Regression Analysis of Amino Acid Substitutions and Factor IX Activity in Hemophilia B

Makoto Utsunomiya, Makoto Sakamoto and Hiroshi Furutani

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

Abstract: Hemophilia B is caused by deficit or decreased activity of factor IX. Mutation in factor IX is made up of a ma jority of amino acid substitution. We examined the relation between activation level of factor IX and category of amino acid substitution with a regression analysis. As parameters, we used four physical-chemical parameters of amino acids and Dayhoff's PAM matrix. In addition, average of activation level with the same a mino acid substitution was used for the analysis. And we analyzed relationship between variation contains cystein e or factor IX's seven regions and activity level.

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

I. Introduction

Hemophilia is a hereditary, X-linked, recessive hemorrhagic disorder[1-2]. 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. Factor IX is a vitamin K dependent plasma protein that participates in the middle phase of blood coagulation. 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. Activity of factor IX in a patient's blood depends on a position of the substitution and combination of original and substituting amino acids. Mutation in factor IX is made up of a majority of point mutation. Substitution of amino acid sequence is the most common form of point mutation. In general, substitution in important site and substitution to different character from original amino acid are supposed to drastic decrease in activity of factor IX. The other way, variation in unimportant place and substitution to similar type of amino acid are supposed to be lightly affected. Cysteine, one of amino acids, has different properties from others. Cysteine contains sulfur, it makes S-S binding with another sulfur[3].

There have been reported a variety of defects in the factor IX gene from hemophilia B patients, and these are summarized in the hemophilia B database[4-5]. 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 analysis, excluding 1431 cases. We adopted 1494 cases. 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. We also adopted two homology matrices. These matrices are symmetric and composed of 20 x 20 elements, corresponding to amino acid pairs. (1)Dayhoff's 120PAM matrix, and (2)Dayhoff's 250PAM matrix. We performed simple and multiple liner regression analysis for the estimation of factor IX activity by using four amino acid parameters and obtained a distance matrix. In addition, we searched relationship between variation contains cysteine or factor IX's seven regions and activity level.

II. Methods

Distance of amino acid. About each four amino acid parameters, the distance between amino acid *i* and *j* is defined by the next expression.

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

PAM matrix. PAM is permutation matrix which Dayhoff(1978) obtained molecular evolution-wise, and evolutionary measure of time that single mutation per 100 amino acids occurs in amino acid sequence[6]. The PAM score is calculated as follows.

The number of which amino acid *i* is substituted for amino acid *j* during 1PAM is m_{ij}, appearance probability of amino acid *i* is f_i. The number of mutation in amino acid *i* is $m_i = \sum_{i \neq j} m_{ij}$.

Total number of mutating amino acid is $m = \sum_i m_i$.

Probability that amino acid *i* mutates is $M_i = \frac{m_i}{100 m f_i}$.

And, probability that amino acid *i* changes to *j* is $M_{ij} = \frac{m_{ij}}{m_i} M_i$, probability that amino acid *i* not changes is $M_{ii} = 1 - M_i$.

Transition probability matrix is $M = \lfloor M_{ij} \rfloor$, $\sum_j M_{ij} = 1.$

Transition probability matrix of *k*PAM is M^k .

Score of *k*PAM is $\frac{f_i M_{ij}^k}{f_i f_j} = \frac{M_{ij}^k}{f_j}$, and therefore elements of PAM matrix are $10 \log \left[\frac{M_{ij}^k}{f_j} \right]$.

Regression analysis. Technique to analyze the relations between two or more parameters. Assume $y = f(x_1, x_2, \cdots, x_p) + \varepsilon$.

Correlation analysis. Technique for analyzing related strength between some variable quantities. Coefficient of correlation *r* is used as criterion of strength of the relation between variable *x* and *y*. *r* is defined by the next expression.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Simple liner regression analysis.

$y_\alpha = \eta_\alpha + \varepsilon_\alpha = \beta_0 + \beta_1 x_\alpha + \varepsilon_\alpha$.

This model is called simple liner regression model. Estimate value of unknown constant number β_0, β_1 are $\hat{\beta}_0, \hat{\beta}_1$. We use least-square method for obtaining $\hat{\beta}_0, \hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x} .$$

Multiple liner regression analysis.

$$y_\alpha = \eta_0 + \varepsilon_\alpha$$
$$= \beta_0 + \beta_1 x_{\alpha 1} + \cdots + \beta_i x_{\alpha i} + \cdots + \beta_p x_{\alpha p} + \varepsilon_\alpha$$

.

