http://www.scirp.org/journal/jbise

J. Biomedical Science and Engineering,
2019, Vol. 12, (No. 7), pp: 354-367

Predicting pH Optimum for Activity of Beta-Glucosidases

Shaomin Yan, Guang Wu

State Key Laboratory of Non-Food Biomass and Enzyme Technology, Guangxi Key Laboratory of Bio-Refinery, National Engineering Research Center for Non-Food Biorefinery, Guangxi Biomass Engineering Technology Research Center, Guangxi Academy of Sciences, Nanning, China

Correspondence to: Guang Wu, hongguanglishibahao@gxas.cn

Keywords: Beta-Glucosidase, Enzyme, pH Optimum, Prediction

Received: June 10, 2019 **Accepted:** July 20, 2019 **Published:** July 23, 2019

ABSTRACT

The working conditions for enzymatic reaction are elegant, but not many optimal conditions are documented in literatures. For newly mutated and newly found enzymes, the optimal working conditions can only be extrapolated from our previous experience. Therefore a question raised here is whether we can use the knowledge on enzyme structure to predict the optimal working conditions. Although working conditions for enzymes can be easily measured in experiments, the predictions of working conditions for enzymes are still important because they can minimize the experimental cost and time. In this study, we develop a 20-1 feedforward backpropagation neural network with information on amino acid sequence to predict the pH optimum for the activity of beta-glucosidase, because this enzyme has drawn much attention for its role in bio-fuel industries. Among 25 features of amino acids being screened, the results show that 11 features can be used as predictors in this model and the amino-acid distribution probability is the best in predicting the pH optimum for the activity of beta-glucosidases. Our study paves the way for predicting the optimal working conditions of enzymes based on the amino-acid features.

1. INTRODUCTION

Enzymatic reactions require many elegant conditions, which are usually determined through experiments. Those elegant experimental conditions are valuable for any new experiments with new enzymes because they can save much time and money for experimenters. On the other hand, many elegant experimental conditions are not always available in literature, so the valuable experience could not be fully useful for fellow researchers.

Still, the modern protein designing produces numerous new enzymes, whose optimal working conditions are totally unknown. Although we can extrapolate our previous experience to new enzymes, they are generally empirical.

With fast development on computational chemistry and bioinformatics, it could be possible to use models to predict the optimal working conditions for enzymatic reactions with newly designed enzymes. This is plausible because currently a lot of information on primary, secondary, tertiary, and quaternary structures is readily available, and many studies have been done on account of structure-function relationship of proteins [1, 2]. Actually the optimal working conditions are adjusted in order to be suitable for enzymatic function, more exactly for enzyme structure, therefore we could assume that there is a certain relationship between enzyme structure and working conditions in enzymatic reaction. This assumption would lay the foundation for predicting the working condition in enzymatic reaction using enzyme structure.

Another effort made by scientific community is to build a comprehensive database to include enzymes with their functional parameters in enzymatic reactions, for example, Km and pH. However, even such comprehensive database cannot include all the parameters for all enzymes simply because many enzymatic parameters are not documented in literature.

Although a measurement of working condition is not difficult during experiments, a measurement is different from a prediction, not only because they are different along the time course, *i.e.* a measurement is related to the past while a prediction is related to the future; but also because they are different in mechanism, *i.e.* a measurement is related to mechanism of enzymatic reaction while a prediction is related to enzyme structure-function relationship.

The β-glucosidase (EC 3.2.1.21) plays an important role in biological processes because it cuts the β-bond linkage in glucose molecules [3], of which celluloses got much recent attention because of interests in its role in biofuels [4]. With such great interest, more efforts are made not only to search for new β-glucosidases but also to mutate current β-glucosidases, so we have more and more β-glucosidases with clear annotations of their primary structures but without their working conditions for enzymatic reactions, for example, pH optimum. This would provide a good case for developing models to predict the pH optimum for the activity of newly mutated and newly found β-glucosidases because the predictions of pH for protein stability have been the research focus for years [5-8]. Therefore, the prediction of pH optimum for enzyme reaction would advance our current knowledge from structure-function relationship to structure-environment relationship.

In this study, we attempted to use the knowledge about amino-acid features from β-glucosidase sequences to predict the pH optimum for the activity of β-glucosidases.

2. MATERIALS AND METHODS

2.1. Data

The β-glucosidases (EC 3.2.1.21) are found in the Comprehensive Enzyme Information System BRENDA [9]. In this databank, only 34 β-glucosidases were found under the category of pH optimum as functional parameter, of which two β-glucosidases are documented with their mutants [10, 11]. Also, two pH values are documented in each of the β-glucosidase A9UIG0, B5TWK3, and P96316, respectively. In total, this databank provides 44 matched β-glucosidases with their pH values, while information on sequences of β-glucosidases was found in the UniProt [12].

2.2. Predictors

For most enzymes, we generally have only their primary structure because the knowledge on secondary, tertiary, and quaternary structures would require considerable amount of experiments. Therefore, the prediction at this stage would focus on using knowledge of amino acids from enzyme sequences. We use several amino-acid properties listed in **Table 1** as predictors. The knowledge in **Table 1** is actually the values reflecting various aspects of amino acids [13], for example, the spatial properties listed from row 3 to row 6 [14, 15]; the hydrophobic properties listed from row 7 to row 11 [16-18]; the electronic properties

Table 1. Features of amino acids used as predictors. 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, phenylalanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine. σ_I: Inductive effect scale; $H_M\Delta PH$: Normalized Mulliken population data for the amino-acid side chains 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.

Amino acid	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
Mass, Dalton	71.09	156.19	114.11	115.09	103.15	129.12	128.14	57.05	137.14	113.16	113.16	128.17	131.19	147.18	97.12	87.08	101.11	186.12	163.18	99.14
Surface Area, Å^2	115	225	160	150	135	190	180	75	195	175	170	200	185	210	145	115	140	255	230	155
Residue Volume, Å^3	88.6	173.4	114.1	111.1	108.5	138.4	143.8	60.1	153.2	166.7	166.7	168.6	162.9	189.9	112.7	89.0	116.1	227.8	193.6	140.0
van der Waals volume, Å^3	67	148	96	91	86	114	109	48	118	124	124	135	124	135	90	73	93	163	141	105
Residue Non-polar Surface Area, Å^2	86	89	42	45	48	69	66	47	129	155	122	164	137	194	124	56	90	236	154	135
Residue Burial, kcal/mol	2.15	2.23	1.05	1.13	1.20	1.73	1.65	1.18	2.45	3.88	3.05	4.10	3.43	3.46	3.10	1.40	2.25	4.11	2.81	3.38
Side Chain Burial, kcal/mol	1.0	1.1	−0.1	−0.1	0.0	0.5	0.5	0.0	1.3	2.7	1.9	2.9	2.3	2.3	1.9	0.2	1.1	2.9	1.6	2.2
Hydropathy index	4.5	4.2	−0.8	−0.9	−3.5	−0.7	−1.6	1.8	−3.9	−3.5	−1.3	2.5	−0.4	−3.2	−3.5	2.8	1.9	4.5	3.8	−3.5
Ranking of amino acid polarities	9	15	16	19	7	18	17	11	10	1	3	20	5	2	13	14	12	6	8	4
pK_a	9.69	9.04	8.80	9.60	10.28	9.67	9.13	9.60	9.17	9.68	9.60	8.95	9.21	9.13	10.60	9.15	9.10	9.39	9.11	9.62
σ_I	0.05	−0.26	−0.14	0.51	−0.01	0.68	−0.10	0.00	−0.01	0.06	0.02	−0.16	0.08	0.04	0.00	−0.03	−0.05	0.06	0.05	0.01
$H_M\Delta PH$	0.05	−0.75	−0.20	1.80	−0.01	1.25	−0.07	0.00	0.21	0.08	0.07	−1.11	−0.04	0.06	0.10	−0.05	−0.03	0.15	0.02	0.09

