The Eighteenth International Symposium on Artificial Life and Robotics 2013 (AROB 18th '13),
Daejeon Convention Center, Daejeon, Korea, January 30-February 1, 2013

Study of factor IX gene using regional structure

Hiroshi Furutani[1], Kenji Sueyoshi[2], Kenji Aoki[3]
Kunihito Yamamori[1], and Makoto Sakamoto[1]

[1] Faculty of Engineering, University of Miyazaki, Japan
[2] Graduate School of Engineering, University of Miyazaki, Japan
[3] Information Technology Center, University of Miyazaki, Japan

e-mail: furutani@cs.miyazaki-u.ac.jp

Abstract: There have been reported a variety of defects in the factor IX gene, which is responsible for hemophilia B, and these are summarized in the hemophilia B database. We analyzed amino acid changing mutations, or missense mutations in the database described with factor IX activity values. We have carried out several kinds of theoretical studies to predict the effect of a missense mutations in Factor IX gene. In this paper, we report results of the analysis using Support Vector Machine. We applied the method of transfer learning, which uses the knowledge of some domain to predict the properties of other domains. As a training set, we use mutations of one of seven regions, and test the obtained parameters by the prediction of mutations in remaining regions.

Keywords: hemophilia, missense mutation , factor IX, SVM

1 INTRODUCTION

As an advance of genome science, there appear many reports describing the analysis of mutations in genes responsible for diseases. Foe example, the Human Gene Mutation Database includes mutations causing or associated with human inherited disease [1]. However, even in recent days, it has been difficult to characterize a genetic disorder, such as hemophilia, in which affected individuals have various types of mutations. It is a time consuming, laborious and expensive task to distinguish a causative mutation from neutral ones. Therefore, it becomes very important to study mutations in genes responsible for diseases by using computer.

We have applied a multiple regression model to predict the effect of a missense mutation in Factor IX gene of hemoplilia B patient [2]. As an extension of this work, we also carried out the same type of analysis using Support Vector Machine [3].

In this study, we apply the method of transfer learning, which uses the knowledge of some domain to predict the properties of other domains.

Genetic defects in blood coagulation proteins are associated with a bleeding disorder known as hemophilia. There are two types of hemophilia, Hemophilia A and Hemophilia B. Hemophilia B is a hereditary, X-linked, recessive hemorrhagic disorder, caused by various types of mutations in factor IX gene. Factor IX (or Christmas factor) is one of the serine proteases of the coagulation system [4, 5]. Mutation in factor IX is made up of a majority of point mutations. Substitution of amino acid sequence is the most common form of point mutations. 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. 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 [6].

We analyzed amino acid changing mutations, or missense mutations in the database described with factor IX activity values. We have introduced distances between 20 amino acids by using the following four physical-chemical properties: molecular volume, hydropathy, polar requirement, and isoelectric point. We carried out an analysis of missense mutations in the database by using SVM. We apply the theory of transfer learning for the analysis of these setting.

2 HEMOPHILIA B DATABASE

We used Hemophila B Mutation Database in this analysis [6]. The latest version of the database is the 13th edition.

There are 2,891 entries in this database, and 34 double mutations and 1 triple mutation are included. Of the 2,891 patients listed, 962 show unique molecular events probably causing the disease. Most of them are point mutations, and the database contains 561 different amino acid substitutions. One hundred and forty-eight residues of factor IX show two or more amino acid substitutions and 99 only one. Among them, the cases of double and triple mutations and female patients were excluded. We adopted 1494 cases as total.

The gene for factor IX contains eight exons and seven introns with an overall size of 33k base pairs[5]. Factor IX is a glycoprotein of 461 amino acids essential for blood coagulation, and made up of seven regions:

Table 1. Used data in database

Domain	Location	Number of mutants
Signal peptide	-46 to -18	18
Propeptide	-17 to -1	107
Gla	1 to 46	99
EGF(1st)	47 to 84	138
EGF(2nd)	85 to 127	95
Activation	128 to 195	241
Catalytic	196 to 415	795

(a) Signal peptide,

(b) Propeptide,

(c) Gla,

(d) EGF(1st),

(e) EGF(2nd),

(f) Activation,

(g) Catalytic.

Signal peptide and propeptide are removed during biosynthesis, and remaining protein circulates in the blood as a mature factor IX.

Activity of factor IX in a patient's blood depends on a position of the substitution and combination of original and substituting amino acids. Classification of the disease is given by in vitro clotting activity,

(i) Severe hemophilia B: <1 % factor IX

(ii) Moderate hemophilia B: 1 % – 5 % factor IX

(iii) Mild hemophilia B: > 5 % factor IX.

3 METHOD

We define the distance of amino acid. For each amino acid parameter, the distance D_{ij} between amino acid i and j is defined by the next expression,

$$D_{ij} = |f_i - f_j|.$$

Here, f_i and f_j are one of four amino-acid parameter of i and j, respectively.