This model is called multiple liner regression model. As in the case with simple liner regression, we find the solution with least-square method.

III. Results

Table 1 is the result of simple liner regression analysis between clotting and molecular volume. If p-value < 0.05 is *. If p-value < 0.01 is **.

Table 1. Correlative relationship

	coefficient	p-value
Molecular volume	−0.04070	**
coefficient of correlation	0.1579	

This is significance on 1% STD.

Table 2 is the result of simple liner regression analysis between clotting and hydropathy.

Table 2. Correlative relationship

	coefficient	p-value
Hydropathy	−0.6107	**
coefficient of correlation	0.1768	

This is significance on 1% STD.

Table 3 is the result of simple liner regression analysis between clotting and polar requirement.

Table 3. Correlative relationship

	coefficient	p-value
Polar requirement	−0.6656	**
coefficient of correlation	0.1599	

This is significance on 1% STD.

Table 4 is the result of simple liner regression analysis between clotting and isoelectric point.

Table 4. Correlative relationship

	coefficient	p-value
Isoelectric point	−0.5053	**
coefficient of correlation	0.1606	

This is significance on 1% STD.

Table 5 is the result of multiple liner regression analysis between clotting and four physical-chemical properties.

Table 5. Correlative relationship

	coefficient	p-value
Molecular volume	-0.02831	*
Hydropathy	-0.3685	**
Polar requirement	-0.1455	
Isoelectric point	-0.4155	**
coefficient of correlation	0.2510	
significant		**

This is significance on 1% or 5% STD other than polar requirement.

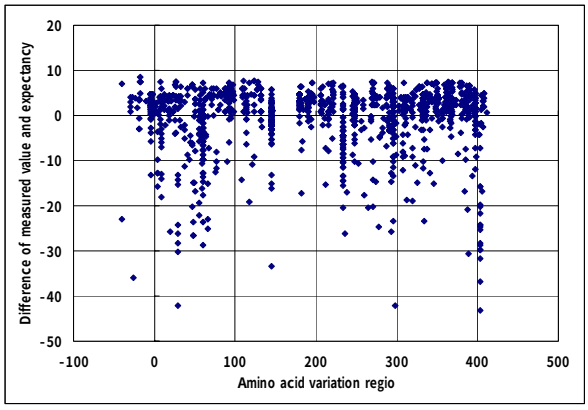

Fig 1. Relations of mutation site and difference

The x-axis is site of origin of amino acid variation. The y-axis is difference of measured value and expectancy of clotting with result of Table 5. The difference at end of factor IX is larger than the central portion.

Table 6 is the result of simple liner regression analysis between clotting and 120PAM.

Table 6. Correlative relationship

	coefficient	p-value
PAM120	0.6207	**
coefficient of correlation	0.1939	

This is significance on 1% STD.

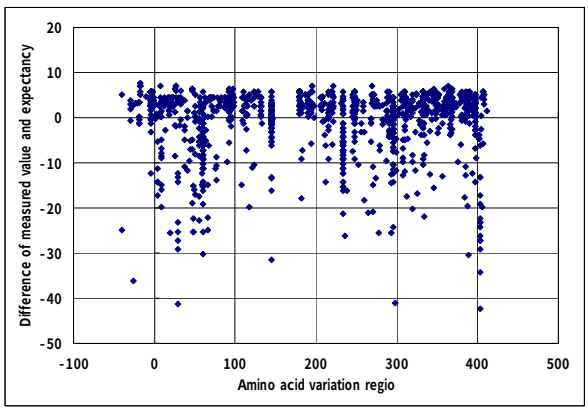

Fig 2. Relations of mutation site and difference

The x-axis is site of origin of amino acid variation. The y-axis is difference of measured value and expectancy of clotting with result of Table 6. The difference at end of factor IX is larger than the central portion.

Table 7 is the result of simple liner regression analysis between clotting and 250PAM.

Table 7. Correlative relationship

	coefficient	p-value
PAM250	0.5370	**
coefficient of correlation	0.1610	

This is significance on 1% STD.

We divided data into variation contains cysteine and variation not contains cysteine. Then, we compared expectancy with measured value of clotting in each data. As a result, the ratio that clotting lowered than expectancy was high in variation contains cysteine.

We divided data into boundary vicinity of factor IX's seven regions and other data. Then, we compared expectancy with measured value of clotting in each data. As a result, the ratio that clotting lowered than expectancy was high in boundary vicinity of signal peptide and propeptide.

