The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

Logistic Regression Analysis for Mutation Data of Hemophilia B

Makoto Utsunomiya, Makoto Sakamoto and Hiroshi Furutani

Faculty of Engineering, University of Miyazaki, Miyazaki City, 889-2192

Abstract: Hemophilia B is hereditary disease caused by defects in coagulation factor IX. The compilation of mutation data in the hemophilia B database makes it possible to study this disease at molecular level. The most common mutations reported in the database are amino acid substitutions. We examined the relation between activation level of factor IX and category of amino acid substitution by using the logistic regression analysis. As parameters, we used four physical-chemical parameters of amino acids. We have good results of discrimination for the severity of the disease.

Keywords: hemophilia B, factor IX, amino acid, logistic regression analysis**.**

I. Introduction

Hemophilia is an X-linked hereditary bleeding disorder caused by a deficient or defective coagulation factor VIII or factor IX [1]. About three-fourths of patients with hemophilia have the A type which is due to deficient factor VIII activity. The B type is less frequent than the A type and is due to deficient factor IX activity [2,3].

Factor IX is a vitamin K dependent plasma protein that participates in the middle phase of blood coagulation [4]. The clinical definition of hemophilia B is based on the individual's factor IX activity level as mild (>5%), moderate (1%-5%), or severe (<1%) [5]. Factor IX is made up of seven regions: (1)Signal peptide, (2)Propeptide, (3)Gla, (4)EGF(1st), (5)EGF(2nd), (6)Activation, and (7)Catalytic. Mutation in factor IX is made up of a majority of point mutations. Substitutions of amino acid sequence are the most common form of point mutation. The effects of amino acid substitutions on the activities of factor IX depends on two factors. One is the combination of native and hemophilic amino acids, the differences of physical-chemical properties of the amino acids. Another is the position of mutation in factor IX protein. In general, substitution in important site and substitution to different character from original amino acid are supposed to the drastic decrease in activity of factor IX. On the other hand, variations in unimportant places and substitution to similar type of amino acid are supposed to be lightly affected. We have introduced distances between 20 amino acids by using the following four physical-chemical properties: (1)Molecular volume, (2)Hydropathy, (3)Polar requirement, and (4)Isoelectric point.

There have been reported a variety of defects in the factor IX gene from hemophilia B patients, and these are summarized in the Haemophilia B Mutation Database [6]. There are 2925 patients data in the database. In this study, we analyzed missense mutations in the database described with factor IX activity values. Among them, the cases with more than single mutations and female patients were excluded from our analysis. We adopted 1494 cases.

We performed the logistic regression analysis for the estimation of factor IX activity by using distances of four amino acid parameters. And, we conducted estimation of factor IX activity and prediction of distinction between severely ill patient and mildly ill patient with result of logistic regression analysis.

II. Methods

Table 1 is a sample of the hemophilia B mutation database. CLOTTING is the factor IX activity level. AA_CHANGE indicates the amino acid substitution.

Table 1. Hemophilia B database

CLOTTING	AA_CHANGE	
	Before	After
4.7	V	L
0	V	F
23	V	A

Distance of amino acid.

We adopted four physical-chemical properties of amino acid.

(1) Molecular volume.
(2) Hydropathy
(3) Polar requirement
(4) Isoelectric point

For each amino acid parameter, the distance between amino acid *i* and *j* is defined by the next expression,

$$D_{ij} = \left| f_i - f_j \right|,$$

f_i and f_j are values of each amino acid parameter in amino acid *i* and *j*.

Logistic regression analysis.

