f(i+3)	0.07	0.095	0.055	0.065	0.068	0.106	0.079	0.167	0.125	0.053

σ_I, inductive effect scale; $H_M\Delta PH$, normalized Mulliken population data in the context of phenol; σ_R, resonance effect scale; σ_α, normalized polarizability index; σ_F, field effect index; A_I, additional scale; f(i), frequency of the 1st residue in turn; f(i+1), frequency of the 2nd residue in turn; f(i+2), frequency of the 3rd residue in turn; f(i+3), frequency of the 4th residue in turn.

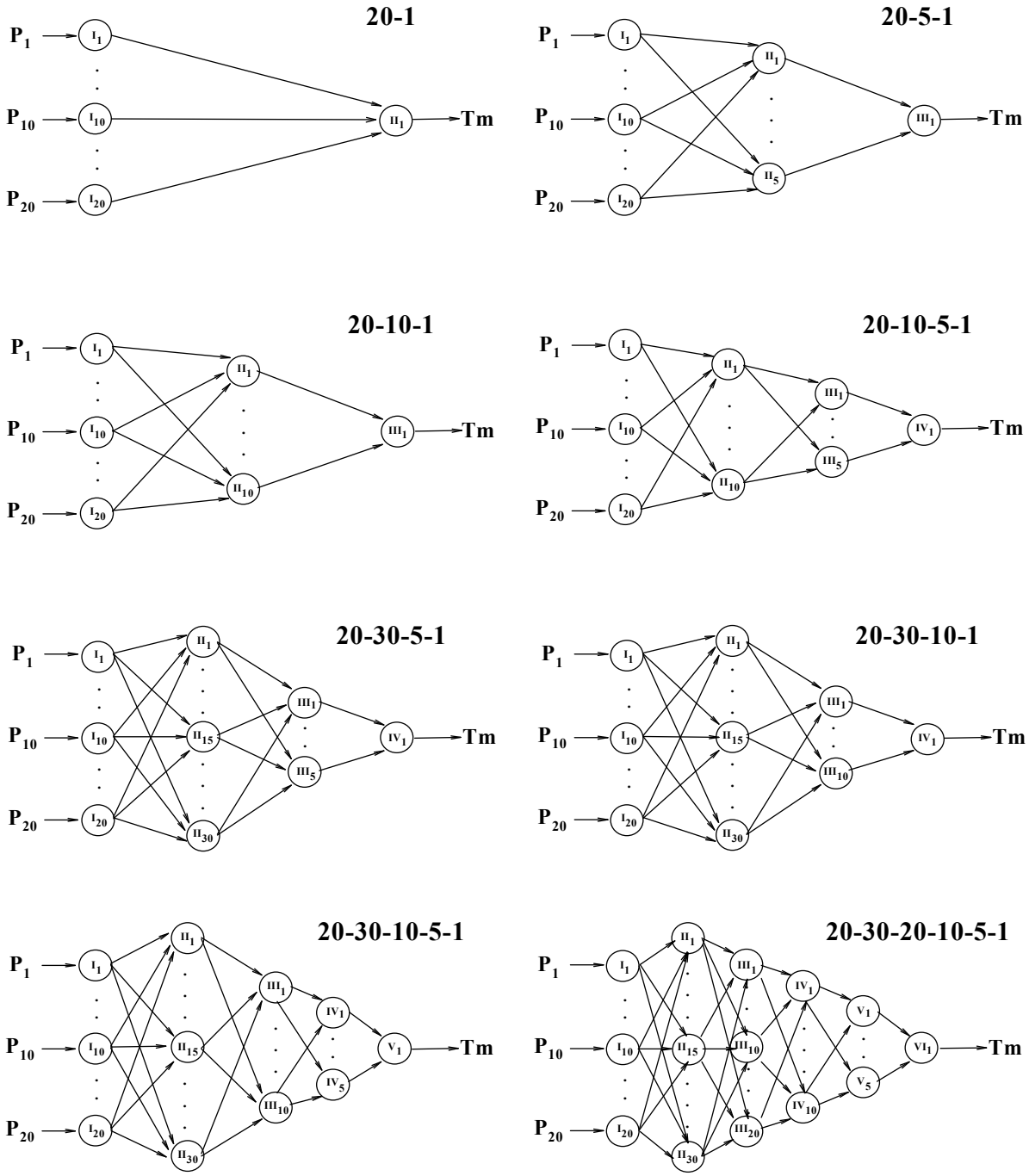

Figure 1. Eight different neural networks used to determine which predictive model is the best.