3.1 Support Vector Machine

Support vector machine (SVM)[7][8] can classify the samples $\boldsymbol{x}_i$ $(i = 1, \cdots, n)$ belonging to unknown class into two classes C_1 or C_2. The classification function $f(\boldsymbol{x})$ is defined as the Equation (1).

$$f(\boldsymbol{x}) = sign\left(g(\boldsymbol{x})\right) = sign(\boldsymbol{w}^t\boldsymbol{x} + b), \quad (1)$$

where $\boldsymbol{w}$ and b are parameters.

Let $\boldsymbol{x}_i$ belong to the class y_i $(= \{1, -1\})$, and if all the samples are correctly classified, Equation (2) will be satisfied.

$$\forall i,\ y_i \cdot \left(\boldsymbol{w}^t\boldsymbol{x}_i + b\right) - 1 \geq 0. \quad (2)$$

When Equation (2) is satisfied, no samples exist between the $H_1 : (\boldsymbol{w}^t\boldsymbol{x} + b) = 1$ and the $H_2 : (\boldsymbol{w}^t\boldsymbol{x} + b) = -1$, and the distance between H_1 and H_2, called as *margin*, becomes $\frac{2}{||\boldsymbol{w}||}$. To obtain the maximum margin, we minimize $\frac{1}{2}||\boldsymbol{w}||^2$. In SVM, it is solved by a Lagrange-multiplier method. To maximize the margin, we rewrite the objective function as Equation (3) in subject to Equation (2),

$$L\left(\boldsymbol{w}, b, \boldsymbol{\alpha}\right) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{n} \alpha_i \left[y_i \left(\boldsymbol{w}^t\boldsymbol{x} + b\right) - 1\right]. \quad (3)$$

where $\alpha \geq 0$ denotes Lagrange-multiplier. Partial differentiations of Equation (3) by $\boldsymbol{w}$ and b are substituted for Equation (3), we obtain Equation (4).

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^t\boldsymbol{x}_j, \quad (4)$$

in subject to

$$\forall_i,\ \alpha_j \geq 0,\ \sum_{i=1}^{n} \alpha_i y_i = 0. \quad (5)$$

Here we denote α_i to maximize Equation (4) as α_i^*. The sample $\boldsymbol{x}_i$ with $\alpha_i^* > 0$ is called as Support Vector (SV), it exists on H_1 or H_2. The optimum of $\boldsymbol{w}$ denoted as $\boldsymbol{w}^*$ is obtained from the partial differentiations of Equation (3) and α_i^* by Equation (6).

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha^* y_i \boldsymbol{x}_i. \quad (6)$$

The optimum of b denoted as b^* is obtained from the Equation (7) with any $\boldsymbol{x}_s$ $(s \in SV)$.

$$b^* = y_s - \boldsymbol{w}^{*t}\boldsymbol{x}_s. \quad (7)$$

Finally, we obtain the discriminant function of SVM for linearly separable problem as Equation (8).

$$f(\boldsymbol{x}) = sign\left(\sum_{i \in SV} \alpha_i^* y_i \boldsymbol{x}_i^t\boldsymbol{x} + b^*\right). \quad (8)$$

4 NUMERICAL EXPERIMENTS

In this discrimnant analysis, we adoped the software package SVM-Light developed by Joachims [9]. Patients' data are devided into two classes, (i) severe and moderate hemophilia, and (ii) mild hemophilia. We carried out two types of experiments, and compared their results. One is

the SVM analysis of learning phase in some region A (self-learning). Another is the SVM learning using data of other region B, and applied the obtained parameters to the discrimnant analysis of region A. We showed the results in the form of Reciever Operating Characteristic (ROC) curve.

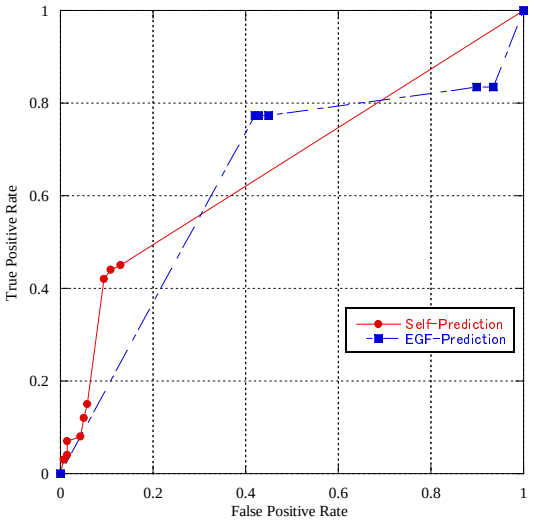

Fig. 1. ROC curves for the discrimination of severity of hemophilia B in EGF region. The solid line shows the SVM result of self-learning in EFF region. The dotted line for SVM learning in ACT region.

Figure 1 shows ROC curves for the discrimination of severity of hemophilia B in EGF region. The solid line shows the SVM result of self-learning. The dotted line shows the ROC curve of EGF data using the parameters obtained by the SVM learning in Activation (ACT) region. Two calculations have almost the same degree of discriminating power of patients' data.

