Journal of Robotics, Networking and Artificial Life, Vol. 2, No. 2 (September 2015), 115-119

# Analysis of Genetic Disease Haemophilia A by Using Machine Learning

**Kenji Aoki**
*Information Technology Center, University of Miyazaki*
*Gakuen-kibanadai-nishi 1-1, Miyazaki City, Miyazaki, 889-2192, Japan*

**Makoto Sakamoto and Hiroshi Furutani**
*Faculty of Engineering, University of Miyazaki*
*Gakuen-kibanadai-nishi 1-1, Miyazaki City, Miyazaki, 889-2192, Japan*
*E-mail: aoki@cc.miyazaki-u.ac.jp, sakamoto@cs.miyazaki-u.ac.jp, furutani@cs.miyazaki-u.ac.jp*
*www.miyazaki-u.ac.jp*

**Abstract**

Haemophilia A is a genetic disease resulting from deficiency of factor VIII. The database of mutations causing haemophilia A has been developed by the world wide collaboration. In this study, we examined the relation between activity of factor VIII and the missense mutation by using machine learning. As parameters, we used four physical-chemical parameters of amino acids. We predicted the severity of haemophilia A by using machine learning in factor VIII. As the result, logistic regression is not better than other methods in the prediction of haemophilia A severity. The result of the prediction improved in order to SVM, bagging, boosting and random forest. These results suggested that we can predict the haemophilia A severity by using these methods, and random forest was the best method in these five methods to predict the haemophilia A severity.

*Keywords*: Haemophilia A, Machine Learning, Factor VIII, Amino-acid, Mutation

## 1. Introduction

The haemophilia is a group of hereditary genetic disorders, in which one of the coagulation factors is deficient [1]. Haemophilia A is the most common form of disorder caused by low concentration of the coagulation factor VIII. Haemophilia B is another form of disorder caused by deficient factor IX. Haemophilia A accounts for about 85% of this disorder, while haemophilia B for 10−12% [2].

Haemophilia A and B are clinically indistinguishable from each other. Diagnosis must be confirmed by specific factor assay. It becomes very important to study mutations in genes responsible for diseases by biological experiment. However, it is a time-consuming, laborious and expensive task. Thus, it is necessary to develop computational method by applying various approaches. We used a multiple regression model to predict the effect of a missense mutation in factor IX gene of haemophilia B patients [3]. In the past, we have demonstrated the calculations using Support Vector Machin (SVM) for the analysis of mutant factor VIII genes [4].

There have been reported a variety of defects in factor VIII gene from haemophilia A patients [5], and these are summarized in the haemophilia A database [6]. This database has data of clotting activity, nucleotide No., position, changed amino-acid and mutation type. In this study, we analyzed amino acid changing mutations, or missense mutations in the database described with factor VIII activity values. We adopted 439 cases from the database. We use the distances between 20 amino acids by using the four

Published by Atlantis Press

physical-chemical properties: Molecular volume, Hydropathy, Polar requirement and Isoelectric point. These distances are the differences between physical-chemical values before mutation and after mutation. In this study, we used some machine learnings to analyze of haemophilia A severity, and we compared these methods.

## 2. Methods

### 2.1. *Haemophilia A Database*

The gene coding for human factor VIII consists of 26 exons and 25 introns, and is located on the X chromosome [5]. Factor VIII is an essential blood-clotting protein, and synthesized as a precursor protein of 2351 amino acids. This includes a signal peptide and a mature protein of 2332 amino acids with domain structure A1-A2-B-A3-C1-C2. Classification of haemophilia A is presented in Table 1.Three A domains display approximately 30% homology to each other. The C domains are structurally related to the C domains of factor V. The B domain exhibits no significant homology with any other known protein. We used Haemophilia A Mutation Database [6]. The part of the database is shown in Table 2. This database includes exon number, amino-acid number, amino-acid change and activity of factor VIII (FVIII:C). Activity of factor VIII in a patient's blood depends on a position of the substitution and combination of original and substituting amino acids.

Table 1. Domain structure and number of data in Factor VIII.

| Domain | Location | Number of data |
|---|---|---|
| A1 | 1 ~ 329 | 111 |
| A2 | 330 ~ 711 | 131 |
| B | 712 ~ 1648 | 18 |
| A3 | 1649 ~ 2019 | 107 |
| C1 | 2020 ~ 2172 | 39 |
| C2 | 2173 ~ 2332 | 33 |
| | total | 439 |

Table 2. Mutation database of haemophilia A.

| Exon Number | Amino-acid Number | Amino-acid Change | FVIII:C (%) |
|---|---|---|---|
| 1 | 3 | Arg Thr | 1 |
| 1 | 6 | Tyr Cys | 6 |
| 1 | 10 | Val Gly | <1 |
| 1 | 11 | Glu Lys | 1.5 |
| 1 | 14 | Trp Gly | 5 |
| ⋮ | ⋮ | ⋮ | ⋮ |

### 2.2. *Machine Learning*

We used five machine learnings for analysis of haemophilia A database. These are logistic regression, support vector machine, bagging, boosting and random forest. We used statistical application software "*R*" and packages for calculations. The packages are 'kernlab', 'ipred', 'ada' and 'randomForest'.