and energy state, i.e., Maxwell-Boltzmann, Fermi-Dirac and Bose-Einstein assumptions [33]. Thus, the amino acid distribution probability for each type of amino acids is different with respect to its number and position in a protein. This difference can be seen in the 8th and 9th columns of Table 1. Hence, we selected totally 25 amino acid properties as predictors (see Table 2), but each time we use a single property in developing a predictive model.

2.3. Predictive Model

We use a feedforward backpropagation neural network [34, 35] to model the relationship between a single amino acid property and the temperature optimum of beta-cellobiosidas, as we have no prior knowledge about whether the

relationship would be linear or nonlinear. In such a situation, the neural network is suitable because it can accommodate any type of relationships. As a neural network has different layers and neurons, we attempt eight neural networks with different layers and neurons (Figure 1). The transfer functions are tan-sigmoid for the input and hidden layers, and linear for the output layer. The training algorithm is the resilient backpropagation, which is the fastest algorithm on pattern recognition in MatLab [36].

2.4. Exhausted Jackknife Validations

In fact, only 20 beta-cellobiosidases are available in the database. This poor data indeed were weakness for model development, but it gives us the possibility to conduct exhausted jackknife validations. Currently, the delete-1 jackknife validation is widely used for its effectiveness compared with other two methods, independent dataset test and subsampling test [37, 38].

For our case, delete-1 jackknife validation is that each time 19 beta-cellobiosidases were used as training group to generate model parameters, and then the omitted beta-cellobiosidase was used to validate the prediction until all 20 beta-cellobiosidases undergone the same procedure. Up until now, the delete-1 jackknife validation is mostly widely used in model development, but we do not know how a jackknife validation works with different deletions because of a huge amount of computations [1]. To answer this question, exhausted jackknife validations were conducted from delete-1 to delete-18.

2.5. Statistics

To search the best predictor, 100 epochs were run in each training in order to converge the fitting. For each beta-cellobiosidase, 100 trainings were conducted to get the mean±SD of predicted temperature optimum to compare with the recorded temperature optimum [39]. For exhausted jackknife validations similar procedures were adopted, but 10 trainings were conducted to get the mean±SD of predicted values for comparison.

3. Results and Discussions

Neural network can theoretically account for a variety of linear and nonlinear relationships between a property of amino acid and a temperature optimum of enzyme, yet we need to find out the suitable number of layers and neurons. Figure 1 showed eight different neural networks used to determine which predictive model was the best.

Technically, the initialization of weights and biases, and the number of training epochs decide whether a neural network can converge. The random initialization function was used to initialize weights and biases, and 100 epochs were used to converge in a single training. The convergence is important because it screens predictors and we can eliminate the one that cannot converge.

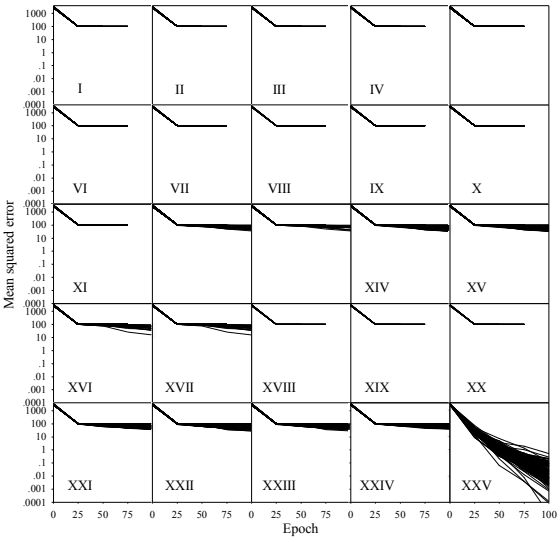

Figure 2. Convergence in terms of mean squared error performance function with respect to each amino acid property.

