Appl Biochem Biotechnol (2011) 165:856–869
DOI 10.1007/s12010-011-9303-2

Searching of Predictors to Predict pH Optimum of Cellulases

Shaomin Yan · Guang Wu

Received: 18 January 2011 / Accepted: 1 June 2011 /
Published online: 14 June 2011

Abstract The optimal working conditions for enzymes are very much elegant, and their determination is often through experimental approach, which generally is costly and time-consuming. Therefore, it is important to develop methods to use as simple as possible information to predict the optimal working condition for enzymes. Cellulase is a very important enzyme widely used in industries. In this study, we attempted to use a 20–1 feedforward backpropagation neural network to screen 24 amino acid properties related to the primary structure of cellulases as predictors to predict the pH optimum in cellulases. The results show that some predictors can predict the pH, especially amino acid distribution probability.

Keywords Cellulase · Enzyme · pH optimum · Prediction

Introduction

The optimal working conditions for enzymes are very much elegant, which leads to great scientific efforts in finding out their optimums, whose determinations generally go through the experimental approaches. However, not only these experiments are costly and time-consuming but also the experimental speed apparently lags the speed of increase of enzymes in database. For example, there are more than 5,000 of different enzymes listed currently in the Comprehensive Enzyme Information System BRENDA, whereas there were only 789 in 2002 [1, 2]. Therefore, we can easily find many enzymes with detailed information on their structures, mainly primary structure, but fewer enzymes with their

S. Yan · G. 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

G. Wu (✉)
DreamSciTech Consulting, 301, Building 12, Nanyou A-zone, Jiannan Road, Shenzhen, Guangdong 518054, China
e-mail: hongguanglishibahao@yahoo.com

optimal working conditions, which are generally defined by various parameters such as *Km*, p*K*i, pH, and so on.

Therefore, it is important to develop methods to predict various parameters, which are the indicators for enzymatic working conditions, in un-annotated enzymes. Still, it is more important to use very simple information related to primary structure to predict the optimal working conditions for enzymes because the prediction based on high-level structure might need more costly and time-consuming experiments to determine the secondary, tertiary, and quaternary structure.

Cellulase is an important enzyme, which hydrolyzes the beta-1,4 linkages of cellulose, and is widely used in industries [3]. Very recently, the cellulase becomes the focus of biofuel industry [4–7] because biofuel has a great prospective in environment-friendly economic development [8–10].

Of various parameters related to enzymatic working condition, we are interested in the prediction of pH optimum because of its importance in enzymatic reactions of cellulases. Furthermore, we hope to use the information from primary structure of cellulases to predict their pH optimum because this approach would be the most cost-effectiveness. Actually, many pieces of information on primary structure of proteins are readily available, for example, amino acid composition, physicochemical property of amino acid, and so on and so forth. In this study, we attempted to find out which piece of information is more useful to predict the pH optimum of cellulases.

Materials and Methods

Data

The data of cellulase (EC 3.2.1.4) are obtained from the Comprehensive Enzyme Information System BRENDA up to October 2010 (http://www.brenda-enzymes.info/php/result_flat.php4?ecno=3.2.1.4), where 32 cellulases have their sequence information under the category of pH optimum as functional parameters, of which two cellulases are documented with their mutants [11]. For example, three reduced-size mutants and 15 site-directed mutageneses were obtained from the cellulose O58925 isolated from *Pyrococcus horikoshii* [12,13]. Also, two pH values are documented in each of the cellulose O58925, Q6BCL3, Q6BCL7, Q6BCM1, and Q9PF60, respectively. In total, this databank provides 55 matched sequences and pH values of cellulases (column 2 Table 1). The amino acid sequences of cellulases are obtained from the Universal Protein Resource [14].

Although we had the intention to find out more cellulases with their pH optimum, the reality is not what we hoped. This once again shows the importance to develop the methods to predict the pH optimum in numerous cellulases because so many cellulases have no documented pH optimum.

Prediction

As the aim of this study is designed to find out the useful information to predict the pH optimum, thus we use the neural network as predictive model and the piece of information from primary structure of cellulases to predict their pH optimum because the relationship between pH optimum and information from primary structure varies greatly.

This model is a 20–1 feedforward backpropagation neural network [15, 16], whose structure is shown in Fig. 1. In this model, the first layer contains 20 neurons corresponding

Table 1 Comparison of recorded pH optimum with predicted pH optimum that is presented as mean±SD of 100 predictions