#### 2.2.1. *Logistic Regression*

Logistic regression is a probabilistic statistical classification (regression) model. It is used for predicting the outcome of a categorical dependent variable based on predictor variables. It is a kind of generalization linear model using a logistic function.

#### 2.2.2. *Support Vector Machine*

SVM (Support Vector Machine) is supervised learning models with associated learning algorithms [7]. It is used for classification and regression analysis. Given a set of training data, SVM builds a model. It assigns new data into one category or the other. It is a non-probabilistic binary linear classifier.

#### 2.2.3. *Bagging*

Bagging is a method for generating multiple versions of a predictor and using these to get an aggregated predictor [8]. The aggregation does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set. Tests on data sets using classification and regression trees show that bagging can give substantial gains in accuracy.

#### 2.2.4. *Boosting*

Boosting is a machine learning based on the idea of creating a highly accurate predictor by combining many

weak rules of thumb [9]. A remarkably rich theory has evolved around boosting with connections to a range of several topics. Boosting algorithms have also made practical success in such fields as biology, vision, and speech processing.

#### 2.2.5. *Random Forest*

Random forests are an ensemble learning method for classification and regression [10]. It operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The method combines bagging idea and the random selection of features, introduced in order to construct a collection of decision trees with controlled variance.

### 2.3. *Dataset*

We used data of haemophilia A Mutation Database. The number of data in A1, A2, B, A3, C1 and C2 domain are 111, 131, 18, 107, 39 and 33 respectively. We used all data in each domain for training data and test data in machine learning. We considered serious illness with less than 1% of factor VIII activity, and slight illness with more than 1% of one. We predicted the serious or slight illness of haemophilia A by machine learning based on these data.

We used a distance between amino acids for each four amino acid physical-chemical parameters (Molecular volume, Hydropathy, Polar requirement and Isoelectric point). The k-th distance between amino acid $A_i$ and $A_j$ is defined as

$$D_{ij}^{(k)} = \left| f_k(A_i) - f_k(A_j) \right| \quad (1)$$

In this study, $A_i$ is a normal amino acid, and $A_j$ is a substituting amino acid.

## 3. Results

At first, we predicted the severity of haemophilia A by logistic regression, SVM, bagging, boosting (adaboost) and random forest using all domain data. Furthermore, we predicted severity of haemophilia A using each domain data.

In all figures, the horizontal axis is the false positive ratio, and the vertical axis is the true positive ratio. False positive means that the predicted result is positive (serious), but observed result is negative (slight). True positive means that both the prediction result and observed result are positive. False negative and true negative are similar to these. We plot the relationship between the false positive ratio and the true positive ratio in the figures. Such a figure is called ROC curve. ROC curve is used for a comparison of the inspection performance, which in the upper left indicates more superior performance.

The result of prediction using all data is shown in Fig. 1. In Fig. 1, there are five curve these are the result of prediction by each machine learning method. The result of using logistic regression lies in more down right than other methods in the prediction of haemophilia A severity. The result of using random forest situated in the most upper left than other curves. Therefore, this result suggested that the predictions by random forest is the most superior performance in five machine learnings.

The result of prediction using one domain data is shown in Fig. 2, 3 and 4. Fig. 2, Fig. 3, Fig. 4 show the ROC curve of prediction using A1, A2, and A3 domain data, respectively. In all figures, the ROC curve using random forest method lies in upper left than other curves. This result is more remarkable in Fig. 2, 3 and 4 than Fig. 1. This result suggested that the predictions by random forest is also the most superior performance in five machine learnings using each domain data. We were not able to get a clear result in other domains, because number of data is too small in these domains.

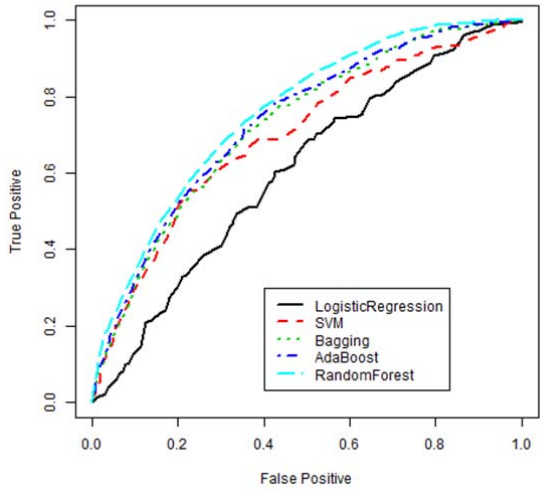


Fig. 1. ROC curve of prediction of haemophilia A severity using all domain data.