Figure 2 displays the convergence with respect to each of 25 different amino acid properties in each panel, where each line represents a training process with random initialization of weights and biases going through 100 epochs in each panel. As can be seen, the convergence for amino acid distribution probability, panel XXV, constantly continued through 100 epochs whereas the convergences for other amino acid properties showed a plateau, i.e., their convergence cannot continue. When a model cannot converge, the simplest suggestion is that the reassumed relationship between predictor and predicted value is not suitable. Actually, all 23 properties of individual amino acids cannot converge perfectly, even they were weighed with amino acid composition. The explanation for this unsuitability is that the properties reflecting individual amino acid characteristics do not have the characteristics of a whole protein, while enzyme works as a whole protein rather than a number of individual amino acids. Therefore, only the amino acid distribution probability- the property combined individual amino acid characteristic and whole protein, can serve as a better predictor.

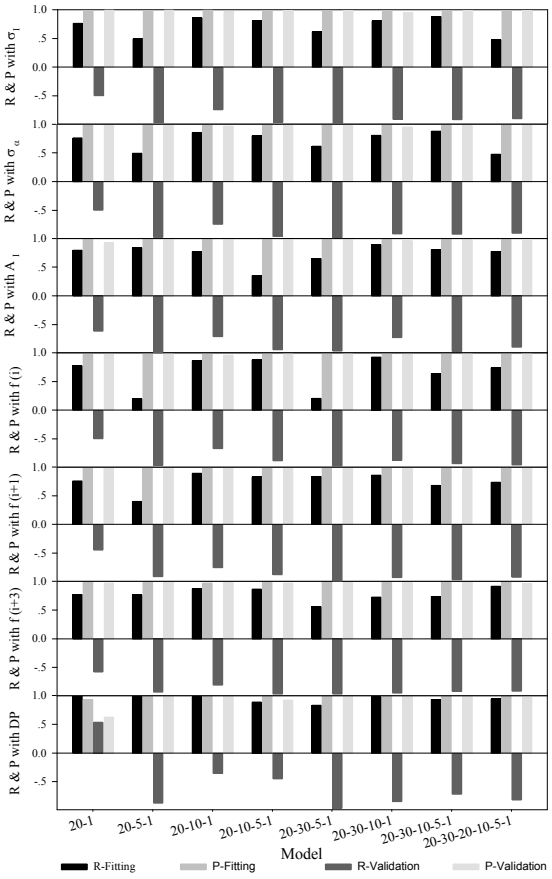

Figure 3. *R* and *P* values obtained from eight different neural networks with seven different predictors in fitting and delete-1 jackknife validation.

Figure 3 illustrates the model selection, i.e., which neural network model works best. For this purpose, the fitted and recorded temperature optimums as well as predicted and recorded ones were evaluated by regression, whose *R* and *P* values serve as indicators for model selection. In this context, the model where *R* is not negative (a bar below zero line) should be chosen because negative *R* indicates a reversed relationship between fitted and recorded as well as predicted and recorded temperature optimums, which may suggest unsuitable model relationship or over-fitting and over-parameterized of model. Hence, 20-1 neural network should be the first choice because it has less cases of negative *R*. In fact, Figure 3 deals with not only model selection but also predictor selection. The seven predictors were selected from Figure 2 because their convergences were better than others. Of these seven predictors (seven panels), the last predictor, amino acid distribution probability (DP), worked best because it combines individual amino acid characteristic and whole protein characteristic. Thus the results in Figure 3 confirmed the results in Figure 2 once again.