Predictor	Accession number	Recorded pH	Predicted pH											
			No.	σ_I×No.	$H_M\Delta$PH×No.	σ_R×No.	σ_α×No.	σ_F×No.	A_1×No.	$f(i)$×No.	$f(i+1)$×No.	$f(i+2)$×No.	$f(i+3)$×No.	Amino acid distribution probability
Training	A4UU22	3.5	3.90±0.86[a]	3.46±0.06[a]	3.48±0.06[a]	3.48±0.04[a]	3.46±0.06[a]	3.46±0.10[a]	3.47±0.06[a]	3.46±0.08[a]	3.45±0.05[a]	3.45±0.15[a]	3.90±0.86[a]	3.50±0.02[a]
	Q97YG7	3.5	3.88±0.83[a]	3.49±0.11[a]	3.40±0.34[a]	3.47±0.23[a]	3.49±0.04[a]	3.50±0.04[a]	3.49±0.08[a]	3.50±0.08[a]	3.50±0.03[a]	3.48±0.13[a]	3.88±0.83[a]	3.49±0.02[a]
	Q25C15	5.0	5.14±0.34[a]	4.98±0.06[a]	4.98±0.06[a]	4.97±0.06[a]	4.98±0.05[a]	5.00±0.07[a]	4.97±0.05[a]	4.98±0.09[a]	4.97±0.05[a]	4.96±0.14[a]	5.14±0.34[a]	5.00±0.03[a]
	O58925 C106A/C159A/C372A/C412A	5.0	5.11±0.37[a]	4.84±0.08[a]	4.85±0.08[a]	4.86±0.09[a]	4.85±0.07	4.85±0.08[a]	4.84±0.07	4.85±0.09[a]	4.85±0.06	4.83±0.14[a]	5.11±0.37[a]	4.93±0.08[a]
	O58925 C372A/C412A	5.0	5.46±0.21	5.39±0.06	5.39±0.07	5.41±0.07	5.37±0.06	5.36±0.08	5.38±0.06	5.38±0.10	5.38±0.06	5.36±0.16	5.46±0.21	5.03±0.06[a]
	P23044	5.0	5.13±0.34[a]	5.03±0.07[a]	5.02±0.06[a]	5.03±0.06[a]	5.04±0.05[a]	5.05±0.11[a]	5.04±0.07[a]	5.03±0.08[a]	5.02±0.06[a]	5.02±0.14[a]	5.13±0.34[a]	5.01±0.02[a]
	Q9PF60	5.0	5.11±0.35[a]	4.94±0.05[a]	4.95±0.05[a]	4.97±0.06[a]	4.92±0.04[a]	4.95±0.06[a]	4.95±0.05[a]	4.98±0.08[a]	4.97±0.04[a]	4.97±0.13[a]	5.11±0.35[a]	5.00±0.03[a]
	O58925	5.5	5.89±0.08	5.93±0.03	5.94±0.03	5.95±0.03	5.92±0.03	5.91±0.04	5.93±0.04	5.92±0.07	5.91±0.03	5.91±0.14	5.89±0.08	6.67±0.14
	O58925 R102A	5.5	5.83±0.08	5.85±0.03	5.83±0.04	5.80±0.04	5.83±0.03	5.83±0.04	5.85±0.03	5.84±0.08	5.84±0.03	5.82±0.13	5.83±0.08	5.49±0.08[a]
	O58925 E201A	5.5	5.83±0.09	5.86±0.04	5.87±0.04	5.86±0.04	5.88±0.03	5.82±0.05	5.87±0.03	5.86±0.08	5.86±0.04	5.85±0.14	5.83±0.09	5.67±0.15[a]
	O58925 E201Q	5.5	5.81±0.09	5.84±0.06	5.84±0.05	5.88±0.05	5.84±0.05	5.82±0.06	5.83±0.06	5.82±0.09	5.82±0.05	5.81±0.14	5.81±0.09	5.48±0.12[a]
	O58925 W377A	5.5	5.84±0.12	5.93±0.07	5.94±0.07	5.94±0.07	5.91±0.08	5.88±0.08	5.92±0.08	5.91±0.09	5.93±0.07	5.90±0.15	5.84±0.12	5.40±0.07[a]
	O58925 D385N	5.5	5.90±0.09	5.94±0.06	5.94±0.06	5.94±0.07	5.94±0.05	5.93±0.07	5.95±0.06	5.94±0.08	5.94±0.04	5.93±0.15	5.90±0.09	5.83±0.21[a]
	O58925 DEL C5	5.5	5.62±0.14[a]	5.56±0.09[a]	5.51±0.10[a]	5.48±0.12[a]	5.54±0.09[aa]	5.58±0.11[a]	5.56±0.09[a]	5.57±0.11[a]	5.55±0.08[a]	5.54±0.15[a]	5.62±0.14[a]	5.56±0.08[a]
	O58925 DEL N5&C5	5.5	5.55±0.17[a]	5.48±0.11[a]	5.42±0.10[a]	5.42±0.10[a]	5.47±0.10[a]	5.49±0.11[a]	5.47±0.11[a]	5.48±0.12[a]	5.47±0.09[a]	5.44±0.17[a]	5.55±0.17[a]	5.49±0.02[a]
	O58925 Y299A	5.5	5.94±0.09	5.99±0.04	5.99±0.04	5.97±0.03	5.98±0.04	5.98±0.05	5.99±0.04	5.97±0.07	5.97±0.03	5.97±0.14	5.94±0.09	5.64±0.07[a]
	O58925 DELTAQ1-G5	5.5	5.82±0.20[a]	5.70±0.13[a]	5.73±0.14[a]	5.67±0.15[a]	5.77±0.14[a]	5.75±0.14[a]	5.72±0.13[a]	5.78±0.14[a]	5.78±0.13	5.75±0.19[a]	5.82±0.20[a]	5.49±0.05[a]
	Q6BCL3	5.8	6.08±0.18[a]	6.07±0.03	6.08±0.03	6.06±0.04	6.07±0.03	6.06±0.06	6.07±0.04	6.07±0.07	6.07±0.03	6.05±0.14[a]	6.08±0.18[a]	6.10±0.01
	Q6BCL7	5.8	5.97±0.22[a]	5.90±0.10[a]	5.89±0.09[a]	5.91±0.11[a]	5.91±0.10[a]	5.93±0.13[a]	5.92±0.11[a]	5.89±0.11[a]	5.90±0.08[a]	5.91±0.18[a]	5.97±0.22[a]	5.80±0.02[a]
	Q46002	6.0	6.20±0.39[a]	6.02±0.04[a]	6.01±0.06[a]	6.00±0.06[a]	6.01±0.04[a]	6.02±0.08[a]	6.02±0.05[a]	6.01±0.08[a]	6.01±0.04[a]	5.99±0.13[a]	6.20±0.39[a]	6.04±0.05[a]
	Q8J0K7	6.0	6.17±0.31[a]	6.01±0.04[a]	6.00±0.04[a]	6.01±0.05[a]	5.99±0.03[a]	6.02±0.09[a]	6.00±0.05[a]	6.01±0.08[a]	6.01±0.04[a]	5.98±0.14[a]	6.17±0.31[a]	6.00±0.01[a]
	B5BNY1	6.0	6.19±0.28[a]	6.02±0.05[a]	6.02±0.04[a]	6.03±0.06[a]	6.04±0.05[a]	6.06±0.14[a]	6.03±0.06[a]	6.03±0.09[a]	6.02±0.05[a]	6.01±0.13[a]	6.19±0.28[a]	6.00±0.01[a]
	Q59963	6.0	6.07±0.22[a]	6.02±0.07[a]	6.03±0.06[a]	6.02±0.05[a]	6.01±0.05[a]	6.03±0.09[a]	6.02±0.06[a]	6.02±0.09[a]	6.03±0.06[a]	6.00±0.14[a]	6.06±0.22[a]	6.00±0.01[a]
	Q9RGE6	6.0	5.82±0.22[a]	5.95±0.06[a]	5.97±0.08[a]	5.93±0.09[a]	5.95±0.06[a]	5.94±0.09[a]	5.96±0.06[a]	5.95±0.10[a]	5.96±0.05[a]	5.94±0.14[a]	5.82±0.22[a]	6.01±0.02[a]
	Q6BCL3	6.4	6.08±0.18[a]	6.07±0.03	6.08±0.03	6.06±0.04	6.07±0.03	6.06±0.06	6.07±0.04	6.07±0.07	6.07±0.03	6.05±0.14	6.08±0.18[a]	6.10±0.01

Table 1 (continued)