Continued

σ_R	0.00	−0.49	−0.06	1.29	0.01	0.57	0.03	0.00	0.22	0.02	0.05	−0.95	−0.12	0.02	0.10	−0.02	0.02	0.09	−0.03	0.08
σ_α	−0.01	−0.08	−0.04	−0.03	−0.03	−0.04	−0.05	0.00	−0.06	−0.04	−0.04	−0.05	−0.05	−0.08	−0.04	−0.02	−0.03	−0.12	−0.09	−0.03
σ_F	0.05	0.27	−0.56	−1.77	0.06	−1.14	−0.35	0.00	−0.58	0.04	−0.03	0.51	−0.30	−0.45	0.02	−0.38	−0.44	−0.24	−0.42	−0.04
A_I	0.05	0.26	0.24	0.51	0.01	0.68	0.10	0.00	0.01	0.06	0.02	0.16	0.08	0.04	0.00	0.03	0.05	0.06	0.05	0.01
P(α-helix)	142	98	101	67	70	151	111	57	100	108	121	114	145	113	57	77	83	108	69	106
P(β-sheet)	83	93	54	89	119	37	110	75	87	160	130	74	105	138	55	75	119	137	147	170
P(turn)	66	95	146	156	119	74	98	156	95	47	59	101	60	60	152	143	96	96	114	50
f(i)	0.060	0.070	0.147	0.161	0.149	0.056	0.074	0.102	0.140	0.043	0.061	0.055	0.068	0.059	0.102	0.120	0.086	0.077	0.082	0.062
f(i+1)	0.076	0.106	0.110	0.083	0.050	0.060	0.098	0.085	0.047	0.034	0.025	0.115	0.082	0.041	0.301	0.139	0.108	0.013	0.065	0.048
f(i+2)	0.035	0.099	0.179	0.191	0.117	0.077	0.037	0.190	0.093	0.013	0.036	0.072	0.014	0.065	0.034	0.125	0.065	0.064	0.114	0.028
f(i+3)	0.058	0.085	0.081	0.091	0.128	0.064	0.098	0.152	0.054	0.056	0.070	0.095	0.055	0.065	0.068	0.106	0.079	0.167	0.125	0.053

listed from row 12 to row 18 [19], and the amino-acid based secondary structure predictions listed from row 17 to 25 row, which are depended on assigning a set of predicted values to a residue and then calculated by applying a simple algorithm [20].

A particular characteristic in **Table 1** is that those values are constant regardless amino-acid position in a protein, neighboring amino acids, protein length, etc. This is understandable because those properties would not be changed in these regards, for example, an amino acid's physicochemical property would not be different no matter where this amino acid is located in a protein. As the amino-acid composition is different one from another in β-glucosidases, we weigh the values listed in **Table 1** by multiplying their amino-acid composition of each β-glucosidase.

Besides the classical knowledge listed in **Table 1**, there is also the amino-acid distribution probability that is based on the occupancy of subpopulations and partitions [21] and reflects the random aspect of amino-acid distribution along a protein (for review and textbook, see [22-26]). The difference is that the amino-acid distribution probability does not give each amino acid a constant value as shown in **Table 1**, but the value subject to the length of enzyme and position of each amino acid. **Table 2** shows such a difference.

2.3. Predictive Model

As the predictors are directly related to 20 types of amino acids, so it is natural to consider a predictive model to couple 20 inputs of knowledge on amino acids with single output with documented pH optimum. As this predictive model advances a step from structure-function to structure-environment relationship, we choose a 20-1 feedforward backpropagation neural network [27, 28] to account this hidden and implicit relationship after large workings on model selection in **Figure 1**.

2.4. Validation of Predictions

The second column in **Table 3** lists all 44 β-glucosidases obtained from the databank, of which 30 were used to generate the model parameters, weights and biases in neural network as the training group, and 14 were used to validate the neural network with generated weights and biases as the validation group. This is a very traditional approach for validation in neural network.

The second approach for validation is the delete-1 observation jackknife, each time we use 43 β-glucosidases as the training group to generate model parameters, and then to validate the prediction in omitted β-glucosidase until all 44 β-glucosidases undergo the same procedure. It is said to be the most effective approach for validation [29] although it is labor-intensive and time-demanding.

The third approach for validation is the cross-validation, through which 44 β-glucosidases were split into 11 subsets containing 4 cases each or 4 subsets containing 11 cases each. Each time, ten or three subsets were used to generate the model parameters, and one subset was used for validation, such a procedure was conducted in turn until each subset has served for validation [29].

2.5. Statistics

For each predictor, we generated 100 sets of model parameters in order that the predictions based on 100 sets of model parameters to have well normally distributed mean±SD to compare with the documented pH optimum of activity for each β-glucosidase [30]. For the data with normal distribution, the Student's t-test was used, and for the data with abnormal distribution, the non-parametric Mann-Whitney U-test was used. $P < 0.05$ is considered statistically significant. For visual comparison, linear regression was also used to evaluate the predicted pH values with their documented ones.

3. RESULTS AND DISCUSSION

It is highly likely that the relationship between feature of amino acids and pH optimum of activity is at least one step beyond the structure-function relationship, which might imply that we need at least two layers in the neural network to account this hidden and implicit relationship (Figure 1). Technically, the development of predictive method includes (1) selection of predictors, and (2) selection of predictive models, while the general and efficient practice is to select predictors at first, and then to select predictive models.

Working with this network model, the next consideration is the training process, which once again guarantees a fair selection of predictors. Technically, both initialization of weights and biases and number of training epochs govern whether the neural network can converge. We used the random initialization

Table 2. Inductive effect scale, amino-acid number and distribution probability in *β*-glucosidase A9UIG0 and Q4U4W7. The amino-acid distribution probability, is computed according to the equation, $r!/(q_0! \times q_1! \times \ldots \times q_n!) \times r!/(r_1! \times r_2! \times \ldots \times r_n!) \times n^{-r}$, 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 and *n* is the number of partitions in the protein for a type of amino acid. The computation can be found in the web site (http://www.gxas.cn/dp.htm).