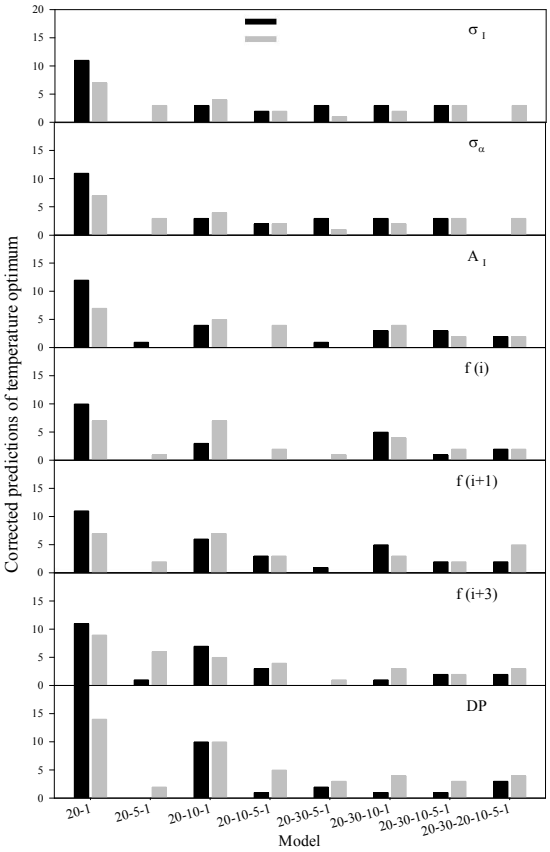

Figure 4. Corrected predictions of temperature optimum using eight different neural networks with seven different predictors in fitting and delete-1 jackknife validation.

As model and predictor selections lie in the very heart of model development, Figure 4 shows the results of model and predictor selections from the aspect of how many corrected predictions of temperature optimum. As can be seen, the 20-1 neural network model worked the best among eight neural network models, and once again amino acid distribution probability--a property combined individual amino acid characteristic and whole protein characteristics, worked the best. Also, Figure 5 showed the statistical comparison with respect to the results shown in Figure 4.

Another issue in developing predictive model is how wide predictions spread. To answer this question, the coefficient of variance of predictions was compared in Figure 6. In this figure, it can be seen that the 20-1 neural network model had the largest coefficients of variance, so its predictions spread in relatively large range around actual

temperature optimum. This in fact is the main consideration that we hope to test other neural network models, because they have a small coefficient of variance. However, taking all these figures together, the priority was given to 20-1 neural network model with amino acid distribution probability as predictor.

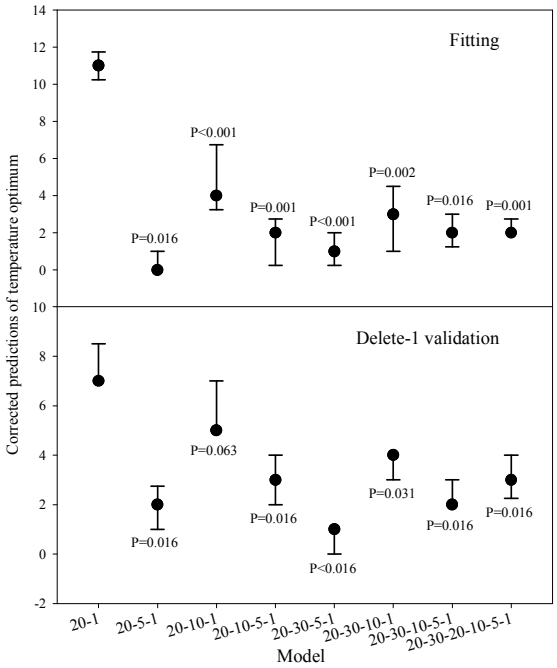

Figure 5. Comparison of corrected predictions of temperature optimum obtained from eight different neural networks with seven different predictors in fitting and delete-1 jackknife validation. The data are presented as median with interquartile. The *P* values are obtained from Mann-Whitney rank sum test compared with Model 20-1.

To this point, the development of model to predict temperature optimum reached its end because we have determined which predictor works best and which neural network works best. Another issue that needs to address is the model validation, which generally divides the dataset into two groups, although it is possible that we do not divide the dataset as the approach used in clinical pharmacology, where different compartment models are used to fit the blood-concentration time curve, in order to see which model works the best [40].

When dividing dataset into two groups, a group of data was used to determine model parameters, while the other was used to validate, i.e., the parameters obtained from the first group of data were used in the model with predictors from the second group for prediction and comparison. In this context, the jackknife validation was quite popular [2]. To the best of our knowledge, no systematical studies so far had been done with respect to use the jackknife validation with deletions increased from one to n-2, where n is the number of samples in dataset. The main obstacle was that such exhausted jackknife validations were very much time-consuming as the number of validations increases in a factorial factor [2].

