DOI:10.13656/j.cnki.gxkx.2011.03.031
广西科学 Guangxi Sciences 2011,18(3):253~260

Simultaneously Predicting Optimum pH Value and Optimum Temperature in Catalytic Reaction of Beta-glucosidase*

贝塔-葡萄糖苷酶催化反应的最适 pH 值和最适温度的同时预测

YAN Shao-min[1], SHI De-qiang[1], NONG Hao[1], WU Guang[2]**
严少敏[1], 师德强[1], 农 浩[1], 吴 光[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, Nanning, Guangxi, 530007, China; 2. DreamSciTech Consulting, Shenzhen, Guangdong, 518054, China)
(1.广西科学院非粮生物质国家重点实验室,国家非粮生物质能源工程技术研究中心,广西生物炼制重点实验室,广西南宁 530007; 2.深圳市追梦科技咨询有限公司,广东深圳 518054)

Abstract: The features of amino acids were used to simultaneously predict optimum pH value and optimum temperature of beta-glucosidases. Firstly, the beta-glucosidases were quantified by different features of amino acids as inputs, and their optimum pH value and optimum temperature were served as outputs; secondly, the training was conducted by means of 20-2 feedforward backpropagation neural network; finally, the validation was performed by three approaches, subset validation, jackknife validation, and cross-validation. The results showed that among 24 features of amino acids only 4 features worked in the prediction model and the amino-acid distribution probability as predictor gave better results. Thus, the method developed in this study paved the way towards the prediction of functional parameters of enzymes based on their amino-acid properties.

Key words: amino acid, beta-glucosidase, optimum pH value, optimum temperature, prediction working condition

摘要: 利用氨基酸特征同时预测贝塔-葡萄糖苷酶反应的最适 pH 值和最适温度。首先,用不同的氨基酸特征对贝塔-葡萄糖苷酶进行量化作为输入,以最适 pH 值和最适温度作为输出;然后用 20-2 前馈反向传播的神经网络进行训练;最后用子集验证、刀切验证和交叉验证三种方法进行验证。结果表明,在 24 个氨基酸特征中只有 4 个特征可以用于模型预测,其中氨基酸分布概率的预测结果优于其它指标,为基于氨基酸属性预测酶的功能参数提供了方法。

关键词: 氨基酸 贝塔-葡萄糖苷酶 最适 pH 值 最适温度 预测 工作条件
中图法分类号: Q814.9, Q939.9 文献标识码: A 文章编号: 1005-9164(2011)03-0253-08

In order to speed up enzymatic reaction, it is important to let the enzymatic reaction be at the optimal conditions. Generally the optimal working conditions are so elegant that we have to conduct many expensive and time-consuming experiments to find them. Although these costly optimal working conditions can serve as references to novel enzymes, the inter-relationship among optimal working conditions is oftentimes difficult to manage.

With the fast development in computational bi-

收稿日期: 2011-01-07
作者简介: 严少敏(1958-), 女, 博士, 研究员, 主要从事计算变异学和模型研究。
*This study was partly supported by Guangxi Academy of Sciences (08YJ6SW06) and Guangxi Science Foundation (0907016、0991013、0991006Z、1004606、1103111、2010GXNSFF013003 and 2010GXNSFA013046).
**通讯作者。

ology and bioinformatics, it would be possible to develop computational methods to predict the optimal working conditions for novel enzymes based on previously obtained optimal conditions.

Actually, many studies so far have been directed to the function-structure relationship of proteins including enzymes. However, this relationship has yet to connect with the optimal conditions for enzymatic reactions, and this relationship is more or less related to secondary, tertiary, and quaternary structures, whose determinations are also costly and time-consuming.

Thus, the challenge arises whether we can use very simple features, such as the features of amino acids, to predict optimal conditions of enzymes, because the prediction would help us economically optimize enzymatic reactions.

As β-glucosidase (EC 3.2.1.21) can cut the β-bond linkage in glucose molecules[1], it plays an important role in biological processes. More notably, it can degrade celluloses, which gives a great perspective in the fermentation of biomass into biofuels and leads to more efforts to not only search for β-glucosidases but also mutate β-glucosidases[2]. Consequently, we can find more and more β-glucosidases with annotations of their primary structures but without optimal conditions for enzymatic reactions.