Amino Acid	Inductive effect scale		Amino-acid number		Distribution probability	
	A9UIG0	Q4U4W7	A9UIG0	Q4U4W7	A9UIG0	Q4U4W7
A	0.05	0.05	82	74	0.0021	0.0015
R	−0.26	−0.26	29	36	0.0043	0.0012
N	−0.14	−0.14	56	46	0.0202	0.0174
D	0.51	0.51	50	50	0.0039	0.0180
C	−0.01	−0.01	8	8	0.1682	0.0841
E	0.68	0.68	35	36	0.0218	0.0224
Q	−0.10	−0.10	28	25	0.0642	0.0051
G	0.00	0.00	94	86	0.0006	0.0027
H	−0.01	−0.01	16	11	0.0715	0.0808
I	0.06	0.06	35	43	0.0194	0.0240
L	0.02	0.02	58	65	0.0002	0.0054
K	−0.16	−0.16	29	29	0.0317	0.0317
M	0.08	0.08	11	19	0.0404	0.0138
F	0.04	0.04	34	33	0.0285	0.0193
P	0.00	0.00	53	65	0.0058	0.0010
S	−0.03	−0.03	62	61	0.0029	0.0112
T	−0.05	−0.05	57	56	0.0023	0.0005
W	0.06	0.06	18	16	0.0023	0.1362
Y	0.05	0.05	41	33	0.0142	0.0174
V	0.01	0.01	70	74	0.0067	0.0008

Table 3. Comparison between documented and predicted pH optimum of activity in 44 β-glucosidases. The predicted pH optimum was presented as mean ± SD of 100 predictions. No., the amino-acid composition. *, no statistical difference between documented and predicted pH.

Group	Accession Number	Documented pH	pH optimum predicted by predictor										
			σ_I × No.	$H_M\Delta PH$ × No.	σ_R x No.	σ_α × No.	σ_F × No.	A_I × No.	f(i) × No.	f(i + 1) × No.	f(i + 2) × No.	f(i + 3) × No.	Distribution probability
Training	Q9AT27	4.0	4.02 ± 0.05*	4.03 ± 0.06*	4.02 ± 0.06*	4.01 ± 0.05*	4.08 ± 0.16*	4.01 ± 0.05*	4.02 ± 0.07*	4.00 ± 0.05*	4.00 ± 0.06*	4.01 ± 0.06*	3.99 ± 0.01*
	Q8TGI8	4.0	4.03 ± 0.05*	4.02 ± 0.06*	4.04 ± 0.06*	4.05 ± 0.07*	4.08 ± 0.15*	4.03 ± 0.05*	4.05 ± 0.09*	4.04 ± 0.07*	4.07 ± 0.10*	4.05 ± 0.09*	4.00 ± 0.01*
	Q12715	4.6	4.58 ± 0.04*	4.60 ± 0.04*	4.60 ± 0.05*	4.59 ± 0.04*	4.56 ± 0.14*	4.59 ± 0.05*	4.58 ± 0.05*	4.59 ± 0.04*	4.58 ± 0.05*	4.59 ± 0.05*	4.61 ± 0.01*
	A1C3J9	5.0	5.00 ± 0.04*	5.00 ± 0.05*	4.99 ± 0.04*	5.00 ± 0.05*	5.06 ± 0.20*	4.99 ± 0.04*	4.99 ± 0.05*	5.00 ± 0.06*	5.00 ± 0.05*	4.99 ± 0.06*	5.00 ± 0.00*
	Q8T0W7	5.0	5.29 ± 0.14	5.22 ± 0.13*	5.24 ± 0.14*	5.28 ± 0.13	5.35 ± 0.24*	5.31 ± 0.14	5.28 ± 0.13	5.24 ± 0.13*	5.27 ± 0.13	5.25 ± 0.13*	5.00 ± 0.01*
	Q4U4W7	5.0	4.98 ± 0.05*	5.00 ± 0.06*	4.99 ± 0.06*	5.01 ± 0.05*	4.98 ± 0.10*	4.99 ± 0.05*	5.02 ± 0.05*	5.02 ± 0.06*	5.02 ± 0.06*	5.01 ± 0.06*	5.00 ± 0.01*
	A9UIG0	5.0	4.99 ± 0.06*	4.98 ± 0.06*	4.96 ± 0.06*	4.95 ± 0.07*	4.93 ± 0.16*	4.97 ± 0.07*	4.95 ± 0.08*	4.97 ± 0.06*	4.94 ± 0.09*	4.96 ± 0.07*	5.01 ± 0.01*
	P94248	5.5	5.54 ± 0.05*	5.55 ± 0.08*	5.54 ± 0.06*	5.54 ± 0.06*	5.62 ± 0.19*	5.54 ± 0.06*	5.54 ± 0.07*	5.52 ± 0.05*	5.56 ± 0.06*	5.54 ± 0.06*	5.50 ± 0.01*
	O08331	5.5	5.48 ± 0.05*	5.54 ± 0.15*	5.48 ± 0.09*	5.48 ± 0.04*	5.56 ± 0.21*	5.48 ± 0.07*	5.48 ± 0.04*	5.48 ± 0.04*	5.47 ± 0.05*	5.48 ± 0.06*	5.50 ± 0.01*
	Q25BW5 D229N	5.5	6.19 ± 0.04	6.22 ± 0.04	6.20 ± 0.04	6.20 ± 0.03	6.21 ± 0.05	6.20 ± 0.03	6.20 ± 0.04	6.19 ± 0.03	6.20 ± 0.03	6.20 ± 0.04	5.70 ± 0.05
	Q9SLA0	5.6	5.76 ± 0.14*	5.95 ± 0.24*	5.92 ± 0.20*	5.75 ± 0.11*	5.97 ± 0.26*	5.79 ± 0.16*	5.79 ± 0.14*	5.76 ± 0.12*	5.83 ± 0.15*	5.76 ± 0.12*	5.60 ± 0.00*
	P49235	5.8	5.75 ± 0.08*	5.78 ± 0.08*	5.78 ± 0.07*	5.78 ± 0.06*	5.76 ± 0.12*	5.74 ± 0.08*	5.76 ± 0.08*	5.78 ± 0.07*	5.77 ± 0.08*	5.77 ± 0.06*	5.80 ± 0.01*
	Q86D78	6.0	5.89 ± 0.07*	5.90 ± 0.11*	5.92 ± 0.08*	5.90 ± 0.07*	5.91 ± 0.13*	5.90 ± 0.08*	5.90 ± 0.08*	5.92 ± 0.09*	5.89 ± 0.09*	5.90 ± 0.08*	6.00 ± 0.01*
	Q2WGB4	6.0	6.07 ± 0.07*	6.09 ± 0.08*	6.08 ± 0.08*	6.11 ± 0.09*	6.10 ± 0.12*	6.10 ± 0.08*	6.09 ± 0.10*	6.07 ± 0.07*	6.07 ± 0.08*	6.09 ± 0.08*	6.00 ± 0.01*
	Q875K3	6.0	5.99 ± 0.06*	5.96 ± 0.09*	5.97 ± 0.06*	5.99 ± 0.03*	5.97 ± 0.17*	5.99 ± 0.05*	6.00 ± 0.04*	5.99 ± 0.04*	5.99 ± 0.04*	5.99 ± 0.04*	6.00 ± 0.01*
	Q25BW5 V173C	6.0	6.26 ± 0.06	6.26 ± 0.07	6.26 ± 0.07	6.28 ± 0.07	6.26 ± 0.07	6.25 ± 0.07	6.26 ± 0.09	6.25 ± 0.07	6.26 ± 0.08	6.27 ± 0.08	6.00 ± 0.01*
	Q25BW5 D229N/ K253A	6.0	6.20 ± 0.05	6.23 ± 0.05	6.22 ± 0.04	6.21 ± 0.04	6.22 ± 0.05	6.21 ± 0.04	6.22 ± 0.04	6.21 ± 0.04	6.22 ± 0.03	6.21 ± 0.04	5.85 ± 0.03
	Q25BW5	6.5	6.20 ± 0.03	6.21 ± 0.04	6.20 ± 0.03	6.21 ± 0.03	6.22 ± 0.04	6.20 ± 0.03	6.20 ± 0.03	6.20 ± 0.02	6.20 ± 0.03	6.20 ± 0.03	6.42 ± 0.03
	P15885	6.5	6.50 ± 0.05*	6.46 ± 0.14*	6.47 ± 0.10*	6.50 ± 0.03*	6.44 ± 0.17*	6.51 ± 0.04*	6.49 ± 0.04*	6.50 ± 0.04*	6.51 ± 0.04*	6.49 ± 0.07*	6.50 ± 0.01*
	Q59976	6.5	6.50 ± 0.06*	6.50 ± 0.06*	6.50 ± 0.06*	6.48 ± 0.06*	6.48 ± 0.11*	6.50 ± 0.06*	6.48 ± 0.06*	6.49 ± 0.06*	6.48 ± 0.07*	6.48 ± 0.06*	6.50 ± 0.01*