Predictor	Accession number	Recorded pH	Predicted pH										
			No.	σ_1×No.	$H_M\Delta$PH×No.	σ_R×No.	σ_α×No.	σ_F×No.	A_1×No.	$f(i)$×No.	$f(i+1)$×No.	$f(i+2)$×No.	$f(i+3)$×No.	Amino acid distribution probability
	Q6BCM1	6.4	6.31±0.21[a]	6.34±0.08[a]	6.34±0.08[a]	6.34±0.11[a]	6.32±0.09[a]	6.31±0.14[a]	6.32±0.11[a]	6.36±0.09[a]	6.34±0.08[a]	6.32±0.15[a]	6.31±0.21[a]	6.40±0.01[a]
	Q9AQF4	6.5	6.50±0.36[a]	6.56±0.05[a]	6.56±0.06[a]	6.57±0.06[a]	6.56±0.04[a]	6.57±0.08[a]	6.55±0.05[a]	6.57±0.09[a]	6.56±0.05[a]	6.54±0.14[a]	6.50±0.36[a]	6.50±0.04[a]
	Q9X273	6.6	6.67±0.38[a]	6.61±0.05[a]	6.60±0.06[a]	6.58±0.05[a]	6.59±0.05[a]	6.62±0.08[a]	6.61±0.05[a]	6.60±0.08[a]	6.59±0.05[a]	6.58±0.14[a]	6.67±0.38[a]	6.60±0.01[a]
	P27033	7.0	7.12±0.47[a]	7.07±0.08[a]	7.12±0.14[a]	7.13±0.15[a]	7.04±0.05[a]	7.20±0.22[a]	7.07±0.06[a]	7.08±0.13[a]	7.07±0.09[a]	7.05±0.14[a]	7.12±0.47[a]	7.01±0.04[a]
	O58925 H297N	7.0	6.06±0.17	6.15±0.08	6.24±0.10	6.27±0.10	6.18±0.08	6.25±0.13	6.15±0.09	6.19±0.11	6.18±0.09	6.19±0.16	6.06±0.17	7.15±0.08[a]
	Q3ZMA8	7.5	7.07±0.68[a]	7.51±0.03[a]	7.50±0.04[a]	7.51±0.05[a]	7.51±0.03[a]	7.50±0.04[a]	7.50±0.03[a]	7.51±0.07[a]	7.50±0.04[a]	7.49±0.13[a]	7.07±0.68[a]	7.49±0.03[a]
	B3GQ73	7.5	7.07±0.49[a]	7.52±0.10[a]	7.50±0.09[a]	7.51±0.09[a]	7.55±0.08[a]	7.47±0.17[a]	7.51±0.09[a]	7.54±0.11[a]	7.53±0.09[a]	7.53±0.15[a]	7.07±0.49[a]	7.51±0.02[a]
	P10476	8.0	7.41±0.73[a]	7.95±0.05[a]	7.91±0.12[a]	7.91±0.10[a]	7.97±0.03[a]	7.86±0.16[a]	7.96±0.04[a]	7.93±0.13[a]	7.93±0.08[a]	7.92±0.15[a]	7.41±0.73[a]	8.00±0.04[a]
	O58925 Y299F	8.5	5.94±0.10	6.05±0.07	6.02±0.07	6.05±0.07	6.05±0.06	6.02±0.07	6.03±0.06	6.02±0.09	6.01±0.06	6.01±0.14	5.94±0.10	6.68±0.14
	P06564	9.0	8.50±0.88[a]	8.98±0.08[a]	8.90±0.30[a]	8.95±0.19[a]	8.99±0.03[a]	8.97±0.12[a]	8.99±0.04[a]	9.00±0.09[a]	9.00±0.03[a]	8.99±0.14[a]	8.50±0.88[a]	8.97±0.03[a]
Validation	O58925	5.0	5.89±0.08	5.93±0.03	5.94±0.03	5.95±0.03	5.92±0.03	5.91±0.04	5.93±0.04	5.92±0.07	5.91±0.03	5.91±0.14	5.89±0.08	6.67±0.14
	O58925 C106A/C159A	5.0	5.46±0.21	5.39±0.06	5.39±0.07	5.41±0.07	5.37±0.06	5.36±0.08	5.38±0.06	5.38±0.10	5.38±0.06	5.36±0.16	5.46±0.21	5.78±0.10
	Q83XK5	5.0	5.31±0.81[a]	4.83±0.62[a]	5.23±0.61[a]	5.25±0.61[a]	4.87±0.66[a]	5.46±0.78[a]	4.78±0.59[a]	4.93±0.76[a]	4.95±0.70[a]	4.93±0.79[a]	5.31±0.81[a]	5.77±1.09[a]
	Q75UV6	5.0	5.89±0.77[a]	6.25±0.65[a]	6.33±0.60	6.42±0.60	6.28±0.59	6.37±0.79[a]	6.18±0.74[a]	6.11±0.75[a]	6.14±0.59[a]	6.20±0.67[a]	5.89±0.77[a]	5.28±1.28[a]
	Q9PF60	5.2	5.11±0.35[a]	4.94±0.05	4.95±0.05	4.97±0.06	4.92±0.04	4.95±0.06	4.95±0.05	4.98±0.08	4.97±0.04	4.97±0.13[a]	5.11±0.35[a]	5.00±0.03
	O58925 E342A	5.5	5.83±0.09	5.86±0.04	5.87±0.04	5.86±0.04	5.88±0.03	5.82±0.05	5.87±0.03	5.86±0.08	5.86±0.04	5.85±0.14	5.83±0.09	5.04±0.25[a]
	O58925 N200A	5.5	5.91±0.09	5.94±0.04	5.95±0.04	5.93±0.04	5.93±0.05	5.92±0.05	5.93±0.05	5.93±0.08	5.93±0.04	5.91±0.13	5.91±0.09	6.65±0.21
	O58925 H297A	5.5	6.08±0.17	6.17±0.09	6.24±0.10	6.26±0.10	6.19±0.09	6.27±0.12	6.16±0.08	6.20±0.11	6.19±0.08	6.20±0.16	6.08±0.17	5.63±0.08[a]
	O58925 E342Q	5.5	5.81±0.09	5.84±0.06	5.84±0.05	5.88±0.05	5.84±0.05	5.82±0.06	5.83±0.06	5.82±0.09	5.82±0.05	5.81±0.14	5.81±0.09	6.19±0.19
	Q6BCM1	5.8	6.31±0.21	6.34±0.08	6.34±0.08	6.34±0.11	6.32±0.09	6.31±0.14	6.32±0.11	6.36±0.09	6.34±0.08	6.32±0.15	6.31±0.21	6.40±0.01
	Q5DIE3	6.0	5.34±1.09[a]	4.38±0.78	4.33±0.72	4.27±0.80	4.19±0.74	4.50±0.77[a]	4.34±0.76	4.53±0.75[a]	4.48±0.81[a]	4.38±0.83[a]	5.34±1.09[a]	5.15±1.72[a]
	Q9STD9 Del 1-90	6.0	6.30±0.88[a]	7.03±0.73[a]	6.69±0.76[a]	6.56±0.80[a]	6.95±0.82[a]	6.13±0.85[a]	6.75±0.74[a]	6.79±0.69[a]	6.97±0.64[a]	6.62±0.76[a]	6.30±0.88[a]	6.73±0.70[a]
	Q8J0K8	6.0	6.09±0.53[a]	5.85±0.53[a]	5.70±0.46[a]	5.69±0.57[a]	5.95±0.46[a]	5.79±0.57[a]	5.79±0.45[a]	6.00±0.50[a]	6.00±0.53[a]	5.96±0.54[a]	6.09±0.53[a]	6.10±1.58[a]
	Q6BCL7	6.4	5.97±0.22[a]	5.90±0.10	5.89±0.09	5.91±0.11	5.91±0.10	5.93±0.13	5.92±0.11	5.89±0.11	5.90±0.08	5.91±0.18	5.97±0.22[a]	5.80±0.02
	A8D0T0	6.5	6.18±0.80[a]	5.95±0.84[a]	6.09±0.64[a]	6.14±0.61[a]	5.88±0.71[a]	6.27±0.69[a]	5.99±0.76[a]	5.99±0.67[a]	6.14±0.71[a]	6.14±0.62[a]	6.18±0.80[a]	4.85±1.04[a]
	O65987	7.0	6.71±0.79[a]	7.12±0.64[a]	6.75±0.59[a]	6.60±0.63[a]	7.17±0.60[a]	6.86±0.57[a]	6.93±0.61[a]	7.27±0.62[a]	7.13±0.59[a]	7.08±0.61[a]	6.71±0.79[a]	5.99±0.69[a]