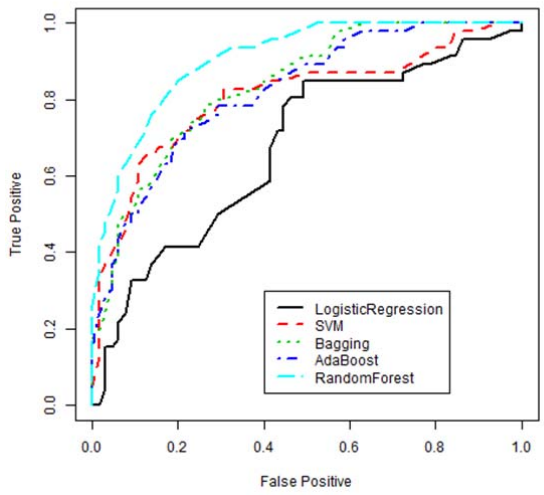


Fig. 2. ROC curve of prediction of haemophilia A severity using A1 domain data.

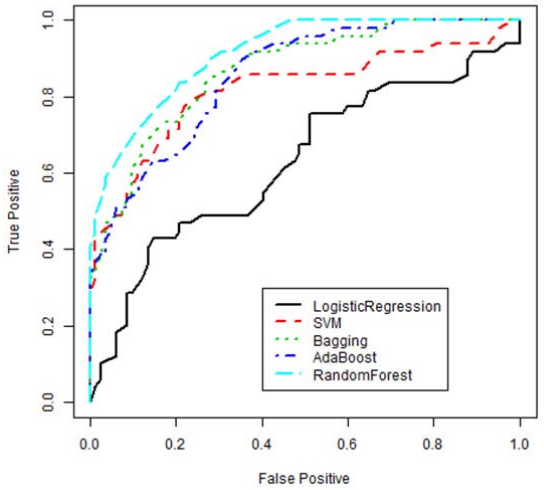


Fig. 3. ROC curve of prediction of haemophilia A severity using A2 domain data.

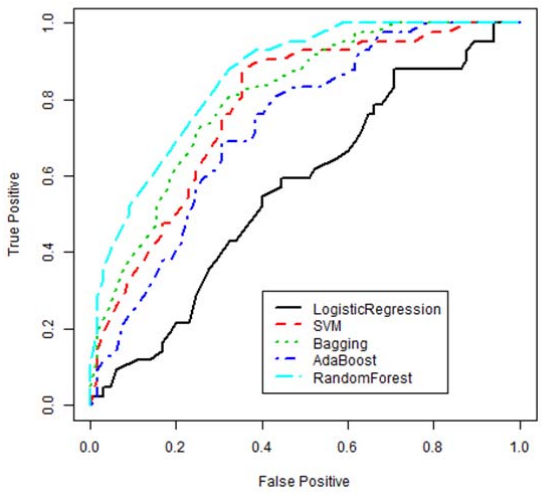


Fig. 4. ROC curve of prediction of haemophilia A severity using A3 domain data.

## References

1. P.H.B. Bolton-Maggs, K.J. Pasi, Haemophilias A and B, *The Lancet* **361** (2003) 1801-1809
2. B. Furie, B.C. Furie, The Molecular Basis of Blood Coagulation, *Cell* **53** (1988) 505-518

3. M. Utsunomiya, M. Sakamoto, H. Furutani, Regression Analysis of Amino Acid Substitutions and Factor IX Activity in Hemophilia B, *Artificial Life and Robotics* **13** (2008) 531-534
4. K. Aoki, K. Yamamori, M. Sakamoto, H. Furutani, SVM Analysis of Hemophilia A by Using Protein Structure, *Lecture Notes in Computer Science, Springer* **8227** (2013) 681-688
5. J. Gitschier, W.I. Wood, T.M. Goralka, et al. Characterization of the Human Factor VIII Gene. *Nature* **312** (1984) 326-330
6. G. Kemball-Cook, The Haemophilia A Mutation, Structure, Test and Resource Site (HAMSTeRS), http://hadb.org.uk/ (11 February 2013 updated)
7. B. Schölkopf, C.J.C. Burges, A.J. Smola, Advances in Kernel Methods, The MIT Press, London (1999)
8. L. Breiman, Bagging Predictors, *Machine Learning* **24**(2) (1996) 123-140
9. J. Friedman, T. Hastie, R. Tibshirani, Additive Logistic Regression: A statistical view of boosting, *Annals of Statistics* **28**(2) (2000) 337-374
10. L. Breiman, Random Forest, *Machine Learning* **45**(1) (2001) 5-32