Actually, there are many features of amino acids available[3,4], which could serve as predictors to predict the optimal conditions for β-glucosidase reactions. However, it is yet to know which feature is really useful. In this study, we attempt to find which of 24 amino-acid features can simultaneously predict optimum pH value and optimum temperature of β-glucosidases.

1 Materials and methods

1.1 Data

The data of β-glucosidases (EC 3.2.1.21) were obtained from the Comprehensive Enzyme Information System BRENDA up to October 2010[5]. Under the functional parameters of pH optimum and temperature optimum, 34 β-glucosidases had their sequence information, and one had documented its mutant[6]. Frequently, an enzyme can have several different optimum values of pH and temperature, which was the case for four β-glucosidases[7]. In total, we found 39 β-glucosidases with their sequence, optimum pH value and optimum temperature in this databank.

1.2 Predictive model

As there may be various linear and nonlinear relationships between the feature of amino acid and optimum pH value and optimum temperature of β-glucosidase, we used a 20-2 feedforward backpropagation neural network displayed in Fig. 1 to account for these relationships[8]. In this model, the first layer contained 20 neurons corresponding to 20 inputs, which could be any piece of features related to 20 types of amino acids in β-glucosidase. The second layer contained two neurons corresponding to two outputs that are optimum pH and optimum temperature. The transfer functions were tan-sigmoid and linear for two layers, and log-sigmoid for output. The training algorithm was the resilient backpropagation, which was the fastest algorithm on pattern recognition in MatLab[9].

1.3 Features of amino acids

The very basic feature of primary structure of β-glucosidase was its amino-acid composition, and then we had the molecular weights of amino acids (row 2, Table 1). Moreover, we had the features of amino acids related to the spatial properties listed in rows 3 ~ 5 in Table 1[10,11], hydrophobic properties listed in rows 6 ~ 10 in Table 1[12], electronic properties listed in rows 11 ~ 17 in Table 1[13], and the secondary structure predictions listed in rows 18 ~ 24 in Table 1[14]. We weighed the values in Table 1 with amino-acid composition because β-glucosidases were different one from another in terms of amino-acid compositions.

Another relatively recent feature about the primary structure was the amino-acid distribution probability, which also represented spatial characteristics on protein[15~17] and was computed according to the occupancy of subpopulations and partitions[18], which gives each type of amino acid a distribution probability in an enzyme (Table 2).

Table 1 Features of amino acids used as predictors for predicting optimum pH value and optimum temperature

Amino acid	A	R	N	D	C	E	Q	G	H	I
Mass(Dalton)	71.09	156.19	114.11	115.09	103.15	129.12	128.14	57.05	137.14	113.16
Surface area(Å^2)	115	225	160	150	135	190	180	75	195	175
Residue volume(Å^3)	166.7	88.6	173.4	114.1	111.1	108.5	138.4	143.8	60.1	153.2
van der Waals volume (Å^3)	67	148	96	91	86	114	109	48	118	124
Residue non-polar surface area(Å^2)	86	89	42	45	48	69	66	47	129	155
Residue burial (kcal/mol)	2.15	2.23	1.05	1.13	1.2	1.73	1.65	1.18	2.45	3.88
Side chain burial(kcal/mol)	1	1.1	−0.1	−0.1	0	0.5	0.5	0	1.3	2.7
Hydropathy index	4.50	4.20	−0.80	−0.90	−3.50	−0.70	−1.60	1.80	−3.90	−3.50
Ranking of amino acid polarities	9	15	16	19	7	18	17	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(a)	142	98	101	67	70	151	111	57	100	108
P(b)	83	93	54	89	119	37	110	75	87	160
P(turn)	66	95	146	156	119	74	98	156	95	47
f(i)	0.06	0.07	0.147	0.161	0.149	0.056	0.074	0.102	0.14	0.043
f(i+1)	0.076	0.106	0.11	0.083	0.05	0.06	0.098	0.085	0.047	0.034
f(i+2)	0.035	0.099	0.179	0.191	0.117	0.077	0.037	0.19	0.093	0.013
f(i+3)	0.058	0.085	0.081	0.091	0.128	0.064	0.098	0.152	0.054	0.056

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)	122	164	137	194	124	56	90	236	154	135
Residue burial (kcal/mol)	3.05	4.1	3.43	3.46	3.1	1.4	2.25	4.11	2.81	3.38
Side chain burial(kcal/mol)	1.9	2.9	2.3	2.3	1.9	0.2	1.1	2.9	1.6	2.2
Hydropathy index	−1.30	2.50	−0.40	−3.20	−3.50	2.80	1.90	4.50	3.80	−3.50
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(a)	121	114	145	113	57	77	83	108	69	106
P(b)	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 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.