	P27035	7.0	6.17±0.49[a]	6.35±0.58[a]	6.22±0.57[a]	6.29±0.59[a]	6.67±0.48[a]	6.21±0.51[a]	6.17±0.52[a]	6.57±0.51[a]	6.51±0.46[a]	6.56±0.48[a]	6.17±0.49[a]		5.84±0.87[a]
	Q9LAJ2	8.0	6.77±0.78[a]	6.60±0.67	7.01±0.77[a]	6.86±0.67[a]	6.74±0.81[a]	6.81±0.62[a]	6.76±0.66[a]	6.77±0.70[a]	6.67±0.74[a]	6.70±0.75[a]	6.77±0.78[a]		6.63±0.64
	Q8NJY6	8.0	4.07±1.12	3.58±0.75	3.72±0.83	3.88±0.85	3.23±0.67	3.65±0.83	3.51±0.77	3.45±0.75	3.38±0.77	3.34±0.77	4.07±1.12		3.75±1.76
	P18126	8.0	6.84±0.72[a]	6.41±0.82[a]	6.30±0.91[a]	6.53±0.85[a]	6.42±0.73	6.74±0.71[a]	6.68±0.74[a]	6.78±0.75[a]	6.72±0.78[a]	6.55±0.75[a]	6.84±0.72[a]		7.65±1.12[a]
Overall performance			37	31	31	31	29	33	31	33	31	35	37	42	

No. the amino acid composition

[a] No statistical difference with the recorded pH

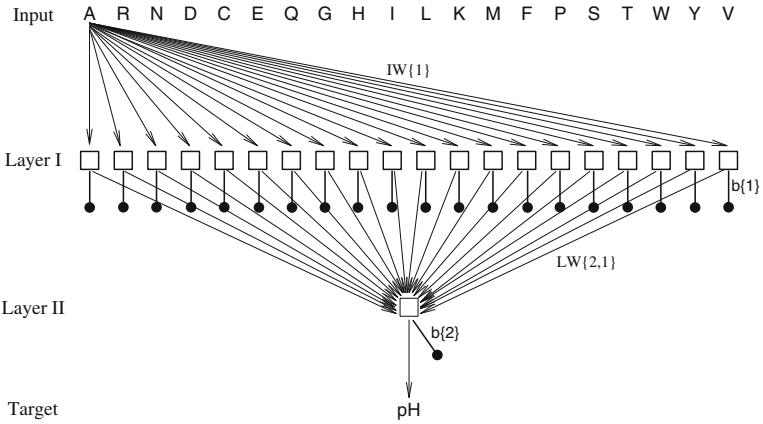

Fig. 1 20–1 feedforward backpropagation neural network to model the relationship between 20 pieces of information on primary structure of cellulase, which are labeled using the symbols of 20 types of amino acids and its pH. Each *square* presents a neuron. IW{1} is the input weights, LW{2,1} is the layer 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

to 20 inputs (or 20 elements of input in neural network terminology), which can be any measure related to 20 types of amino acids. The second layer contains a single neuron corresponding to the single output, pH. The transfer functions are tan-sigmoid and linear for two layers. The training algorithm is the resilient backpropagation, which is the fastest algorithm on pattern recognition in MatLab [17].

Possible Predictors

The following pieces of information, which we consider useful, are listed in Table 2. These pieces of information can be classified according to their size, charge, hydrophilicity or hydrophobicity, and functional groups, which are important indicators for protein structure and protein–protein interactions [18]. Several pieces of information could be particularly considered to be related to primary structure of enzymes such as the spatial properties [19, 20] listed in rows 2–5 in Table 2, hydrophobic properties [21–23] listed in rows 6–10 in Table 2, electronic properties [24] listed in rows 11–17 in Table 2, and the secondary structure predictions [25] listed in rows 18–24 in Table 2.

Actually, all these pieces of information is to use a particular number to replace a certain amino acid in proteins; naturally, each amino acid should have a fixed value using these pieces of information. On the other hand, we have developed a measure, which results in different value for the same type of amino acid [26–30] based on occupancy of subpopulations and partitions [31] according to the following 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 characteristics of this equation is that we view the amino acid position along a cellulase as a distribution; thus, each type of amino acids has its own distribution probability, which is different according to their position in cellulase, their composition, and the length of cellulase as example shown in Table 3.