Continued

	Q9H227	6.5	6.50 ± 0.05*	6.53 ± 0.08*	6.52 ± 0.06*	6.47 ± 0.05*	6.51 ± 0.09*	6.50 ± 0.04*	6.51 ± 0.05*	6.51 ± 0.05*	6.50 ± 0.05*	6.49 ± 0.05*	6.48 ± 0.02*
	Q746L1	6.5	6.54 ± 0.06*	6.54 ± 0.08*	6.54 ± 0.09*	6.53 ± 0.04*	6.60 ± 0.19*	6.54 ± 0.08*	6.53 ± 0.06*	6.54 ± 0.05*	6.54 ± 0.08*	6.54 ± 0.08*	6.50 ± 0.01*
	B9K7M5	6.5	6.23 ± 0.13	6.19 ± 0.17*	6.20 ± 0.16*	6.19 ± 0.14	6.18 ± 0.18*	6.20 ± 0.16*	6.23 ± 0.14*	6.27 ± 0.14*	6.24 ± 0.15*	6.26 ± 0.15*	6.50 ± 0.01*
	Q25BW5 K253A	6.5	6.21 ± 0.03	6.22 ± 0.04	6.22 ± 0.04	6.22 ± 0.03	6.23 ± 0.04	6.22 ± 0.03	6.22 ± 0.03	6.22 ± 0.03	6.22 ± 0.03	6.22 ± 0.03	6.53 ± 0.04*
	Q25BW5 M177L	6.5	6.20 ± 0.05	6.21 ± 0.05	6.21 ± 0.06	6.20 ± 0.05	6.23 ± 0.06	6.20 ± 0.05	6.19 ± 0.05	6.20 ± 0.04	6.20 ± 0.05	6.20 ± 0.04	6.49 ± 0.02*
	Q9H227 V168Y	6.5	6.54 ± 0.04*	6.56 ± 0.07*	6.55 ± 0.06*	6.53 ± 0.04*	6.53 ± 0.09*	6.54 ± 0.04*	6.55 ± 0.04*	6.55 ± 0.04*	6.55 ± 0.04*	6.54 ± 0.04*	6.52 ± 0.02*
	Q25BW5 H231D	6.5	6.34 ± 0.08	6.31 ± 0.08	6.33 ± 0.08	6.35 ± 0.07	6.32 ± 0.08	6.32 ± 0.08	6.32 ± 0.08	6.34 ± 0.07	6.32 ± 0.07	6.33 ± 0.07	6.50 ± 0.04*
	Q08IT7	7.0	6.93 ± 0.11*	6.82 ± 0.17*	6.84 ± 0.16*	6.93 ± 0.09*	6.77 ± 0.23*	6.91 ± 0.12*	6.90 ± 0.09*	6.90 ± 0.09*	6.87 ± 0.12*	6.91 ± 0.10*	7.00 ± 0.01*
	Q6QGY5	7.0	6.88 ± 0.09*	6.76 ± 0.17*	6.80 ± 0.14*	6.86 ± 0.11*	6.70 ± 0.23*	6.85 ± 0.11*	6.84 ± 0.11*	6.86 ± 0.09*	6.84 ± 0.11*	6.86 ± 0.08*	7.00 ± 0.01*
	Q47RE2	7.2	7.08 ± 0.10*	7.09 ± 0.14*	7.11 ± 0.11*	7.10 ± 0.10*	6.96 ± 0.26*	7.08 ± 0.12*	7.11 ± 0.13*	7.11 ± 0.08*	7.08 ± 0.12*	7.10 ± 0.11*	7.20 ± 0.01*
Validation	Q08638	3.2	6.59 ± 0.26	6.50 ± 0.27	6.55 ± 0.28	6.69 ± 0.29	6.55 ± 0.32	6.58 ± 0.29	6.74 ± 0.27	6.78 ± 0.26	6.67 ± 0.26	6.71 ± 0.30	5.42 ± 0.99
	B5TWK3	4.5	5.32 ± 0.66*	5.40 ± 0.61*	5.31 ± 0.68*	5.23 ± 0.73*	5.42 ± 0.64*	5.52 ± 0.68*	4.95 ± 0.73*	5.11 ± 0.67*	4.96 ± 0.68*	4.89 ± 0.68*	5.18 ± 0.44*
	B5TWK3	5.0	5.32 ± 0.66*	5.40 ± 0.61*	5.31 ± 0.68*	5.23 ± 0.73*	5.42 ± 0.64*	5.52 ± 0.68*	4.95 ± 0.73*	5.11 ± 0.67*	4.96 ± 0.68*	4.89 ± 0.68*	5.18 ± 0.44*
	B6ZKM3	5.0	5.67 ± 0.67*	5.91 ± 0.60*	6.04 ± 0.59*	5.71 ± 0.68*	5.89 ± 0.56*	5.74 ± 0.66*	5.46 ± 0.61*	5.70 ± 0.59*	5.55 ± 0.64*	5.55 ± 0.67*	5.92 ± 0.97*
	Q2UUD6	5.0	5.04 ± 0.54*	5.30 ± 0.61*	5.19 ± 0.59*	4.66 ± 0.69*	5.27 ± 0.58*	5.13 ± 0.60*	4.75 ± 0.60*	4.86 ± 0.66*	4.82 ± 0.63*	4.75 ± 0.55*	5.24 ± 0.23*
	Q9SPK3	5.0	6.39 ± 0.36	6.35 ± 0.36	6.52 ± 0.39	6.65 ± 0.43	6.18 ± 0.37	6.45 ± 0.46	6.32 ± 0.39	6.35 ± 0.40	6.29 ± 0.40	6.37 ± 0.41	5.90 ± 0.55*
	A9UIG0	6.0	4.99 ± 0.06	4.98 ± 0.06	4.96 ± 0.06	4.95 ± 0.07	4.93 ± 0.16	4.97 ± 0.07	4.95 ± 0.08	4.97 ± 0.06	4.94 ± 0.09	4.96 ± 0.07	5.01 ± 0.01
	O61594	6.0	6.06 ± 0.38*	5.89 ± 0.41*	5.85 ± 0.41*	6.13 ± 0.40*	5.97 ± 0.36*	6.03 ± 0.40*	6.01 ± 0.44*	5.93 ± 0.42*	5.85 ± 0.43*	5.91 ± 0.43*	6.36 ± 0.54*
	Q12601	6.0	5.41 ± 0.66*	5.49 ± 0.72*	5.51 ± 0.72*	5.66 ± 0.70*	5.71 ± 0.66*	5.35 ± 0.82*	5.38 ± 0.83*	5.59 ± 0.76*	5.43 ± 0.73*	5.36 ± 0.82*	5.15 ± 0.56*
	P26208	6.0	5.26 ± 0.62*	5.52 ± 0.58*	5.20 ± 0.61*	4.78 ± 0.76*	5.60 ± 0.58*	5.25 ± 0.63*	5.02 ± 0.64*	5.07 ± 0.69*	5.20 ± 0.65*	5.08 ± 0.69*	7.13 ± 0.90*
	P10482	6.0	6.27 ± 0.37*	6.22 ± 0.38*	6.11 ± 0.39*	6.08 ± 0.38*	6.18 ± 0.44*	6.28 ± 0.39*	5.98 ± 0.40*	6.09 ± 0.41*	5.97 ± 0.45*	6.07 ± 0.43*	5.30 ± 1.09*
	P96316	6.2	5.66 ± 0.80*	5.93 ± 0.82*	5.85 ± 0.81*	5.24 ± 0.91*	5.97 ± 0.82*	5.60 ± 1.03*	5.51 ± 0.91*	5.73 ± 0.89*	5.68 ± 0.87*	5.61 ± 0.81*	6.58 ± 0.69*
	Q9C3Z9	6.4	4.96 ± 0.74*	5.29 ± 0.75*	4.98 ± 0.73*	4.71 ± 0.90*	5.30 ± 0.76*	5.00 ± 0.80*	4.59 ± 0.82	4.77 ± 0.77	4.74 ± 0.91*	4.70 ± 0.82	4.81 ± 0.66
	P96316	6.6	5.66 ± 0.80*	5.93 ± 0.82*	5.85 ± 0.81*	5.24 ± 0.91*	5.97 ± 0.82*	5.60 ± 1.03*	5.51 ± 0.91*	5.73 ± 0.89*	5.68 ± 0.87*	5.61 ± 0.81*	6.58 ± 0.69*
Total			32	34	34	32	34	33	32	33	33	33	38