Table 2 Amino-acid composition and distribution probabilities of Q9AT27 β-glucosidase.

Amino acid	Number	Distribution probability
A	54	0.015
R	19	0.017
N	21	0.027
D	42	8.903e-3
C	6	0.039
E	26	0.040
Q	23	0.040
G	35	5.699e-3
H	10	0.191
I	32	0.037
L	55	0.012
K	23	7.082e-4
M	14	0.055
F	24	0.040
P	26	4.712e-5
S	49	1.650e-5
T	48	8.611e-3
W	12	4.432e-3
Y	22	0.051
V	31	9.435e-3

1.4 Model development

Of 39 β-glucosidases listed in Table 3, 23 served as training group to generate the neural network model parameters, weights and biases, and 16 served as validation group. This is a very traditional approach.

A more recent approach is the jackknife, and the jackknifing of delete-1 observation was used[19], i.e. one β-glucosidase of 39 β-glucosidases did not attend the training, while the generated model parameters were used to predict optimum pH value and optimum temperature of omitted β-glucosidase, until that each β-glucosidase went through this jackknifing process.

The third approach is the cross-validation, i.e. 39 β-glucosidases were split into 3 subsets containing 13 β-glucosidases each, then one subset did not attend the training, but the generated model parameters were used to predict optimum pH value and optimum temperature in omitted subset, until that each subset went through this cross-validation process. Another way to divide 39 β-glucosidases was to split β-glucosidases into 13 subsets containing 3 β-glucosidases each, and to go to the same process as described above.

The above three approaches were applied to each predictor listed in Table 1 in order to compare their predictions statistically.

1.5 Statistics

For each predictor, one hundred trainings were conducted, and the obtained 100 sets of weights and biases were used to predict optimum pH value and optimum temperature 100 times, and their mean and standard deviation were used to compare the recorded optimum pH value and optimum temperature for each β-glucosidase[20].

2 Results

Fig. 1 is the scheme of neural network for model development. This model is particularly designed to simultaneously account for the features of amino-acids and optimum pH value and optimum temperature of β-glucosidases.

For training of neural network, the initialization of weights and biases and number of training epochs govern whether the neural network can converge. We used the random initialization function to initialize weights and biases, and 350 training epochs. Fig. 2 displays training processes in 23 β-glucosidases with different features of amino acids (Table 1). In this figure, each line represents a training process from the beginning to the end, and we can find the convergence reached within 350 training epochs with any random initialization, which lays the foundation to guarantee our training process.

In Fig. 3, we can see that the percentage of correctly predicted β-glucosidases improves with respect to training epochs. Actually, what we need to see is whether this percentage is stable along the training process, which is the case in Fig. 3, so we can exclude the possibility of over-fitting or over-generalization using this neural network model. Also, Fig. 3 demonstrates that predictions of optimum pH value and optimum temperature using some features of amino acids can reach a pretty good level. As we used three different approaches to develop predictive models, Fig. 3 is only related to one approach, and Fig. 4 shows the correctly predicted percentage improves with respect to training epochs in the other two approaches. In general, all three approaches reach the similar results.

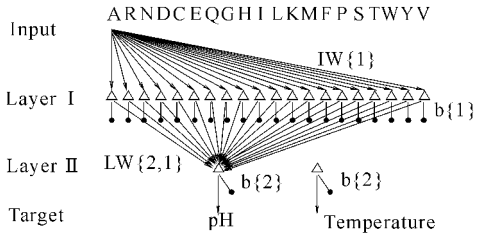

Fig. 1 20-2 feedforward backpropagation neural network to model the relationship between 20 amino-acid features of β-glucosidase and optimum pH value and optimum temperature. Each tri-circle represents a neuron.

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.