Preliminary tests indicated that the predictive model could not work if we simply use the values listed in Table 2 to replace each amino acid in cellulases. This is understandable

Table 2 Possible predictors related to amino acids of cellulases

Amino acid	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
Mass, Da	71.09	156.19	115.09	114.11	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	150	160	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	116.1	227.8	193.6	140
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	47	86	135	155	164	124	48	137	39+155	37+199	38+116	43+86	90	56	66	42	69	45	122	89
Residue burial, kcal/mol	1.18	2.15	3.38	3.88	4.1	3.1	1.2	3.43	3.46	4.11	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	0	1	2.2	2.7	2.9	1.9	0	2.3	2.3	2.9	1.6	1.3	1.1	0.2	0.5	-0.1	0.5	0.1	1.9	1.1
Hydropathy index	1.8	4.5	−3.5	−3.5	2.5	−3.5	−3.5	−0.4	−3.2	4.5	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	9	15	16	19	7	17	18	11	10	1	3	20	5	2	13	14	12	6	8	4
pK_a	9.69	9.04	8.8	9.6	10.28	9.67	9.13	9.6	9.17	9.68	9.6	8.95	9.21	9.13	10.6	9.15	9.1	9.39	9.11	9.62
σ_1	0.05	−0.26	−0.14	0.51	−0.01	0.68	−0.1	0	−0.01	0.06	0.02	−0.16	0.08	0.04	0	−0.03	−0.05	0.06	0.05	0.01
$H_M\Delta$PH	0.05	−0.75	−0.2	1.8	−0.01	1.25	−0.07	0	0.21	0.08	0.07	−1.11	−0.04	0.06	0.1	−0.05	−0.03	0.15	0.02	0.09
σ_R	0	−0.49	−0.06	1.29	0.01	0.57	0.03	0	0.22	0.02	0.05	−0.95	−0.12	0.02	0.1	−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	−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	−0.58	0.04	−0.03	0.51	−0.3	−0.45	0.02	−0.38	−0.44	−0.24	−0.42	−0.04
A_1	0.05	0.26	0.24	0.51	0.01	0.68	0.1	0	0.01	0.06	0.02	0.16	0.08	0.04	0	0.03	0.05	0.06	0.05	0.01
P(a)	142	98	67	101	70	151	111	57	100	108	121	114	145	113	57	77	83	108	69	106
P(b)	83	93	89	54	119	37	110	75	87	160	130	74	105	138	55	75	119	137	147	170
P(turn)	66	95	156	146	119	74	98	156	95	47	59	101	60	60	152	143	96	96	114	50
$f(i)$	0.06	0.07	0.161	0.147	0.149	0.056	0.074	0.102	0.14	0.043	0.061	0.055	0.068	0.059	0.102	0.12	0.086	0.077	0.082	0.062
$f(i+1)$	0.076	0.106	0.083	0.11	0.05	0.06	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	0.191	0.179	0.117	0.077	0.037	0.19	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	0.091	0.081	0.128	0.064	0.098	0.152	0.054	0.056	0.07	0.095	0.055	0.065	0.068	0.106	0.079	0.167	0.125	0.053

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

Table 3 Amino acid distribution probabilities for 20 types of amino acids of cellulase A4UU22

Amino acid	Amino acid composition	Amino acid distribution probability
A	40	0.02247
R	8	0.06729
N	22	0.03073
D	23	0.01726
C	4	0.56250
E	9	0.01967
Q	12	0.12410
G	38	0.00961
H	8	0.25234
I	15	0.02616
L	36	0.02173
K	8	0.25234
M	2	0.50000
F	14	0.10307
P	17	0.03658
S	21	0.07952
T	23	0.03956
W	7	0.12852
Y	10	0.11431
V	23	0.07596

because each cellulase has its own amino acid composition: If we simply use the values in Table 2, we would ignore the different amino acid compositions of cellulases. For this reason, we multiply the values listed in Table 2 by their amino acid composition for each cellulase. However, this is not the case for the amino acid distribution probability because their values are different for different cellulases.

Validation of Predictions

Table 1 lists a total of 55 cellulases in databank, of which 35 were used to generate the weights and biases in neural network as training group and 20 were used to validate the neural network with trained weights and biases as validation group. This is a traditional way used in neural network.

Cross-validation was also used, where the data were split into 11 subsets. Each subset had five cases and was held out in turn as the validation set [32, 33].

Finally, the delete-1 observation jackknife was used, leaving out one observation at a time from the sample set [34, 35] because it is most effective in comparison with independent dataset test and subsampling test and is widely used. Each predictor went through this predictive model with same procedures in order to compare its output statistically.

Statistics

For each predictor, 100 trainings were conducted in 20–1 neural network. Then the obtained 100 sets of weights and biases were used to predict the pH optimum 100 times, and their mean and standard deviation were used to compare the recorded pH optimum for each

cellulase [36]. For visual comparison, linear regression was also used to evaluate the predicted pH values with their recorded ones.

Results and Discussion

The neural network in Fig. 1 theoretically can account for various relationships between information of primary structure and pH optimum. Thus, this neural network can guarantee the screening of various predictors, no matter whether the relationship between predictors and pH is linear or nonlinear.

Actually, the development of methods for predictions would include the choice of predictive models and choice of predictors. Naturally, the combination of both choices would be innumerable; therefore, we decided to use a predictive model to screen predictors, and then we can use a predictor to screen various predictive models. On

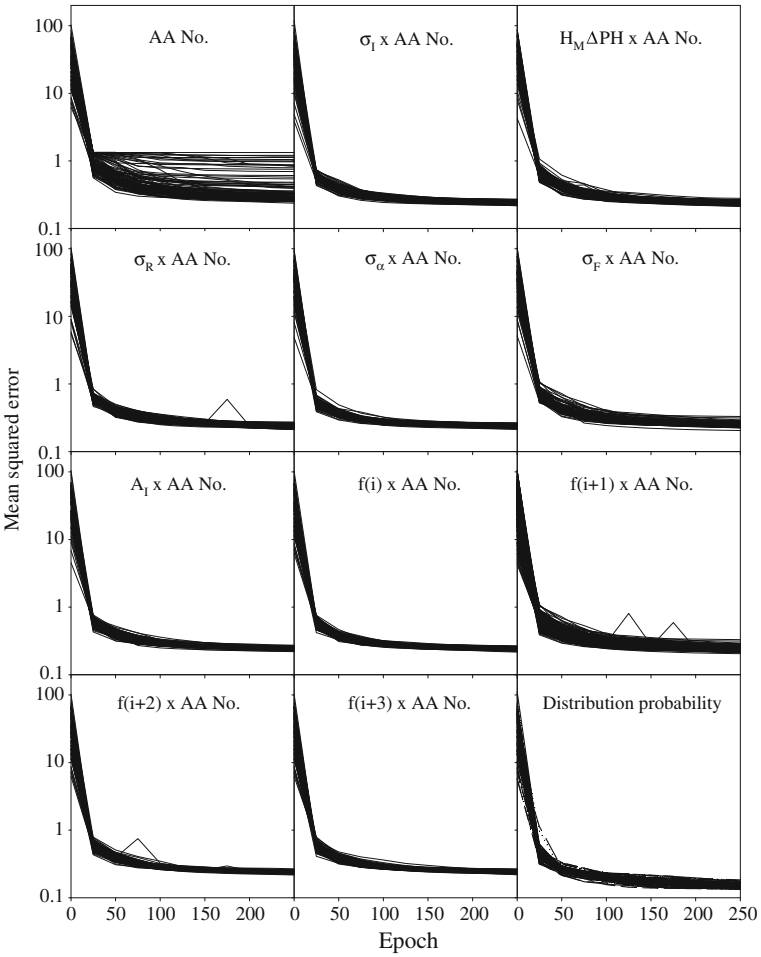

Fig. 2 Convergence of mean squared error performance function with 100 different initial weights and biases generated by random initialization function