In this study, we applied the logistic regression analysis, which a collection of p independent variables denoted by $x_1, x_2, \cdots, x_p$ [7]. This study adopts 4 independent variables $x_k = D_{ij}(k)$ corresponding to four amino acid parameters $(k = 1,2,3,4)$. The logistic regression model is given by

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}},$$

where $x = (x_1, x_2, x_3, x_4)$. The conditional probability given independent variables x is defined by

$$\pi(x) = P(Y = 1 \mid x),$$

where Y is the severity of a patient with $Y = 0$ for mild case and $Y = 1$ for severe case. Thus in this study, $\pi(x)$ represents the probability that a patient with data x shows severe bleeding symptom. The logit of the multiple logistic regression model $g(x)$ is given by the equation

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_4 x_4 .$$

Assume that we have a sample of n independent observations $(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i})$, $i = 1,2,\cdots,n$, where y_i is the severity of the *i*th patient. As in the univariate case, fitting the model requires the estimates of the parameters $(\beta_0, \beta_1, \cdots, \beta_4)$. The method of estimation is the maximum likelihood.

In this study, we defined severity of hemophilia B as follows. Clotting $\geqq$ 5; mild. Clotting < 5; severe. We used sensitivity and specificity. These are statistical measures of the performance of a binary classification test. Table 2 is relationships among terms in this work. Sensitivity is probability that the test recognizes severe patient as severe. Specificity is probability that the test recognizes mild patient as mild.

$$Sensitivity = \frac{TP}{FN + TP}$$

$$Specificity = \frac{TN}{TN + FP}$$

Table 2. Relationships of among terms

		expected severity		
		mild	severe	
actual severity	mild	True Negative	False Positive	Specificity
	severe	False Negative	True Positive	Sensitivity

III. Results

Case 1.

Table 3 is the result of the logistic regression analysis. We used converted for with the equation of y.

$$y = 1 - (clotting/100)$$

In the following we use the notations of the significance level; ** for p-value < 0.01, * for p-value < 0.05 and blank for p-value > 0.05.

Table 3. Result of logistic regression analysis

	coefficient	p-value
β_0	2.1711	**
Molecular volume	0.0063	
Hydropathy	0.1029	
Polar requirement	0.0299	
Isoelectric point	0.1018	
likelihood		-280.67

We defined the expected value of $\pi(x)$ as follows. If $\pi(x) \leq 0.95$; mild case. If $\pi(x) > 0.95$; severe case. And, we performed chi-square test for its result. The result is significance on 1% STD. Table 4 is organized list for these data.

Table 4. Frequency of mild and severe case

		expected severity		
		mild	severe	total
actual severity	mild	315	150	465
	severe	277	752	1029
	total	592	902	1494
p-value		**		

Specificity is 67.7%. Sensitivity is 73.1%.

We defined as follows. If $\pi(x) \leq 0.97$; mild case. If $\pi(x) > 0.97$; severe case. And, we performed chi-square test for its result. The result is significance on 1% STD. Table 5 is organized list for these data.

Table 5. Frequency of mild and severe case

		expected severity		
		mild	severe	total
actual severity	mild	442	23	465
	severe	848	181	1029
	total	1290	204	1494
p-value		**		

Specificity is 95.1%. Sensitivity is 17.6%.

We defined as follows. If $\pi(x) \leq 0.93$; mild case. If $\pi(x) > 0.93$; severe case. And, we performed chi-square test for its result. The result is significance on 1% STD. Table 6 is organized list for these data.

Table 6. Frequency of mild and severe case

		expected severity		
		mild	severe	total
actual severity	mild	110	355	465
	severe	90	939	1029
	total	200	1294	1494
p-value		**		

Specificity is 23.7%. Sensitivity is 91.3%.

Case 2.

Table 7 is the result of logistic regression analysis. We converted clotting by the method. If clotting $\geqq$ 5; $y = 0$. If clotting < 5; $y = 1$.

Table 7. Result of logistic regression analysis

	coefficient	p-value
β_0	-1.044	**
Molecular volume	0.0136	**
Hydropathy	0.1844	**
Polar requirement	0.0958	
Isoelectric point	0.2669	**
likelihood		-743.75

The parameters are significance on 1% STD other than polar requirement.

We defined the expected value of $\pi(x)$ as follows. If $\pi(x) \leq 0.5$; mild case. If $\pi(x) > 0.5$; severe case. We performed chi-square test for its result. The result is significance on 1% STD. Table 8 is organized list for these data.