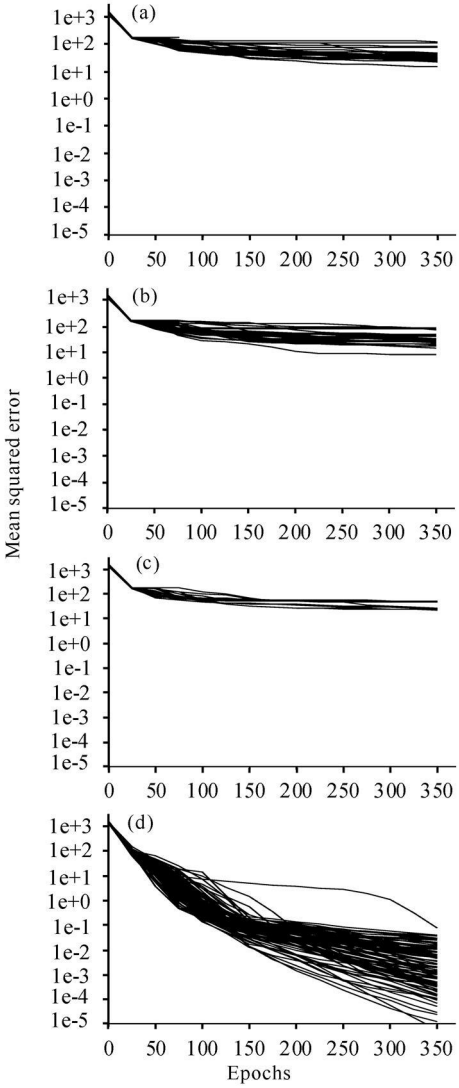

Fig. 2 Convergence of mean squared error performance function with 100 different initial weights and biases generated by random initialization function in training

(a)$\sigma_1\times$ No., (b)$\sigma_\alpha\times$ No., (c)f(i)$\times$ No., (d)DP.

No.: amino-acid composition, σ_1: inductive effect scale, σ_α: normalized polarizability index, f(i): frequency of the 1st residue in turn, DP: amino-acid distribution probability.

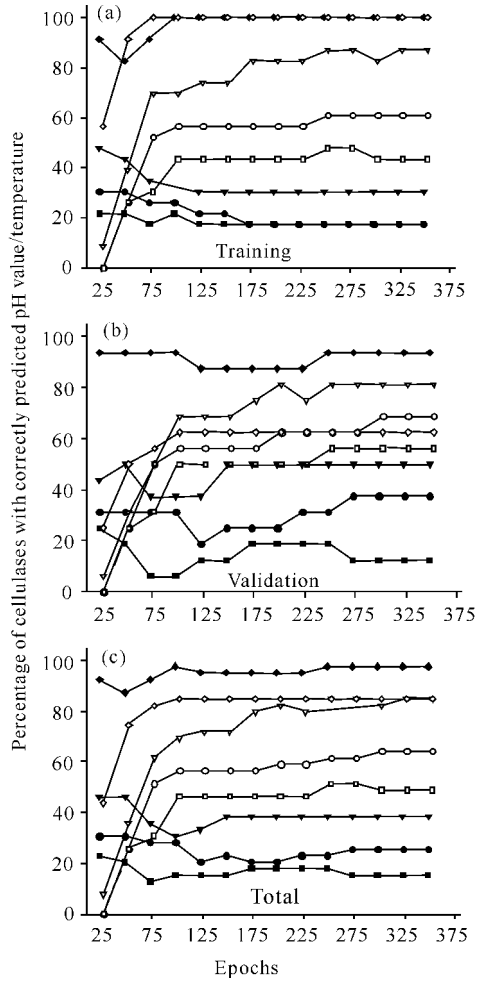

Fig. 3 Percentage of correctly predicted optimum pH value and optimum temperature by different features of amino acids

(a) Training, (b) Validation, (c) Total.

The training and validation groups contain 23 and 16 β-glucosidases.

—●—: pH by $\sigma_1\times$ No., —▼—: pH by $\sigma_\alpha\times$ No., —■—: pH by f(i)$\times$ No., —◆—: pH by DP, —○—: Tm by $\sigma_1\times$ No., —▽—: Tm by $\sigma_\alpha\times$ No., —□—: Tm by f(i)$\times$ No., —◇—: Tm by DP.

pH: optimum pH value, Tm: optimum temperature, No.: amino-acid composition, σ_1: inductive effect scale, σ_α: normalized polarizability index, f(i): frequency of the 1st residue in turn, DP: amino-acid distribution probability, $\times$: multiplication.

Table 3 details the statistical comparison between predicted and recorded optimum pH value and optimum temperature. The data generated in Table 3 are based on the very traditional approach to divide the training and validation groups, where the prediction should be good if there is no statistical difference between recorded and predicted optimum pH value and optimum temperature, respectively. The prediction based on amino-acid distribution probability is clearly better than the predictions based on other features of amino acids.