the other hand, we also could use the opposite approach, i.e., to use a predictor to screen various predictive models and then use the selected predictive model to screen predictors. However, the latter approach appears less attractive.

As we need to choose a predictive model at first, we decided to use the neural network model because the information listed in Table 2 can be related to every aspect of primary structure. Therefore, it is very hard to define whether the relationship between pH and information from primary structure is a cause-consequence or phenomenological relationship, linear or nonlinear relationship, and explicit or implicit (pattern) relationship, continuous or discrete (in particular example on–off) relationship. Taking this into consideration, we seem to treat the relationship in a black box, which can be accommodated by the feedforward backpropagation neural network to varying extents.

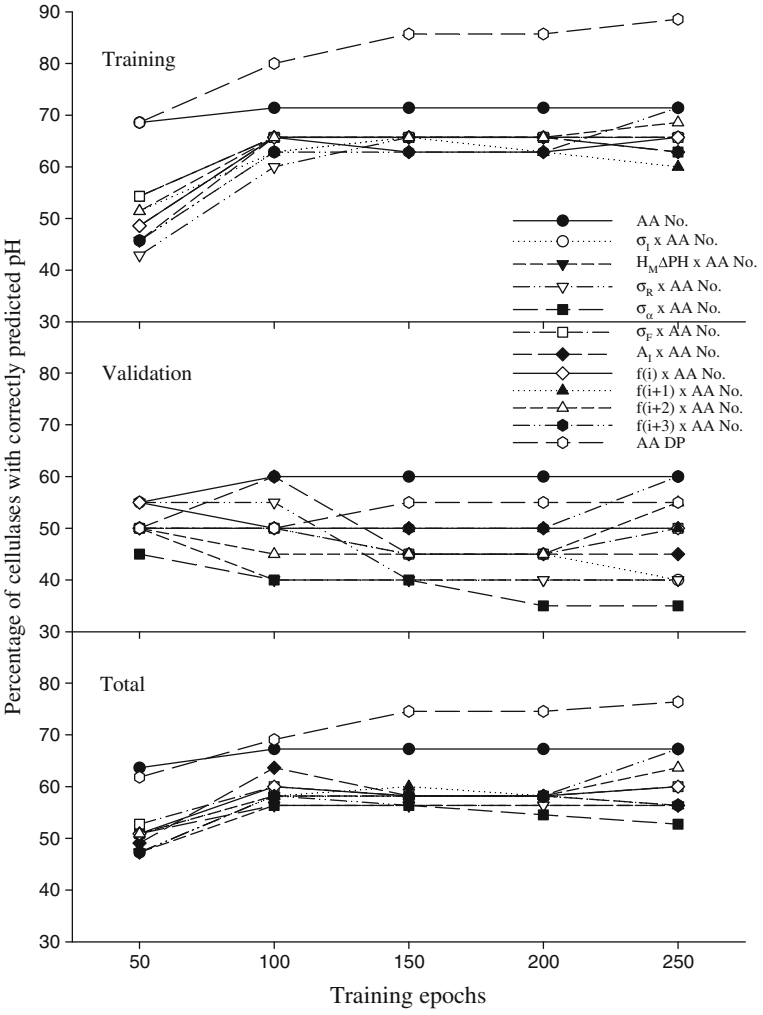

Fig. 3 Percentage of cellulases with correctly predicted pH. The training and validation groups contain 35 and 20 cellulases. *AA No.* amino acid composition, *AA DP* amino acid distribution probability

Furthermore, the second consideration should be directed to training process, which once again guarantees a fair screening of predictors. Technically, 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 250 training epochs for convergence. Figure 2 displays the performance of convergence, where each line represents a training process contains random initialization of weights and biases with 250 training epochs. As seen, the convergence can be reached within 250 training epochs with any random initialization, which lays the foundation to guarantee our training process. However, a detailed scrutiny of each panel in Fig. 2 reveals that different predictors have different profiles of its convergence; for example, the convergence of amino acid composition (top-left panel) is not as narrow as others.

Actually, Fig. 2 demonstrated the training performance, while we can also look at the training process from another viewpoint, i.e., the percentage of correctly predicted pH during the training process. Accordingly, Fig. 3 shows such an analysis: (a) The predictions improve with the increase of training epochs mainly in training group, (b) the efficient epochs for most predictors are about 100, and (c) the amino acid distribution probability works better than other predictors in training group and total group, which is the combination of training and validation groups. So far, Figs. 2 and 3 verified the training process, which laid the base for this study.

Sometimes, the pre-processing of data is needed in order that the input and target data become normally distributed. We also attempted this pre-processing, but the trainings do not converge for all predictors screened, which suggested that the neural network still performs well in some non-normal distribution environment [37].