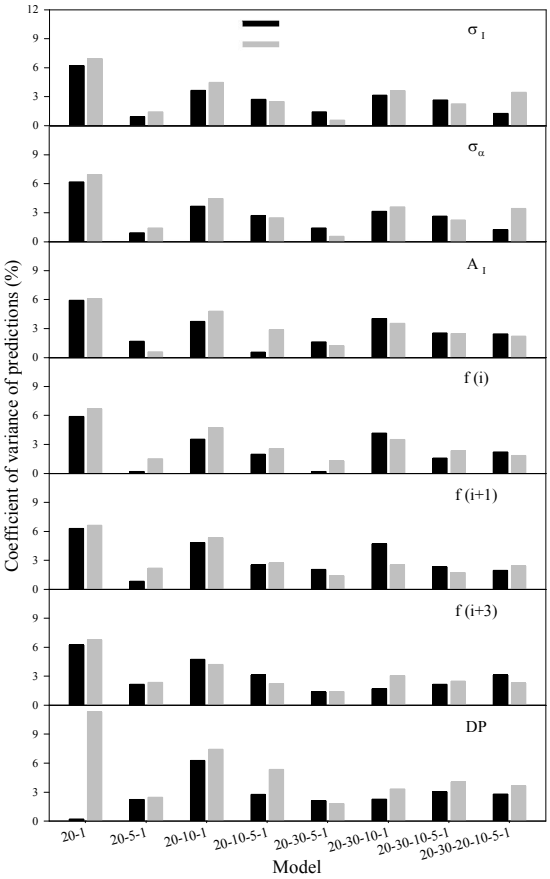

Figure 6. Coefficient of variance of predictions (%) obtained from eight different neural networks with seven different predictors in fitting and delete-1 jackknife validation.

Technically, the exhausted jackknife validation can only be conducted in such a small dataset like beta-cellobiosidases because the number of validations is equal to $C_n^r = \frac{n!}{r!(n-r)!}$, where *n* is the number of dataset and *r* is the number of deletion. For delete-1 jackknife validation, we have n = 20 and r = 1, so $C_{20}^{1} = \frac{20!}{1!(20-1)!} = \frac{20 \times 19!}{19!} = 20$, while for delete-10 jackknife validation, we have n = 20 and r = 10, so

$$C_{20}^{10} = \frac{20!}{10!(20-10)!}$$
$$= \frac{20\times19\times18\times17\times16\times15\times14\times13\times12\times11\times10!}{10!\times10!}$$
$$= \frac{670442572800}{3628800} = 184756$$

Figure 7 showed the number of jackknife validations with respect to different deletions, where we can see that the number was huge. This was the reason why we have not yet get a conclusion on which jackknife validation works best with respect to different deletions.

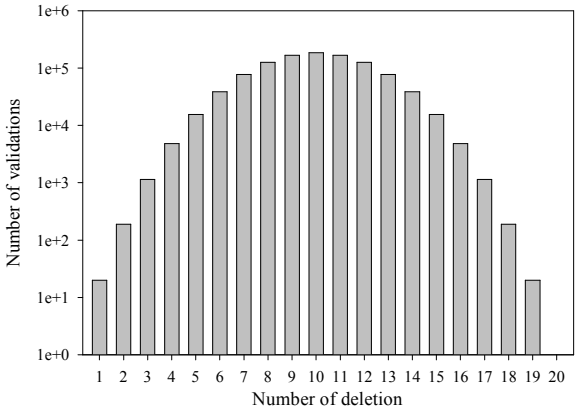

Figure 7. Numbers of jackknife validations with different deletions for dataset of 20 beta-cellobiodases

Figure 8 demonstrated the statistical comparisons on jackknife validations with different deletions. The top panel telled that no difference exists from delete-6 to delete-10 jackknife validations and the predictions based on delete-1 jackknife validation were quite good. The next to top panel compared the coefficient of variance for all jackknife validations, which showed the delete-18 jackknife validation worked really poor. The next to bottom panel indicated the comparison of correlation coefficient, *R* value, between predicted and actual temperature optimum, which favored the delete-1 jackknife validation. Equally the *P* value in bottom panel also favored the delete-1 jackknife validation.

In conclusion, we had investigated two issues: (i) which predictor serves better in predicting temperature optimum of beta-cellobiosidases, and the answer is the amino acid distribution probability that combined individual amino acid characteristic and whole protein characteristic, and (ii) with which deletion jackknife validation works better, and the answer is the delete-1 jackknife validation, when taking all the factors into consideration [1].