Table 3 Statistical comparison between recorded and predicted optimum pH value and optimum temperature (mean±*SD*, *n*= 100).

Group	Accession number	Recorded pH value	Optimum pH value predicted by predictor				Recorded Tm	Optimum temperature predicted by predictor			
			σ_1× No.	σ_α× No.	f(i) x No.	DP		σ_1× No.	σ_α× No.	f(i)× No.	DP
Training	Q9AT27	4	5.83±0.10	5.86±0.14	5.83±0.09	4.05±0.13*	50	46.39±4.92*	44.64±5.73*	47.43±4.24*	50.00±0.02*
	Q8TGI8	4	5.70±0.21	5.57±0.32	5.73±0.18	4.35±0.35*	71.5	52.80±6.86	56.58±8.82*	51.74±6.40	71.48±0.07*
	Q12715	4.6	5.64±0.28	5.54±0.34	5.71±0.21	4.66±0.25*	70	56.34±11.80*	58.90±11.34*	52.66±8.53	69.91±0.76*
	A1C3J9	5	5.82±0.08	5.85±0.09	5.81±0.07	5.00±0.13*	40	47.47±3.05	45.71±4.08*	47.98±2.97	40.00±0.08*
	Q8T0W7	5	5.75±0.11	5.77±0.13	5.78±0.07	5.07±0.23*	50	51.94±5.17*	51.19±3.35*	50.02±2.70*	50.10±0.78*
	Q4U4W7	5	5.74±0.14	5.55±0.35*	5.71±0.21	5.03±0.12*	50	50.77±5.17*	57.29±10.90*	52.13±7.98*	50.01±0.03*
	A9UIG0	5	5.66±0.26	5.54±0.34*	5.71±0.21	4.93±0.25*	70	55.32±10.19*	58.71±11.27*	52.69±8.67*	70.02±0.10*
	P94248	5.5	5.85±0.11	5.90±0.16	5.83±0.10	5.51±0.11*	45	45.80±5.43*	43.37±6.56*	47.26±4.54*	45.00±0.04*
	O08331	5.5	5.73±0.16*	5.62±0.26*	5.73±0.16*	5.51±0.15*	65	53.63±7.55*	56.21±8.20*	52.03±6.73*	65.00±0.02*
	Q9SLA0	5.6	5.84±0.11	5.89±0.13	5.83±0.09	5.64±0.12*	40	46.14±5.19*	44.31±5.44*	47.43±4.12*	40.00±0.02*
	P49235	5.8	5.83±0.09*	5.83±0.11*	5.81±0.06*	5.67±0.19*	50	46.56±4.56*	46.12±4.46*	48.19±3.03*	50.00±0.04*
	Q86D78	6	5.85±0.11*	5.91±0.16*	5.83±0.10*	6.04±0.12*	35	45.69±5.60*	43.31±6.55*	47.28±4.48	35.00±0.03*
	Q2WGB4	6	5.83±0.08	5.88±0.12*	5.82±0.09	5.99±0.09*	37	46.94±3.78	44.76±5.01*	47.65±3.77	37.00±0.03*
	Q875K3	6	5.82±0.11*	5.87±0.12*	5.83±0.09*	5.98±0.09*	40	46.73±5.47*	44.24±5.57*	47.38±4.26*	40.00±0.02*
	Q25BW5	6.5	5.86±0.15	5.91±0.18	5.83±0.09	6.47±0.13*	30	45.91±5.47	44.39±5.90	47.45±4.09	30.01±0.05*
	P15885	6.5	5.85±0.12	5.91±0.16	5.83±0.10	6.46±0.11*	30	45.68±5.49	43.18±6.75*	47.26±4.52	30.00±0.03*
	Q59976	6.5	5.86±0.19	5.91±0.19	5.83±0.09	6.54±0.11*	30	46.03±6.29	43.89±7.00*	47.45±4.09	30.00±0.03*
	Q9H227	6.5	5.84±0.09	5.88±0.14	5.83±0.10	6.43±0.12*	50	46.08±4.95*	44.48±5.64*	47.36±4.30*	50.01±0.04*