Table 1 shows the comparison of recorded pH optimum with predicted pH optimum for each cellulase involved in this study. As we would consider a predictor workable if there is no statistical difference between recorded and predicted pH optimum, so the predicted pH optimum is marked with asterisk if no statistical difference was found between recorded and predicted pH optimum. The last row of Table 1 shows the overall performance, where we can see that the amino acid distribution probability works best, followed by amino acid composition and $f(i+3)$, then $f(i+2)$, and so on. Consequently, we used the regression

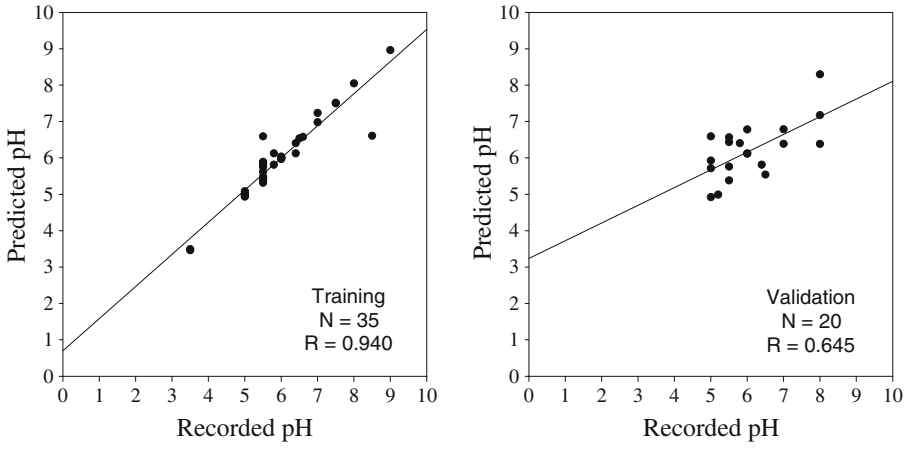

Fig. 4 Linear regression between recorded pH and predicted pH in training and validation groups, respectively, with the amino acid distribution probability as predictor

between recorded pH and predicted pH to visualize the predictive performance using the amino acid distribution probability as predictor (Fig. 4).

The above results are exclusively based on the traditional training and validation [17], while the jackknife validation became popular over recent years; thus, we also used the delete-1 jackknife and 11-fold cross-validation to conduct the training and validation as shown in Fig. 5, where we can see that the best predictor is the amino acid composition.

The structure–function relationship has so far been the objective of many studies [38, 39]; however, the relationship between enzymatic optimal working conditions and information related to primary structure of enzyme is understudied. Our study would take a small step to fill this gap.

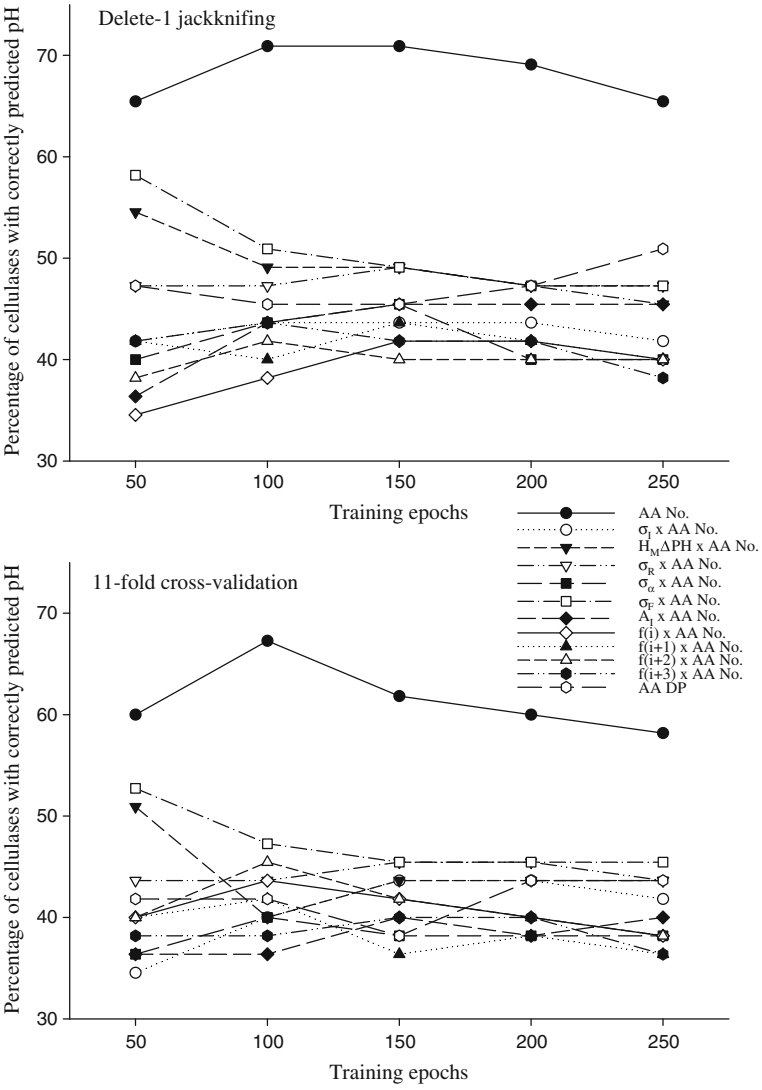

Fig. 5 Percentage of cellulases with correctly predicted pH using the delete-1 jackknifing and 11-fold validation. *AA No.* amino acid composition, *AA DP* amino acid distribution probability

Among different predictors analyzed in this study, the results show that amino acid distribution probability and amino acid composition work better than others, which may be reasonable because both predictors are not constant for each type of amino acids in a cellulase, and this characteristic may be more suitable to reflect the relationship between the enzymatic structure and function.

Actually, the neural network has different structures to accommodate the relationship defined between inputs and target; although the simple the better, we did attempt several more complicated structures such as 20–10–1 or 20–30–15–1, which result in much worse predictions than the results obtained from 20–1 neural network. In Fig. 6, we can see the coefficient of variation (SD/mean×100%) in 20–1 structure is best (upper panel); however, the predictions become more unstable with the increase in complex of model structures, and another evidence can be found in the lower panel. In the future, we may use other predictive models including unstructured models to screen predictors as well as predictive models.

In conclusion, this study demonstrates that some predictors do have a promising prospective to predict the pH optimum of cellulases. Clearly, many more studies are needed in order to explore a cost-effective way to predict various enzymatic parameters in cellulases.