Figure 2 presents two ROC curves in ACT region. The solid line is the results of self-learning, while the dotted line shows SVM learning in EGF region. In this case, the transfer-learning is better than the self-learning.

Figure 3 presents two ROC curves in GLA region. The solid line shows the SVM prediction of self-learning in GLA region. The dotted line is the result of SVM learning in ACT region. This figure presents that the transfer-learning has little power of discrimination.

5 SUMMARY

We applied the method of transfer learning, which uses the knowledge of some domain to predict the properties of other domains. As a training set, we use the clotting activity of mutations in one of seven regions, and test the obtained parameters by the discrimination of mutations in other regions. In some cases, the transfer-learning has the almost the same

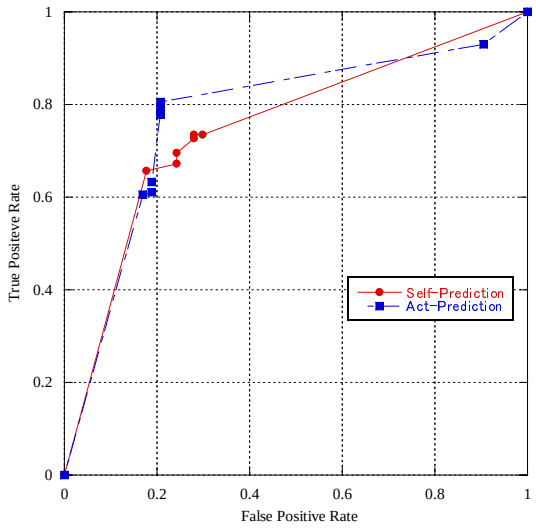

Fig. 2. ROC curves for the discrimination of severity of hemophilia B in ACT region. The solid line shows the SVM result of self-learning in ACT region. The dotted line for SVM learning in EGF region.

power of the discrimination of patients' data. However, there are many cases that the transfer-learning has no power of discrinimination. Thus the next step of this research is to study the mechanism why some transfer-learning has a discrimination power.

REFERENCES

[1] Sternson, P.D., Mort, M., Ball, E. V., Howells, K., Phillips, A.D., Thomas, N.S., Cooper, D.N.: The Human Gene Mutation Database: 2008 update. Genome Medicine, **1:13**, (2009).

[2] Utsunomiya, M., Sakamoto, M., Furutani, H.: Regression Analysis of Amino Acid Substitutions and Factor IX Activity in Hemophilia B. Artificial Life and Robotics, **13** (2008) 531–534

[3] Aoki, K., Yamamori K., Sakamoto. M., Furutani H.: Analysis of Genetic Disease Hemophilia B by Using Support Vector Machine. Proceedings of the 19th International Conference on Neural Information Processing, Lecture Notes in Computer Science, Vol.7666(4), (2012) 476–483

[4] Furie, B., Furie, B.C.: The Molecular Basis of Blood Coagulation. Cell, **53** (1988) 505–518

[5] Yoshitake, S., Schach, B. G., Foster D. C., Davie, E. W. and Kurachi, K.: Nucleotide Sequence of the Gene for Human Factor IX. Biochemistry. **24** (1985) 3736–3750

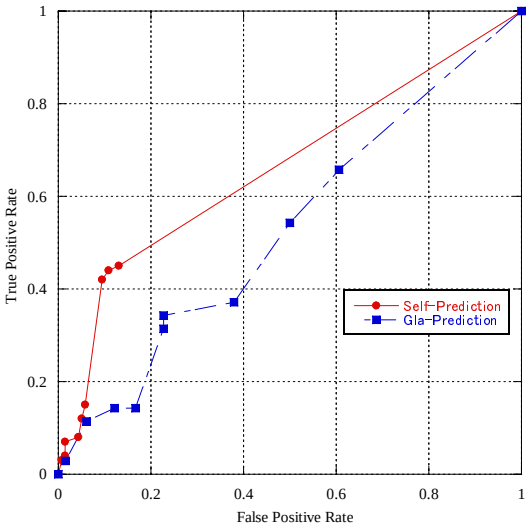

Fig. 3. ROC curves in GLA region. The solid line shows the SVM result of self-learning in GLA region. The dotted line for SVM learning in ACT region.

[6] Hemophilia B Mutation Database: A database of point mutations and short additions and deletions in the factor IX gene. version 13 (2004).

[7] Schölkopf B., Burges, C.J.C., Smola, A.J.: Advances in Kernel Methods, The MIT Press, London (1999)

[8] Schölkopf B., Tsuda K., Vert J.P.: Kernel Methods in Computational Biology, A Bradford Book, The MIT Press, London (2004)

[9] Joachims, T.: Making large-Scale SVM Learning Practical, in Advances in Kernel Methods - Support Vector Learning, The MIT Press, London (1999)