	Q746L1	6.5	5.70±0.24	5.68±0.34	5.73±0.21	6.44±0.21*	88	57.95±15.12*	62.97±16.16*	53.58±10.37	87.99±0.04*
	B9K7M5	6.5	5.68±0.25	5.68±0.37	5.73±0.21	6.35±0.21*	95	58.47±14.94	64.33±17.67*	53.58±10.59	94.96±0.18*
	Q08IT7	7	5.85±0.13	5.89±0.15	5.82±0.08	6.93±0.12*	30	46.20±5.24	44.36±5.71	47.74±3.61	30.00±0.03*
	Q6QGY5	7	5.85±0.11	5.89±0.14	5.83±0.10	6.98±0.11*	40	45.91±5.09*	43.97±5.75*	47.29±4.46*	40.01±0.04*
	Q47RE2	7.2	5.85±0.15	5.90±0.17	5.83±0.09	7.14±0.13*	25	46.08±5.94	43.87±6.77	47.39±4.22	24.99±0.06*
Validation	Q08638	3.2	5.69±0.24	5.67±0.37	5.74±0.21	5.71±1.26*	85	58.43±15.30*	64.99±18.80*	53.66±10.85	67.71±14.73*
	B5TWK3	4.5	5.62±0.31	5.51±0.36	5.71±0.22	5.12±0.90*	22	58.70±17.26	61.04±16.14	52.92±9.54	59.08±9.07
	Q12715	4.6	5.64±0.28	5.54±0.34	5.71±0.21	4.66±0.25*	65	56.34±11.80*	58.90±11.34*	52.66±8.53*	69.91±0.76
	B5TWK3	5	5.62±0.31	5.51±0.36*	5.71±0.22	5.12±0.90*	37	58.70±17.26*	61.04±16.14*	52.92±9.54*	59.08±9.07
	B6ZKM3	5	5.81±0.14	5.79±0.27	5.78±0.09	5.92±1.59*	30	48.50±5.77	51.01±10.65*	49.31±3.72	49.29±17.14*
	Q2UUD6	5	5.68±0.20	5.57±0.32*	5.72±0.18	4.96±0.66*	60	54.37±10.84*	56.98±11.75*	51.92±7.17*	60.20±4.11*
	Q9SPK3	5	5.84±0.11	5.90±0.15	5.83±0.09	6.09±0.98*	50	46.03±5.38*	43.75±6.48*	47.32±4.38*	62.34±9.92*
	A9UIG0	6	5.66±0.26*	5.54±0.34*	5.71±0.21*	4.93±0.25	70	55.32±10.19*	58.71±11.27*	52.69±8.67*	70.02±0.10*
	O61594	6	5.83±0.09*	5.87±0.11*	5.82±0.08	5.53±1.29*	30	46.64±4.28	44.75±5.14	47.88±3.89	60.57±10.65
	Q12601	6	5.80±0.19*	5.78±0.25*	5.79±0.12*	5.26±1.00*	35	48.04±5.40	48.37±8.17*	48.83±4.06	60.97±10.07
	P26208	6	5.85±0.18*	5.89±0.26*	5.81±0.10*	6.84±1.22*	65	47.14±6.83	48.22±8.77*	48.77±4.16	59.84±14.00*
	P10482	6	5.80±0.13*	5.79±0.16*	5.79±0.05	5.54±1.55*	80	49.34±5.86	50.58±5.71	49.24±2.64	57.36±17.67*
	P96316	6.2	5.68±0.23	5.57±0.34*	5.71±0.20	5.76±1.23*	35	55.81±14.34*	61.65±16.69*	53.46±12.27*	65.98±11.55
	Q9C3Z9	6.4	5.70±0.20	5.59±0.32	5.71±0.20	5.36±1.05*	50	54.70±12.02*	60.20±16.63*	53.20±10.99*	73.78±13.01*
	Q9H227 V168Y	6.5	5.84±0.10	5.88±0.14	5.83±0.10	6.51±0.30*	50	46.03±4.97*	44.40±5.66*	47.35±4.32*	47.99±2.88*
	P96316	6.6	5.68±0.23	5.57±0.34	5.71±0.20	5.76±1.23*	45	55.81±14.34*	61.65±16.69*	53.46±12.27*	65.98±11.55*
Total performance	—	—	9	15	7	38	—	24	33	20	33