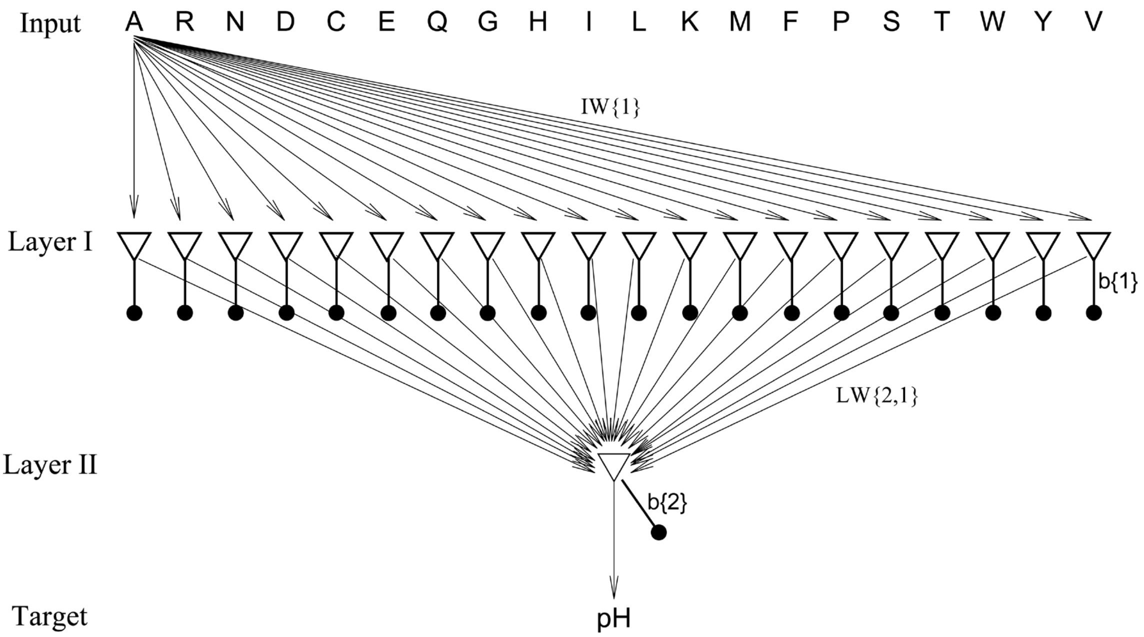

Figure 1. 20-1 feedforward backpropagation neural network to account for the relationship between features of primary structure of *β*-glucosidase, labeled with amino acid abbreviations, and pH optimum. Each triangle presents a neuron. IW{1} is the input weights, LW{2,1} is the layer weights to the second layer from the first layer. b{1} and b{2} are the biases related to each neuron at the first and second layers.

function to initialize weights and biases, and 250 training epochs for convergence. Figure 2 displays the performance of convergence in the training group, where each line represents a training process with random initialization of weights and biases running 250 training epochs. Different predictor has different profiles of its convergence. As seen, the convergence of 11 predictors can be reached within 250 training epochs with any random initialization, which guarantees our training process. However, the convergence of other predictors is not possible after 100 epochs, as the amino-acid composition shown in the top-left panel of Figure 2.

Figure 3 shows that the percentage of correctly predicted pH optimum ranges from about 70% to 90% with respect to different features. Actually, Figure 3 gives us a basic concept on which predictor has better effect on predicting pH optimum of activity. Accordingly, the amino-acid distribution probability is the better one than others. This is because the amino acid features except for amino-acid distribution probability are not subject to their positions in amino acid sequence and their neighboring amino acids, whereas the amino-acid distribution probability is sensitive to these conditions.

Proteins evolve to function in specific cellular environment; thus pH of activity is subject to evolutionary pressure [5]. If the pH level could change the conformation of *β*-glucosidase; then different pH levels could have slight difference in conformation of *β*-glucosidase. In this context, the explanation for the results in Figure 3 would be that the amino-acid distribution probability as it reflects the randomness in enzyme would more accurately reflect the changes in the conformation of *β*-glucosidase due to different pH levels, while the other predictors due to the fact that they have constant values, for example, physicochemical property, would less accurately reflect the changes in conformation of *β*-glucosidase.

Table 3 shows the comparison between documented and predicted pH optimum for each *β*-glucosidase found in the database [9]. We should consider a predictor workable if there is no statistical difference between documented and predicted pH optimum, and the last row of Table 3 shows the overall performance, where we can see that the amino-acid distribution probability gives better predictive results than other predictors, whose results are similar.

