Vol.2011-MPS-85 No.7
2011/9/15

サポートベクターマシンを用いた血友病Bデータベースの解析 –アミノ酸置換の第IX因子活性への影響–

青木謙二[†1] 末吉健二[†2] 石橋太郎[†2]
坂本眞人[†2] 古谷博史[†2]

血友病Bは血液凝固因子の一つである第IX因子の異常によって引き起こされる遺伝性疾患である．血友病Bについては，患者の遺伝子異常がデータベース化され，血友病Bを分子レベルで組織的に研究することが可能である．本論文は，第IX因子のアミノ酸ミスセンス変異において，アミノ酸置換に伴う物理化学的パラメータ（分子体積，疎水性，極性，等電点，残基間接触エネルギー）の変化量を用いてサポートベクターマシンにより第IX因子の活性度を予測し，重症度を推定するものである．この結果，ロジスティック回帰分析による推定結果よりも優れた結果が得られた．また，パラメータとして過去の研究で用いられていなかったアミノ酸の残基間接触エネルギーを加えることにより，よりよい推定結果が得られることが分かった．

Analysis of the database in hemophilia B by Support Vector Machine –Influence of Amino-acid substitution on factor IX–

Kenji Aoki,[†1] Kenji Sueyoshi,[†2] Taro Ishibashi,[†2]
Makoto Sakamoto[†2] and Hiroshi Furutani[†2]

Hemophilia B is caused by decreased activity of factor IX. Mutation in factor IX is made up of a majority of amino acid substitutions. Hemophilia B is possible for a patient's gene abnormality to be put in a database and is studied systematically with a molecular level. In this paper, in the amino acid missense mutation of factor IX, we predict the activity of factor IX by Support Vector Machine (SVM) using the amount of change of the physicochemical parameters accompanying amino acid substitution, and presume severity of illness. As a result, the SVM was superior to Logistic Regression Analysis. Moreover, the better presumed result was obtained by adding the interresidue contact energy.

1. はじめに

近年, 分子遺伝学の急速な進歩により多くの遺伝病の病因が分子レベルで明らかにされつつある．血友病もそのような病気の一つであり，その原因となる遺伝子の異常が多数報告されている．これらの情報はデータベース化され，血友病を分子レベルで組織的に研究することや, 血友病の病因や重症度の統計的手法を用いた解析に利用されている[1)]．

血友病は血液凝固因子欠乏症の1つで伴性劣性遺伝による先天性疾患である[2)]．血友病には大きく分けて血友病Aと血友病Bの2つがあり，このうち血友病Bは血液凝固因子である第IX因子の欠損あるいは活性低下に起因している．第IX因子遺伝子の突然変異の大部分は点突然変異（ミスセンス変異）であり，DNA配列の大幅な欠失や挿入を持つ症例は少数である[3)]．点突然変異ではアミノ酸の置換が最も多く，活性度は変異の起きた部位，アミノ酸置換の種類（もとのアミノ酸と置換したアミノ酸の組み合わせ）に依存してさまざまな値をとる．一般に，重要な部位における変異や性質の異なるアミノ酸への置換は，凝固因子活性の大幅な低下をもたらすものと予想される．

サポートベクターマシンは1963年にV.N. Vapnikが考案した最適分離超平面（Optimal Separating Hyperplan）を起源とし，1995年にカーネル法と組み合わせることのより非線形の識別手法へと拡張された識別器である[4)5)]．サポートベクターマシンは，他の分野と同様に遺伝子解析の領域でも幅広く利用されている．Zienらはサポートベクターマシンを用いて翻訳開始部位の予測を行い，良好な結果を得ている[6)]．また，Brownらは DNA マイクロアレイの発現データから遺伝子機能を分類する問題にこれを適用した[7)]．この他，タンパク質の細胞内での局在部位の予測[8),9)]，alternative splicing の予測[10)] などサポートベクターマシンの様々な応用が報告されている．このように，サポートベクターマシンは，現在知られている手法の中でも最もパターン認識性能の優秀な学習モデルの一つである．

本論文では，第IX因子のアミノ酸ミスセンス変異において，アミノ酸置換に伴う物理化学的パラメータ（分子体積，疎水性，極性，等電点）のアミノ酸間距離に加えて，アミノ酸残基間接触エネルギー[11)12)] を用いることにより，サポートベクターマシンによる第IX因