No.: amino-acid composition, σ_1: inductive effect scale, σ_α: normalized polarizability index, f(i): frequency of the 1st residue in turn, DP: amino-acid distribution probability, Tm: temperature, ×: multiplication.

3 Discussion

The model used in this study can account for any possible interaction between pH value and temperature if such an interaction would exist. Statistically, the two-way ANOVA could detect a possible interaction between pH value and temperature although the available data should be well designed

for this purpose, which are not the case for the data in this study. So, the neural network model has a big advantage over other models, which usually account for a single predicted variable.

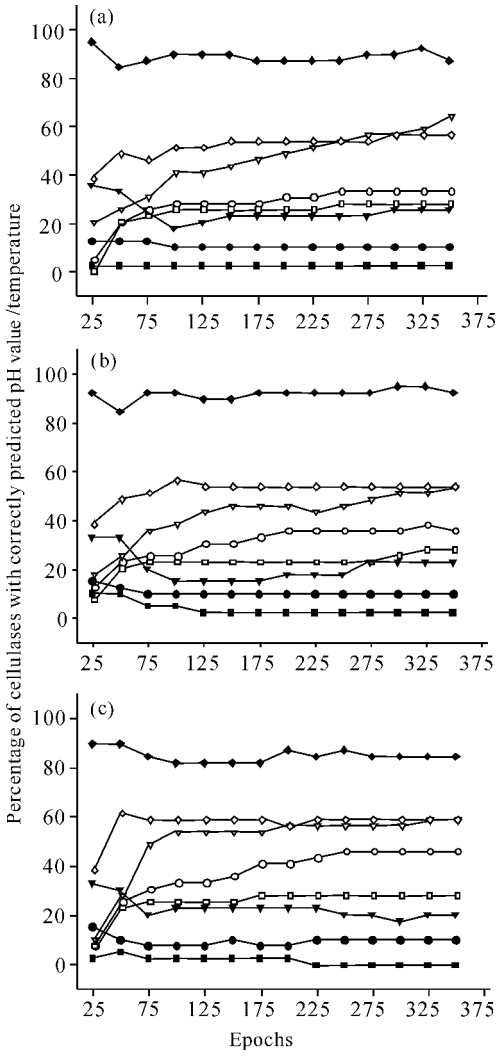

Fig. 4 Percentage of correctly predicted optimum pH value and optimum temperature by different amino-acid features

(a) Jack Knifing of delete-1 glucosidase, (b) 13-fold cross-validation, (c) 3-fold cross-validation.

—●—: pH by $\sigma_1 \times$ No., —▼—: pH by $\sigma_\alpha \times$ No., —■—: pH by f(i)× No., —◆—: pH by DP, —○—: Tm by $\sigma_1 \times$ No., —▽—: Tm by $\sigma_\alpha \times$ No., —□—: Tm by f(i)× No., —◇—: Tm by DP.

pH: optimum pH value, Tm: optimum temperature, No.: amino-acid composition, σ_1: inductive effect scale, σ_α: normalized polarizability index, f(i): frequency of the 1st residue in turn, DP: amino-acid distribution probability, ×: multiplication.

Actually, the prediction of optimal working condition for enzymes is an understudied area, thus it is important to develop methods along this line of studies. Experimentally and practically, it is important to develop methods to use as simple information as possible to predict the optimal working condition for enzymes.

For an experimentalist, it would be easier to measure optimum pH value as well as optimum temperature than to predict. However, it is only the model that can provide the basis for generalization. Moreover, the model would provide the basis for simulation of catalytic reaction using computer. Thus, our study can be considered as a small step towards such direction.

The results suggest that the amino-acid distribution probability appears better than other features of amino acids, which is reasonable because it is mainly related to amino-acid spatial distribution. Nevertheless, more studies are needed in order to better predict the optimal working conditions in different enzymes.

Acknowledgements

The authors wish to thank the Library of Guangxi Zhuang Autonomous Region for purchasing the book, Biometry.

References:

[1] Jeng W Y, Wang N C, Lin M H, et al. Structural and functional analysis of three beta-glucosidases from bacterium *Clostridium cellulovorans*, fungus *Trichoderma reesei* and termite *Neotermes koshunensis* [J]. J Struc Biol, 2011, 173: 46-56.

[2] Sticklen M. Plant genetic engineering to improve biomass characteristics for biofuels[J]. Curr Opin Biotechnol, 2006, 17: 315-319.

[3] Fasman G D. Handbook of biochemistry: section D physical chemical data[M]. Third ed. London and New York: CRC Press, 1976.

[4] Kawashima S, Pokarowski P, Pokarowska M, et al. Aaindex: amino acid index database, progress report 2008 [J]. Nucleic Acids Res, 2008, 36: D202-D205.