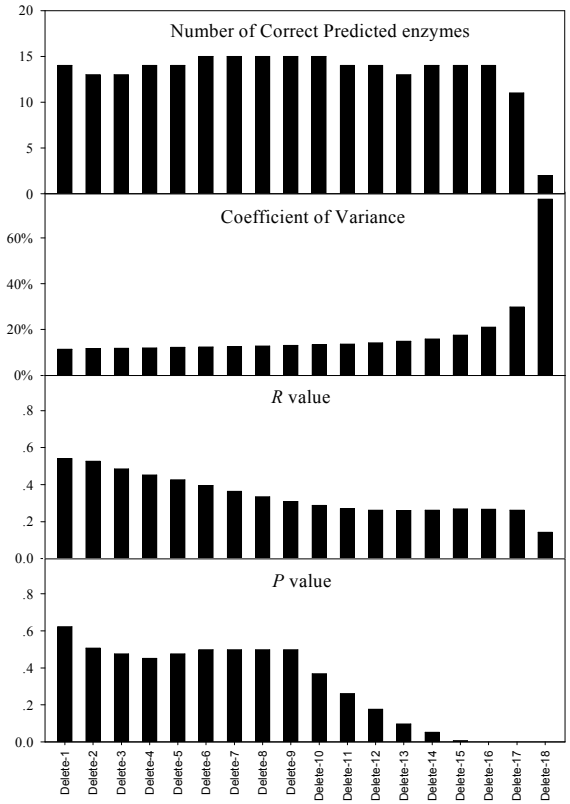

Figure 8. Statistical analyses on exhausted jackknife validations with amino acid distribution probability as predictor

Acknowledgements

This study was partly supported by Guangxi Science Foundation (11107021-5-2, 12237022, 1347004-1, 2013GXNSDA019007, 13-051-08 and 13-051-50,) and by BaGui Scholars Program Foundation.

***Corresponding Author:**
Guang Wu
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.
E-mail: hongguanglishibahao@yahoo.com

References

[1] Yan S, Shi D, Nong H, Wu G. Simultaneously predicting pH and temperature optimum in catalytic reaction of beta-glucosidase. Guangxi Sci 2011; 18: 253-260.
[2] Yan S, Wu G. Searching of predictors to predict pH of cellulases. Appl Biochem Biotech A:

Enzyme Eng Biotech 2011; 165: 856-869.
[3] Yan S. Wu G. Exhausted jackknife validation exemplified by prediction of temperature optimum in enzymatic reaction of cellulases. Appl Biochem Biotech A: Enzyme Eng Biotech 2012; 166: 997-1107.
[4] Yan SM, Wu G. Prediction of Michaelis-Menten constant of beta-glucosidases using nitrophenyl-beta-D-glucopyranoside as substrate. Protein Pept Lett 2011; 18: 1053-1057.
[5] Yan S, Wu G. Prediction of optimal pH and temperature of cellulases using neural network. Protein Pept Lett 2012; 19: 29-39.
[6] Yan S, Shi D, Nong H, Wu G. Predicting Km values of beta-glucosidases using cellobiose as substrate. Interdiscip Sci Comput Life Sci 2012; 4: 46-53.
[7] Yan S, Wu G. Prediction of Michaelis-Menten constant in beta-cellobiosidase's reaction with lactoside as substrate. Enzyme Eng 2011;1:102.
[8] Yan S, Wu G. Prediction of turnover number of cellulose 1,4-beta-cellobiosidase. Protein Pept Lett 2013;20:255-264.
[9] Fox JM, Levine SE, Clark DS, Blanch HW. Initial- and processive-cut products reveal cellobiohydrolase rate limitations and the role of companion enzymes. Biochemistry 2012;51:442-452.
[10] Matano Y, Hasunuma T, Kondo A. Display of cellulases on the cell surface of Saccharomyces cerevisiae for high yield ethanol production from high-solid lignocellulosic biomass. Bioresour Technol 2012;108:128-133.
[11] Zhang J, Zhong Y, Zhao X, Wang T. Development of the cellulolytic fungus Trichoderma reesei strain with enhanced beta-glucosidase and filter paper activity using strong artificial cellobiohydrolase 1 promoter. Bioresour Technol 2010;101:9815-9818.
[12] Liu YS, Baker JO, Zeng Y, Himmel ME, Haas T, Ding SY. Cellobiohydrolase hydrolyzes crystalline cellulose on hydrophobic faces. J Biol Chem 2011;286:11195-11201.
[13] Ma L, Zhang J, Zou G, Wang C, Zhou Z. Improvement of cellulase activity in Trichoderma reesei by heterologous expression of a beta-glucosidase gene from Penicillium decumbens. Enzyme Microb Technol 2011;49:366-371.
[14] Colussi F, Garcia W, Rosseto FR, de Mello BL, de Oliveira Neto M, Polikarpov I. Effect of pH and temperature on the global compactness, structure, and activity of cellobiohydrolase Cel7A from Trichoderma harzianum. Eur Biophys J 2012;41:89-98.
[15] Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J Theor Biol 2011;273:236-247.
[16] Levitin A. Introduction to the Design and Analysis of Algorithms. 1st Edition, Pearson Education, NJ, USA, 2003.
[17] The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 2010;38:D142-D148.
[18] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 2008;36:D202-D205.
[19] Yang XY, Shi XH, Meng X, Li XL, Lin K, Qian ZL, Feng KY, Kong XY, Cai YD. Classification of transcription factors using protein primary structure. Protein Pept Lett 2010;17:899-908.
[20] Burlingame AL, Carr SA. Mass Spectrometry in the Biological Sciences. Humana Press, Totowa, NJ, 1996.
[21] Zamyatin AA. Protein volume in solution. Prog Biophys Mol Biol 1972;24:107-123.
[22] Darby NJ. Creighton TE. 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 1993;232:873-896.
[23] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105-132
[24] Trinquier G, Sanejouand YH, Hausman RE. Which effective property of amino acids is best preserved by the genetic code? Protein Eng 1998;11:153-169.
[25] Cooper GM. The Cell: A Molecular Approach. Washington, D.C: ASM Press, 2004; p 51.
[26] Dwyer DS. Electronic properties of amino acid side chains: quantum mechanics calculation of substituent effects. BMC Chem Biol 2005;5:2.
[27] Chou PY, Fasman GD. Prediction of secondary structure of proteins from amino acid sequence. Adv. Enzymol Relat Subj Biochem 1978;47:45-148.
[28] Wu G, Yan S. Randomness in the primary structure of protein: methods and implications. Mol Biol Today 2002;3:55-69.
[29] Wu G, Yan S. Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint. Acta Pharmacol Sin 2006; 27: 513-526.
[30] Wu G, Yan S. Fate of influenza A virus proteins. Protein Pept Lett 2006; 13:377-384.
[31] Yan S, Wu G. Creation and application of computational mutation. J Guangxi Acad Sci 2010; 17: 145-150.
[32] Wu G, Yan S. Lecture Notes on Computational

Mutation. Nova Science Publishers, New York, 2008.
[33] Feller W. An Introduction to Probability Theory and Its Applications. 3rd ed. Vol, I. Wiley, New York, 1968.
[34] Hagan MT, Demuth HB, Beale MH. Neural Network Design. PWS Publishing Company, Boston, MA, 1996.
[35] Demuth H, Beale M. Neural Network Toolbox for Use with MatLab. User's Guide. Version 4, 2001.
[36] MathWorks Inc. MatLab – The Language of Technical Computing. (version 6.1.0.450, release 12.1), 1984-2001.
[37] Chou KC, Shen HB. Review: Recent progresses in protein subcellular location prediction. Anal Biochem 2007;370:1-16.
[38] Chou KC, Shen HB. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natural Sci 2010;2:1090-1103.
[39] Sokal RR, Rohlf FJ. Biometry: The Principles and Practices of Statistics in Biological Research. 3rd ed WH Freeman, New York, 1995; p 203-218.
[40] Wu G, Cossettini P, Furlanut M. Prediction of blood cyclosporine concentrations in haematological patients with multidrug resistance by one-, two- and three-compartment models using Bayesian and non-linear least squares methods. Pharmacol Res 1996;34:47–57.

9/13/2013