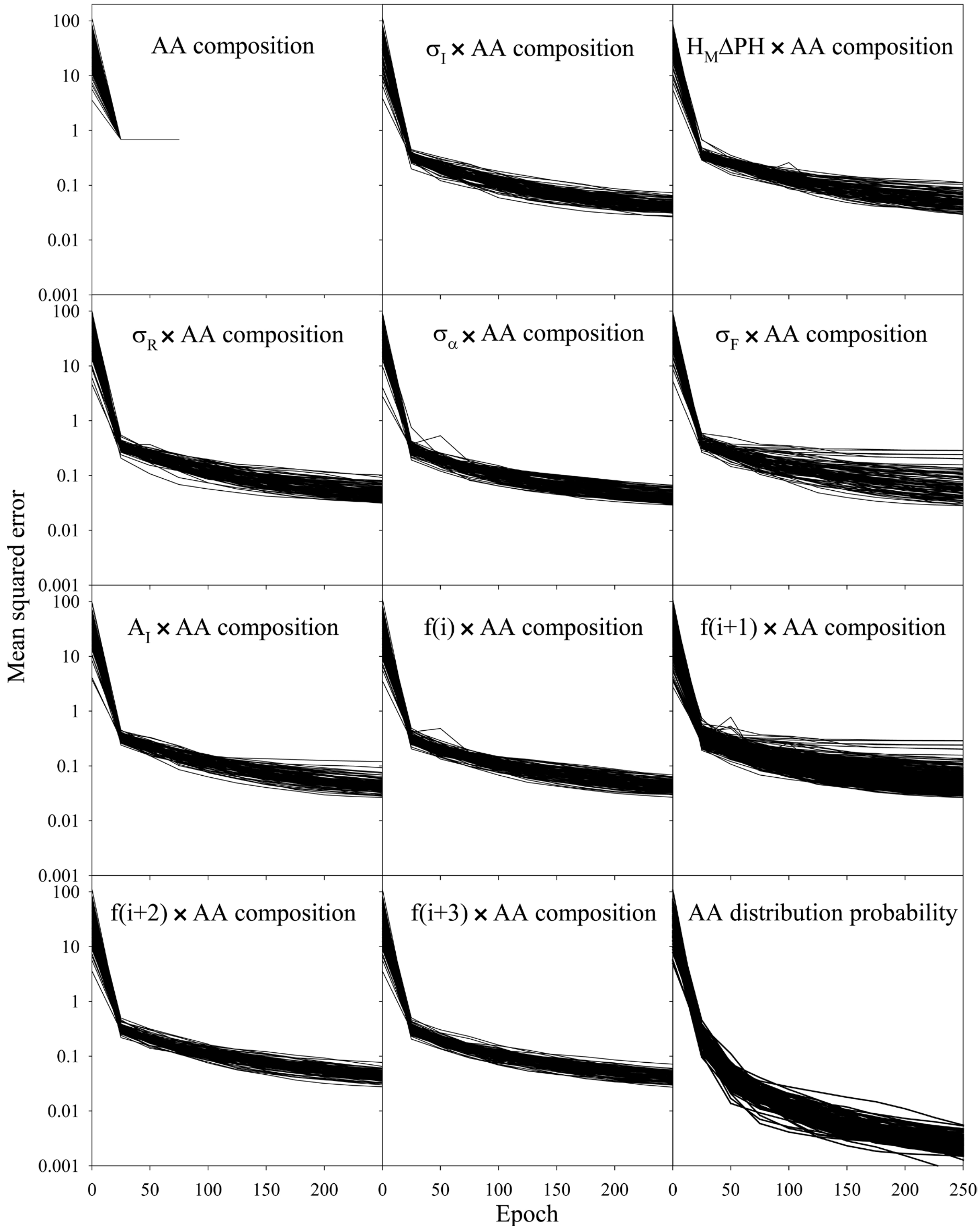

Figure 2. Convergence of mean squared error performance function with 100 different initial weights and biases generated by random initialization function.

In Figure 4, we used the regression between documented and predicted pH optimum of activity to visualize the predictive performance using the amino-acid distribution probability as predictor in order to confirm our observation visually.

To furthermore validate the above findings, we used the delete-1 jackknife validation and 3-fold, 10-fold cross-validation to treat these predictors as shown in Figure 5, where we once again find that the

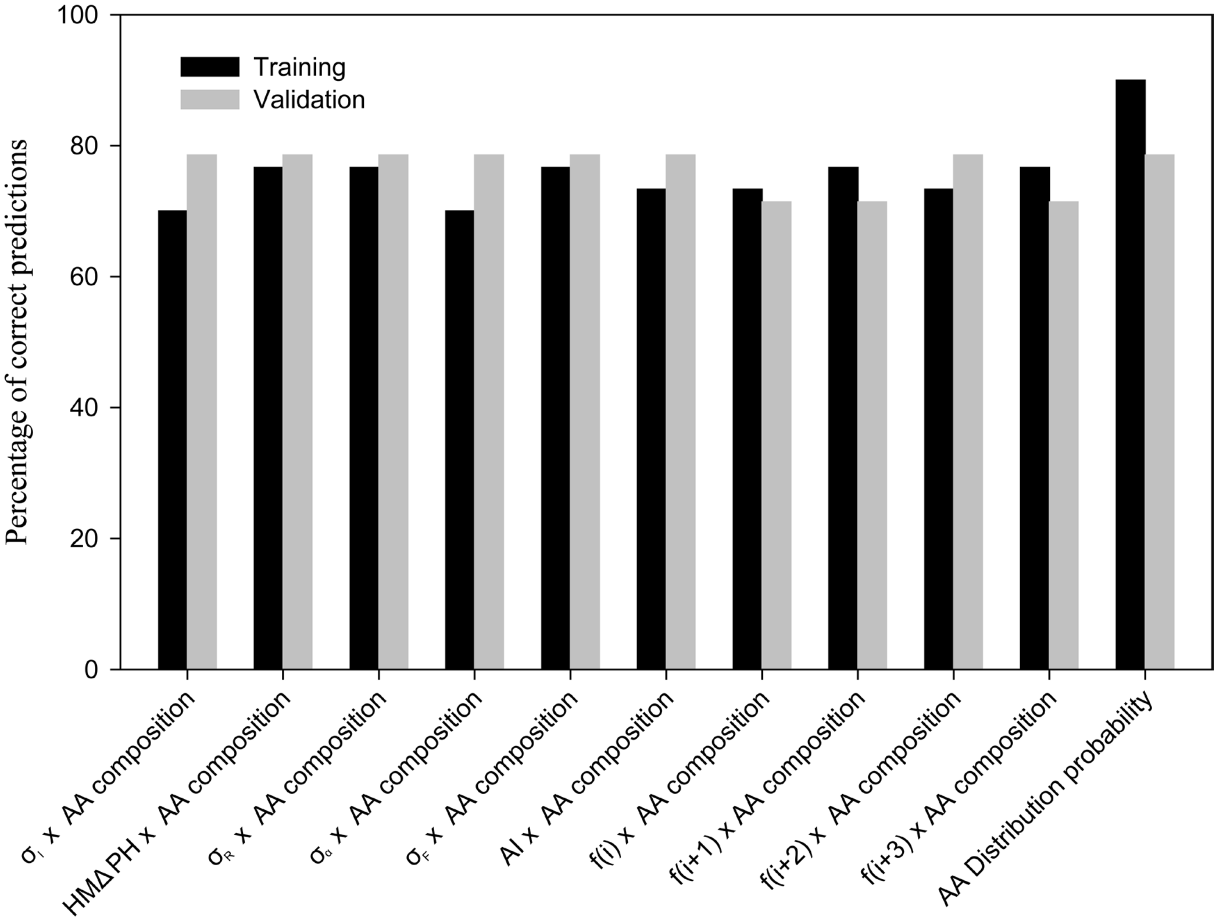

Figure 3. Percentage of correct predictions using different predictors.