[5] Department of Bioinformatics and Biochemistry, Technical University of Braunschweig. The comprehensive enzyme information system BRENDA[EB/OL]. [2010-10] http://www.brenda-enzymes.info/php/result_flat.php4?ecno=3.2.1.4, 2010.

[6] Berrin J G, Czjzek M, Kroon P A, et al. Substrate (aglycone) specificity of human cytosolic beta-glucosidase [J]. Biochem J, 2003, 373: 41-48.

[7] Harnpicharnchai P, Champreda V, Sornlake W, et al. A thermotolerant beta-glucosidase isolated from an endophytic fungi, *Periconia* sp. with a possible use for biomass conversion to sugars[J]. Protein Expr Purif, 2009, 67: 61-69.

[8] Demuth H, Beale M. Neural network toolbox for use with MatLab. User's guide, version 4, 2001.

[9] MathWorks Inc. MatLab-the language of technical computing (version 6. 1. 0. 450, release 12. 1), 1984-2001 [CP]. 2001.

[10] Zamyatin A A. Protein volume in solution[J]. Prog Biophys Mol Biol, 1972, 24: 107-123.

[11] Darby N J, Creighton T E. Dissecting the disulphide-coupled folding pathway of bovine pancreatic trypsin inhibitor. Forming the first disulphide bonds in analogues of the reduced protein[J]. J Mol Biol, 1993, 232: 873-896.

[12] Cooper G M. The cell: a molecular approach[M]. Washington, D. C: ASM Press, 2004: 51.

[13] Dwyer D S. Electronic properties of amino acid side chains: quantum mechanics calculation of substituent effects[J]. BMC Chem Biol, 2005, 5: 2.

[14] Chou P Y, Fasman G D. Prediction of secondary structure of proteins from amino acid sequence[J]. Adv Enzymol Relat Subj Biochem, 1978, 47: 45-148.

[15] Wu G, Yan S M. Randomness in the primary structure of protein: methods and implications[J]. Mol Biol Today, 2002, 3: 55-69.

[16] Wu G, Yan S M. Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint[J]. Acta Pharmacol Sin, 2006, 27: 513-526.

[17] Wu G, Yan S M. Lecture notes on computational mutation[M]. New York: Nova Science Publishers, 2008.

[18] Feller W. An introduction to probability theory and its applications[M]. 3rd ed. Vol. I. New York: Wiley, 1968.

[19] Chou K C, Shen H B. Cell-PLoc 2. 0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms[J]. Natural Sci, 2010, 2: 1090-1103.

[20] Sokal R R, Rohlf F J. Biometry: The principles and practices of statistics in biological research[J]. 3rd ed. W H Freeman: New York, 1995, 203-218.

(责任编辑: 陈小玲)

科学家研制新型催化剂让二氧化碳变成低成本液态燃料

太阳是地球上主要的能量来源，更好地利用丰富的阳光是所有新能源专家试图摘取的“圣杯”。科学家很早就知道如何将水和二氧化碳转变为氢气和一氧化碳。但是如何高效、批量、而且低廉地转换一直困扰着科学家。其中的一个“拦路虎”，是转换过程需要昂贵且稀有的铂或铱等元素来作催化剂，以促使反应发生。最近科学家将目光投向了二氧化铈，金属铈的氧化物二氧化铈常用于自洁烤箱内壁，可作催化剂使用，铈储量丰富，转换成本低。

经过反复实验尝试，科学家们研究开发出一种太阳能反应器。该太阳能反应器采用低成本的二氧化铈作为催化剂集中太阳的热量，当将二氧化铈加热至约 1500 摄氏度高温时，会自动地从其结构内释放出氧气；接着将其冷却，氧气离开后留下的空白需要新氧气来填满。在约为 900 摄氏度的较低温度时，铈、氢气和碳都需要氧气，但是铈的需求更强烈，于是，它就会从水和二氧化碳中“掠夺”氧气来填满这些空白，因此，水和二氧化碳就变成了氢气和一氧化碳。大量的氢气和一氧化碳结合在一起可形成液态燃料，为汽车、手提电脑和全球定位系统(GPS)供电。

但是，目前这个将太阳光、二氧化碳和水转变为液态燃料的反应器的转换效率不足 1%。科学家表示，理论上反应器的转换效率可达 15%以上。此外，科学家也希望能找到比二氧化铈更好的燃料，降低发生反应所需要的高温和低温。

(据科学网)