We calculated average of clotting in each amino acid substitution that after variation is the same as before variation. Type1: we replaced calculated clotting with original clotting. Type2: if before and after of amino acid substitution are same, we made one data from those data. Therefore, the number of data decreased to 126.

Table 8. Sample data

CLOTTING	AA_CHANGE	AA_CHANGE - b	AA_CHANGE - a
1	2	N	I
5	3	S	P
23	4	G	S
20	4	G	S
20	4	G	S
17	4	G	S
11	4	G	S
4	5	K	E

In this instance, the clotting changes in 18.2 that data of before variation is G and after variation is S.

Table 9 is the result of multiple liner regression analysis between average of clotting (type1) and four physical-chemical properties.

Table 9. Correlative relationship

	coefficient	p-value
Molecular volume	-0.02831	**
Hydropathy	-0.3685	**
Polar requirement	-0.1455	*
Isoelectric point	-0.4155	**
coefficient of correlation	0.4970	
significant		**

This is significance on 1% STD other than polar requirement.

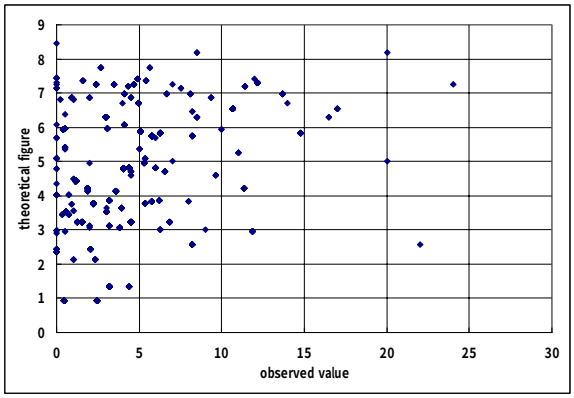

Fig 3. Observed value and theoretical figure

The x-axis is observed value that average with type1. The y-axis is theoretical figure of clotting with result of Table 9.

Table 10 is the result of multiple liner regression analysis between average of clotting (type2) and four physical-chemical properties.

Table 10. Correlative relationship

	coefficient	p-value
Molecular volume	0.0005939	
Hydropathy	-0.5087	*
Polar requirement	-0.006099	
Isoelectric point	-0.4324	
coefficient of correlation	0.3061	
significant		*

Altogether and hydropathy are significance on 5% STD.

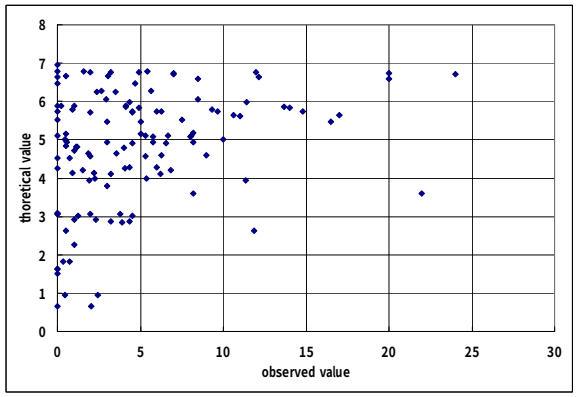

Fig 4. Observed value and theoretical figure

The x-axis is observed value that average with type2. The y-axis is theoretical figure of clotting with result of Table 10.

IV. Summary

This analysis shows that the highest correlation with factor Ⅸ activity is given by Dayhoff's 120PAM matrix in the simple liner regression analysis. The multiple liner regression analysis, using four physical-chemical properties of amino acids, gives higher correlation coefficient. We found that variation containing cysteine and mutation in a particular sphere of seven regions are more likely to have severe disease.

The result of multiple liner regression analysis using average of same pair of before amino acid substitution and after the substitution, we got significant result and high correlation coefficient in the analysis with type1.

REFERENCES

[1] Furutani H (1995), A Method to Estimate Effect s of Amino Acid Substitutions in Blood Coagulatio n Factor IX from Hemophilia B Patients. Proceedin gs of MEDINFO 95, 909

[2] 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

[3] 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

[4] F.Giannelli, P.M.Green, K.A.High, et al (1991), Haemophilia B: database of point mutations and short additions – second edition. *Nucleic Acids Research*, Vol.19, Supplement, 2193-2196

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

[6] BRUCE S. WEIR (1996), Genetic Data Analysis II. *Sinauer Associates Inc*.