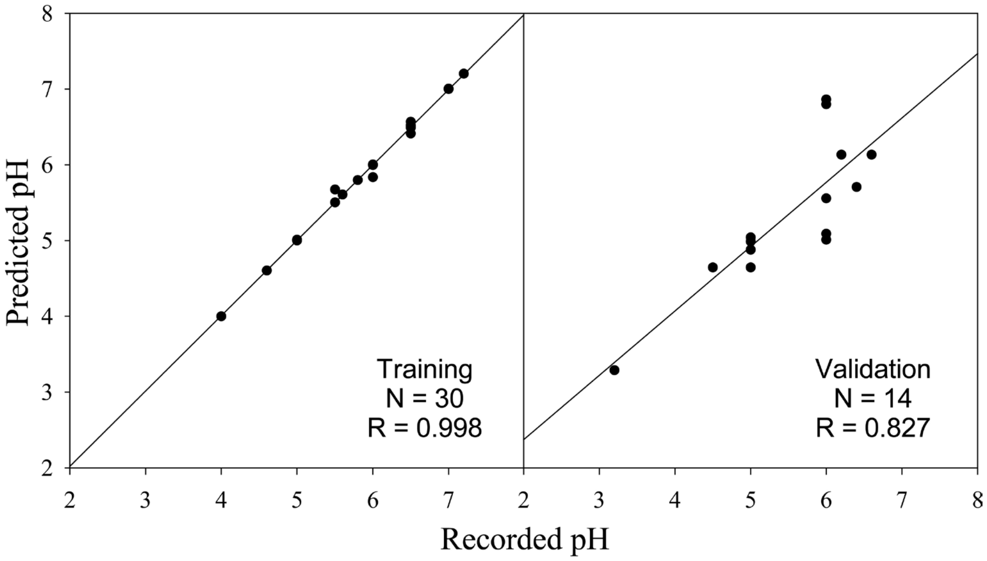

Figure 4. Linear regression between documented and predicted pH optimum of activity in training and validation groups with the amino-acid distribution probability as predictor. For training group, predicted pH = 0.9947 × recorded pH + 0.0308 ($P < 0.001$). For validation group, predicted pH = 0.8492 × recorded pH + 0.6748 ($P = 0.003$).

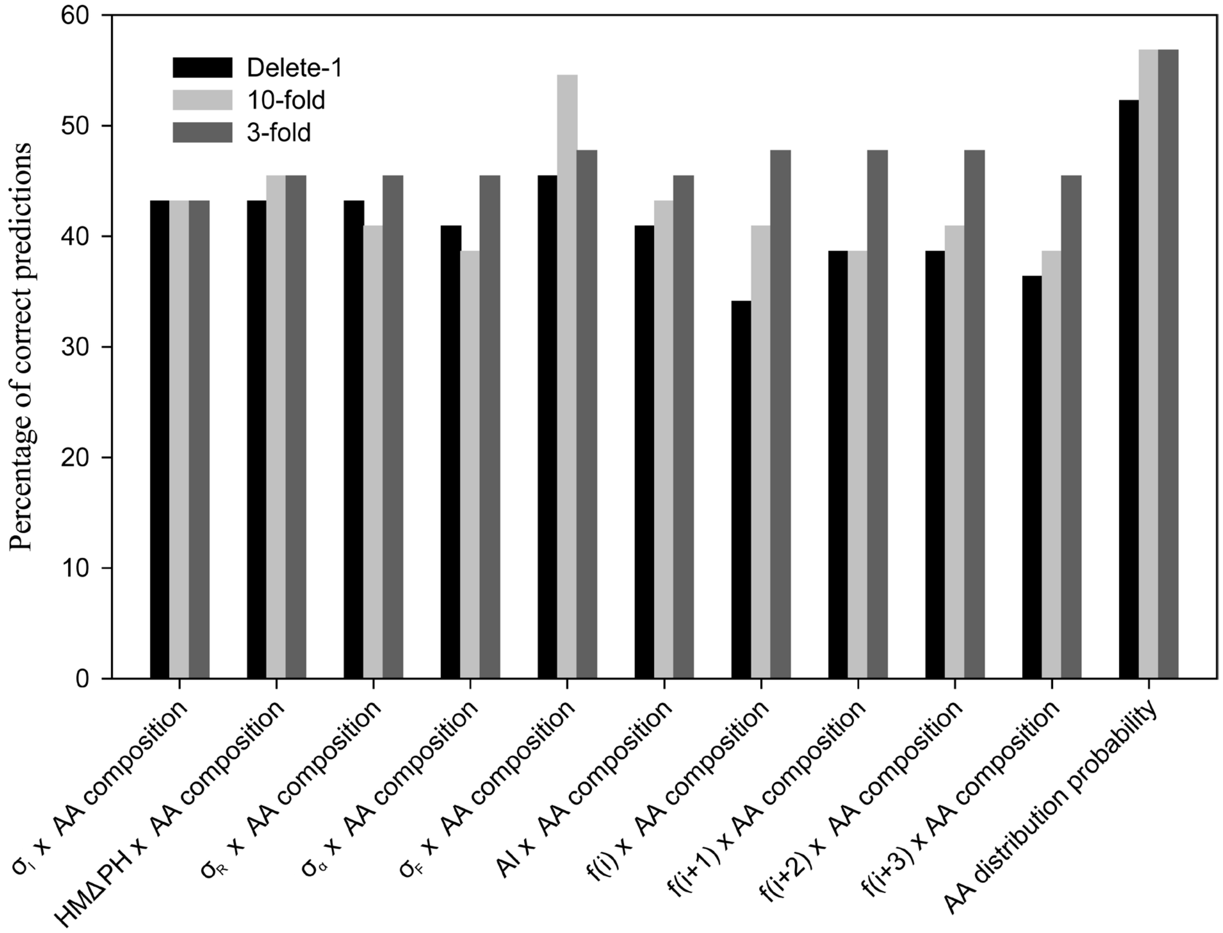

Figure 5. Percentage of correct predictions with delete-1 jackknife validation, 10-fold and 3-fold cross-validation using different predictors. AA represents amino acid.

best predictor is the amino-acid distribution probability.

The predictors used in this study include some related to the amino-acid based secondary structure of *β*-glucosidase; however, these features do not render better predictions than others. This may open the possibility to use the amino-acid features to predict various working conditions for enzymes, even the possibility to use the information about primary structure to predict the changing environments. This is so because other studies have confirmed that a physicochemical metric of charge distribution correlates better with subcellular pH [6]. The amino-acid composition is one of two factors that influence the pH of maximal protein stability [7] and can empirically model the pH optimum of protein-protein binding [8].

If we pay our attention to validation, the technique detail would puzzle us, that is, the Jackknife validation is worse than traditional validation by comparing **Figure 3** with **Figure 5**. This is interesting because the Jackknife uses almost all the samples to generate model parameters, but produces a worse prediction. This is very counter-intuitive, because the current knowledge indicates that the larger the trained data, the larger the chance that predicted sample would be included, the better the prediction. Clearly, this technique should require many more studies to deal with.

Currently it is not very clear whether different pH optimums would suggest that *β*-glucosidases would have different structures. If so, our prediction still falls into the so-called structure-function relationship; if not, our prediction would suggest a more sophisticated mechanism between structure and enzymatic working condition, *i.e.* structure-environment relationship.