Table 8. Frequency of mild and severe case

		expected severity		
		mild	severe	total
actual severity	mild	187	278	465
	severe	108	921	1029
	total	295	1199	1494
p-value		**		

Specificity is 40.2%. Sensitivity is 89.5%.

We defined as follows. If $\pi(x) \leq 0.7$; mild case. If $\pi(x) > 0.7$; severe case. And, we performed chi-square test for its result. The result is significance on 1% STD. Table 9 is organized list for these data.

Table 9. Frequency of mild and severe case

		expected severity		
		mild	severe	total
actual severity	mild	339	126	465
	severe	330	699	1029
	total	669	825	1494
p-value		**		

Specificity is 72.9%. Sensitivity is 67.9%.

Case 3.

Table 10 is the result of logistic regression analysis. We converted clotting by the method. If clotting $\geqq$ 5; $y = 0$. If clotting < 5; $y = -\frac{1}{5}(clotting) + 1$.

Table 10. Result of logistic regression analysis

	coefficient	p-value
β_0	-1.2200	**
Molecular volume	0.0123	**
Hydropathy	0.0786	*
Polar requirement	0.0900	*
Isoelectric point	0.1591	**
likelihood		-743.75

Molecular volume and isoelectric point are significance on 1% STD. Hydropathy and polar requirement are significance on 5% STD.

We defined the expected value of $\pi(x)$ as follows. If $\pi(x) \leq 0.5$; mild case. If $\pi(x) > 0.5$; severe case. And, we performed chi-square test for its result. The result is significance on 1% STD. Table 11 is organized list for these data.

Table 11. Frequency of mild and severe case

		expected severity		
		mild	severe	total
actual severity	mild	335	130	465
	severe	316	713	1029
	total	651	843	1494
p-value		**		

Specificity is 72.0%. Sensitivity is 69.3%.

We defined as follows. If $\pi(x) \leq 0.7$; mild case. If $\pi(x) > 0.7$; severe case. And, we performed chi-square test for its result. The result is significance on 1% STD. Table 12 is organized list for these data.

Table 12. Frequency of mild and severe case

		expected severity		
		mild	severe	total
actual severity	mild	440	25	465
	severe	848	181	1029
	total	1288	206	1494
p-value		**		

Specificity is 94.6%. Sensitivity is 17.6%.

IV. Summary

This analysis shows that molecular volume and isoelectric point are important parameters for prediction of hemophilia B severity. In case 1, specificity is 67.7% and sensitivity is 73.1% if we set 0.95 as limits of severity. In case 2, specificity is 72.9% and sensitivity is 67.9% if we set 0.7 as limits of severity. In case 3, specificity is 72.0% and sensitivity is 69.3% if we set 0.5 as limits of severity.

As future work, we would like to develop a method which takes into account the site dependence of mutation.

REFERENCES

[1] Iris Plug, Eveline P. Mauser-Bunschoten, Annette H. J. T. Brocker-Vriends, et al (2006), Bleeding in carriers of hemophilia. BLOOD, 1 JULY 2006 – VOLUME 108, NUMBER 1

[2] Furutani H (1995), A Method to Estimate Effects of Amino Acid Substitutions in Blood Coagulation Factor IX from Hemophilia B Patients. Proceedings of MEDINFO 95, 909

[3] H Furutani (1993), Analysis of Correlation between Amino Acid Substitution and Factor IX Activity in Hemophilia B (in Japanese). Iryoujouhougaku Vol.13 No.4, 211-220

[4] S Yoshitake, Barbara G. Schach, Donald C. Foster, et al (1985), Nucleotide Sequence of the Gene for Human Factor IX. *Biochemistry* 1985, 24, 3736-3750

[5] Da-Yun Jin, Tai-Ping Zhang, Tong Gui, et al (2004), BLOOD, 15 SEPTEMBER 2004・VOLUME 104, NUMBER 6

[6] fixhome : http://www.kcl.ac.uk/ip/petergreen/haemBdatabase.html

[7] David W. Hosmer, Stanley Lemeshow (2000), Applied Logistic Regression Second Edition. WILEY-INTERSCIENCE