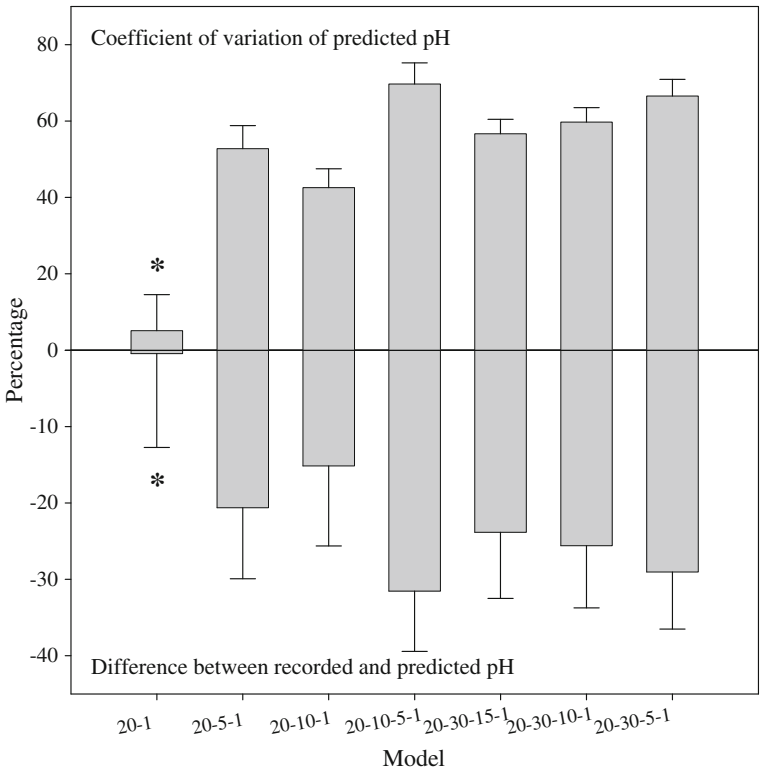

Fig. 6 Coefficient of variation of predicted pH and difference between recorded and predicted pH in different neural network models. The data are presented as mean±SD. There is statistically significant difference among different models ($p<0.001$, one-way ANOVA), and * indicates statistical difference with other models at $p<0.01$ level (Holm–Sidak method)

Acknowledgments This study was partly supported by Guangxi Natural Science Foundation (0832016, 0991013, 0991077, 10-046-06, 11-031-11, 2010GXNSFF013003 and 2010GXNSFA013046). The authors wish to thank the Library of Guangxi Zhuang Autonomous Region for purchasing the book, Biometry.

References

1. Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F., & Schomburg, D. (2002). *Trends in Biochemical Sciences, 27*, 54–56.
2. Pharkya, P., Nikolaev, E. V., & Maranas, C. D. (2003). *Metabolic Engineering, 5*, 71–73.
3. Duan, C. J., & Feng, J. X. (2010). *Biotechnology Letters, 32*(12), 1765–1775.
4. Gonçalves, A. R., Benar, P., Costa, S. M., Ruzene, D. S., Moriya, R. Y., Luz, S. M., et al. (2005). *Appliance Biochem Biotechnol, 121–124*, 821–826.
5. Hahn-Hägerdal, B., Galbe, M., Gorwa-Grauslund, M. F., Lidén, G., & Zacchi, G. (2006). *Trends in Biotechnology, 24*(12), 549–556.
6. Sticklen, M. (2006). *Current Opinion in Biotechnology, 17*, 315–319.
7. Dashtban, M., Schraft, H., & Qin, W. (2009). *International Journal of Biological Sciences, 5*, 578–595.
8. Dhepe, P. L. (2008). *ChemSusChem, 1*, 969–975.
9. Carroll, A., & Somerville, C. (2009). *Annual Review of Plant Biology, 60*, 165–182.
10. Sánchez, C. (2009). *Biotechnology Advances, 27*, 185–194.
11. Zhou, W., Irwin, D. C., Escovar-Kousen, J., & Wilson, D. B. (2004). *Biochemistry, 43*, 9655–9663.
12. Kang, H. J., & Ishikawa, K. J. (2007). *Microbiolog Biotechnol, 17*, 1249–1253.
13. Kim, H. W., Takagi, Y., Hagihara, Y., & Ishikawa, K. (2007). *Bioscience, Biotechnology, and Biochemistry, 71*, 2585–2587.
14. The UniProt Consortium. (2010). *Nucleic Acids Research, 38*(Database issue), D142–D148.
15. Hagan, M. T., Demuth, H. B., & Beale, M. H. (1996). *Neural network design*. Boston: PWS.
16. Demuth, H., & Beale, M. (2001). *Neural network toolbox for use with MatLab. User's guide, version 4*. Natick: The MathWorks.
17. Inc, Math. Works. (2001). *MatLab—the language of technical computing (version 6.1.0.450, release 12.1), 1984–2001*. Natick: The MathWorks.
18. Burlingame, A. L., & Carr, S. A. (1996). *Mass spectrometry in the biological sciences*. Totowa: Humana.
19. Zamyatin, A. A. (1972). *Progress in Biophysics and Molecular Biology, 24*, 107–123.
20. Darby, N. J., & Creighton, T. E. (1993). *Journal of Molecular Biology, 232*, 873–896.
21. Kyte, J., & Doolittle, R. F. (1982). *Journal of Molecular Biology, 157*(1), 105–132.
22. Trinquier, G., Sanejouand, Y. H., & Hausman, R. E. (1998). *Protein Engineering, 11*(3), 153–169.
23. Cooper, G. M. (2004). *The cell: A molecular approach* (p. 51). Washington, D.C.: ASM.
24. Dwyer, D. S. (2005). *BMC Chemical Biology, 5*, 2.
25. Chou, P. Y., & Fasman, G. D. (1978). *Advances in Enzymology and Related Subjects of Biochemistry, 47*, 45–148.
26. Wu, G., & Yan, S. (2008). *Amino Acids, 35*, 365–373.
27. Wu, G., & Yan, S. (2008). *Lecture notes on computational mutation*. New York: Nova Science.
28. Yan, S., & Wu, G. (2009). Biopolymers. *Peptide Science, 92*, 399–404.
29. Yan, S., & Wu, G. (2010). *Annals of Biomedical Engineering, 38*, 984–992.
30. Yan, S., & Wu, G. (2010). *Computer Methods in Biomechanics and Biomedical Engineering, 13*, 403–411.
31. Feller, W. (1968). *An introduction to probability theory and its applications, vol I* (3rd ed.). New York: Wiley.
32. Chou, K. C., & Zhang, C. T. (1995). *Critical Reviews in Biochemistry and Molecular Biology, 30*, 275–349.
33. Zhou, G. P. (1998). *Journal of Protein Chemistry, 17*, 729–738.
34. Chou, K. C., & Shen, H. B. (2007). *Analytical Biochemistry, 370*, 1–16.
35. Chou, K. C., & Shen, H. B. (2010). *Natural Science, 2*, 1090–1103.
36. Sokal, R. R., & Rohlf, F. J. (1995). *Biometry: The principles and practices of statistics in biological research* (3rd ed., pp. 203–218). New York: Freeman.
37. Guh, R. S. (2002). *International Journal Quality Reliab Manag, 19*, 97–112.
38. Pryor, S. W., & Nahar, N. (2010). *Applied Biochemistry and Biotechnology, 162*, 1737–1750.
39. Sudha, T. B., Thanikaivelan, P., Ashokkumar, M., & Chandrasekaran, B. (2011). *Applied Biochemistry and Biotechnology, 163*, 247–257.