In conclusion, this study suggests that we can use the features of amino acids of *β*-glucosidases to predict their pH optimum of activity. Among 25 amino-acid features screened, 11 can be serve as predictors to estimate the pH optimum, including 6/7 of the electronic properties and 4/7 of the amino-acid based secondary structure predictions, and the amino-acid distribution probability reaches better prediction than other predictors. However, the amino-acid composition, electronic properties and secondary structure predictions themselves cannot work in the neural network model, and they must be weighted with the amino-acid composition. Thus, the amino-acid distribution probability reveals its advantage in the prediction, indicating the random mechanisms may underline in enzyme reaction. The model provides the possibility to use the amino-acid features to predict various working conditions for enzymes.

FUND

This study was supported by National Natural Science Foundation of China (31560315), and Key Project of Guangxi Scientific Research and Technology Development Plan (AB17190534).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Wang, T., Wu, M.B., Lin, J.P. and Yang, L.R. (2015) Quantitative Structure-Activity Relationship: Promising Advances in Drug Discovery Platforms. *Expert Opinion on Drug Discovery*, **10**, 1283-1300. https://doi.org/10.1517/17460441.2015.1083006

2. Xue, L.C., Dobbs, D., Bonvin, A.M. and Honavar, V. (2015) Computational Prediction of Protein Interfaces: A Review of Data Driven Methods. *FEBS Letters*, **589**, 3516-3526. https://doi.org/10.1016/j.febslet.2015.10.003

3. Jeng, W.Y., Wang, N.C., Lin, M.H., Lin, C.T., Liaw, Y.C., Chang, W.J., Liu, C.I., Liang, P.H. and Wang, A.H.J. (2011) Structural and Functional Analysis of Three Beta-Glucosidases from Bacterium *Clostridium cellulovorans*, fungus *Trichoderma reesei* and Termite *Neotermes koshunensis*. *Journal of Structural Biology*, **173**, 46-56. https://doi.org/10.1016/j.jsb.2010.07.008

4. Sticklen, M. (2006) Plant Genetic Engineering to Improve Biomass Characteristics for Biofuels. *Current Opinion in Biotechnology*, **17**, 315-319. https://doi.org/10.1016/j.copbio.2006.05.003

5. Talley, K. and Alexov, E. (2010) On the pH-Optimum of Activity and Stability of Proteins. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **78**, 2699-706. https://doi.org/10.1002/prot.22786

6. Garcia-Moreno, B. (2009) Adaptations of Proteins to Cellular and Subcellular pH. *Journal of Biology*, **8**, 98. https://doi.org/10.1186/jbiol199

7. Alexov, E. (2004) Numerical Calculations of the pH of Maximal Protein Stability. The Effect of the Sequence Composition and Three-Dimensional Structure. *European Journal of Biochemistry*, **271**, 173-185. https://doi.org/10.1046/j.1432-1033.2003.03917.x

8. Mitra, R.C., Zhang, Z. and Alexov, E. (2011) In Silico Modeling of pH-Optimum of Protein-Protein Binding. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **79**, 925-936. https://doi.org/10.1002/prot.22931

9. Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J. and Schomburg, D. (2017) BRENDA in 2017: New Perspectives and New Tools in BRENDA. *Nucleic Acids Research*, **45**, D380-D388. https://doi.org/10.1093/nar/gkw952

10. Berrin, J.G., Czjzek, M., Kroon, P.A., McLauchlan, W.R., Puigserver, A., Williamson, G. and Juge, N. (2003) Substrate (Aglycone) Specificity of Human Cytosolic Beta-Glucosidase. *Biochemical Journal*, **373**, 41-48. https://doi.org/10.1042/bj20021876

11. Tsukada, T., Igarashi, K., Fushinobu, S. and Samejima, M. (2008) Role of Subsite +1 Residues in pH Depen-

dence and Catalytic Activity of the Glycoside Hydrolase Family 1 β-Glucosidase BGL1A from the Basidiomycete *Phanerochaete chrysosporium*. *Biotechnology and Bioengineering*, **99**, 1295-1302. https://doi.org/10.1002/bit.21717

12. UniProt Consortium (2019) UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Research*, **47**, D506-D515. https://doi.org/10.1093/nar/gky1049
13. Burlingame, A.L. and Carr, S.A. (1996) Mass Spectrometry in the Biological Sciences. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-4612-0229-5
14. Zamyatin, A.A. (1972) Protein Volume in Solution. *Progress in Biophysics & Molecular Biology*, **24**, 107-123. https://doi.org/10.1016/0079-6107(72)90005-3
15. Darby, N.J. and Creighton, T.E. (1993) Dissecting the Disulphide-Coupled Folding Pathway of Bovine Pancreatic Trypsin Inhibitor. Forming the First Disulphide Bonds in Analogues of the Reduced Protein. *Journal of Molecular Biology*, **232**, 873-896. https://doi.org/10.1006/jmbi.1993.1437
16. Kyte, J. and Doolittle, R.F. (1982) A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, **157**, 105-132. https://doi.org/10.1016/0022-2836(82)90515-0
17. Trinquier, G., Sanejouand, Y.H. and Hausman, R.E. (1998) Which Effective Property of Amino Acids Is Best Preserved by the Genetic Code? *Protein Engineering*, *Design and Selection*, **11**, 153-169. https://doi.org/10.1093/protein/11.3.153
18. Cooper, G.M. (2004) The Cell: A Molecular Approach. ASM Press, Washington DC, 51.
19. Dwyer, D.S. (2005) Electronic Properties of Amino Acid Side Chains: Quantum Mechanics Calculation of Substituent Effects. *BMC Chemical Biology*, **5**, 2. https://doi.org/10.1186/1472-6769-5-2
20. Chou, P.Y. and Fasman, G.D. (1978) Prediction of Secondary Structure of Proteins from Amino Acid Sequence. *Advances in Enzymology and Related Subjects of Biochemistry*, **47**, 45-148. https://doi.org/10.1002/9780470122921.ch2
21. Feller, W. (1968) An Introduction to Probability Theory and Its Applications. 3rd Edition, Wiley, New York.
22. Wu, G. and Yan, S.M. (2002) Randomness in the Primary Structure of Protein: Methods and Implications. *Molecular Biology Today*, **3**, 55-69.
23. Wu, G. and Yan, S. (2006) Mutation Trend of Hemagglutinin of Influenza a Virus: A Review from Computational Mutation Viewpoint. *Acta Pharmacologia Sinica*, **27**, 513-526. https://doi.org/10.1111/j.1745-7254.2006.00329.x
24. Wu, G. and Yan, S. (2006) Fate of Influenza a Virus Proteins. *Protein & Peptide Letters*, **13**, 377-384. https://doi.org/10.2174/092986606775974474
25. Yan, S. and Wu, G. (2010) Creation and Application of Computational Mutation. *Journal of Guangxi Academy of Sciences*, **17**, 145-150.
26. Wu, G. and Yan, S. (2008) Lecture Notes on Computational Mutation. Nova Science Publishers, New York.
27. Demuth, H. and Beale, M. (2001) Neural Network Toolbox for Use with MatLab. User's Guide, Version 4.
28. MathWorks Inc (1984-2001) MatLab—The Language of Technical Computing.
29. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition. *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024
30. Sokal, R.R. and Rohlf, F.J. (1995) Biometry: The Principles and Practices of Statistics in Biological Research. 3rd Edition, W. H. Freeman, New York, 203-218.