†1 宮崎大学情報基盤センター
Information Technology Center, University of Miyazaki
†2 宮崎大学大学院工学研究科
Graduate School of Engineering, University of Miyazaki

子の活性度の予測を行い，重症度の推定を行った．また，同様のパラメータを用いたロジスティック回帰との比較を行うことにより，サポートベクターマシンの重症度推定における優位性を検証した．

2. 方　　法

血友病 B の患者データベースをもとに分子体積，疎水性，極性，等電点のアミノ酸間距離，およびアミノ酸残基間接触エネルギーを求めた．これらのデータを用いて，サポートベクターマシンおよびロジスティック回帰を適用し，第 IX 因子の活性度予測を行い，重症度を推定した．指定した結果とデータベースの活性度を比較し，各手法における推定の感度および特異度を求め，各手法の推定性能を評価した．

2.1 アミノ酸間距離

アミノ酸における各種の物理化学的パラメータから単鎖アミノ酸置換における自由エネルギーの変化量をアミノ酸間距離として求めることができる．アミノ酸間距離は式 (1) で表される．

$$D_{ij} =\mid f_i - f_j \mid \tag{1}$$

ここで f_i, f_j は各パラメータのアミノ酸 i および j における値を示す．アミノ酸の物理化学的パラメータには，分子体積（Mv: Moleculer volume），疎水性（Hy: Hydropathy），極性要求（Po: Polar requirement），等電点（Is: Isolectric point）を用いた[1)]．

2.2 アミノ酸残基間接触エネルギー

アミノ酸置換に伴う自由エネルギーの変化量はアミノ酸の残基間接触エネルギーから推定を行うことができる．アミノ酸の残基間接触エネルギー $e^{'}_{ij}$,e_{ij} は式 (2), 式 (3), 式 (4) より導かれる．

$$exp(-2e^{'}_{ij}) = \frac{N_{ij}^2}{N_{ii}N_{jj}} \frac{C_{ii}C_{jj}}{C_{ij}^2} \ for\ i,j \neq 0 \tag{2}$$

$$exp(-2e^{'}_{i0}) = \frac{N_{i0}^2}{N_{ii}N_{00}} \frac{C^{'}_{ii}C^{'}_{00}}{C^{'2}_{i0}} \tag{3}$$

$$e_{ij} = e^{'}_{ij} + e^{'}_{00} - e^{'}_{i0} - e^{'}_{j0} \tag{4}$$

アミノ酸の残基間接触エネルギーの平均値は, 式 (5), 式 (6) より導かれる．

$$\delta(\delta_d G)_{i\rightarrow j} \approx [\sum_k (e_{jk} - e_{ik}) n_{pik} - (f_j - f_i n_p^d)] \tag{5}$$

$$< n_{p,jk} >= \frac{N_{ik}}{N_i} \tag{6}$$

本論文では，アミノ酸残基間接触エネルギーの上流側を LCE，下流側を RCE，平均値を ACE と表記する．

2.3 血液凝固第 IX 因子の領域区分

血液中の第 IX 因子は 415 個のアミノ酸から構成されており[13)]，第 IX 因子タンパク質は 7 つの領域から構成され，各エクソンに分割してコードされている．その 7 つの領域はそれぞれ，Signal peptide(シグナルペプチド)，Propeptide(プロペプチド)，Gla，1stEGF(第 1EGF 領域)，2ndEGF(第 2EGF 領域)，Activation(活性化ペプチド)，Catalytic(触媒領域) と呼ばれている．

本研究では，第 IX 因子の中で細胞内で切断されるシグナルペプチド領域とプロペプチド領域, そして成長因子である EGF 領域を除く，活性度との関連が高いと考えられる Gla 領域, Activation（Act）領域，Catalytic（Cat）領域の 3 つの領域に特に注目し解析を行った．

2.4 ロジスティック回帰

ロジスティック回帰は，医学や社会科学で使われることが多い，ベルヌーイ分布に従う変数の統計的回帰モデルの一種である．まず，n 個の説明変数 $\boldsymbol{x}$ をベクトル表現して式 (7) に表す．

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_r)^T \tag{7}$$

説明変数 $\boldsymbol{x}$ の観測値が与えられているという条件のもとで，ある事象の発生する確率 $p(\boldsymbol{x})$ を式 (8) のように定義する．

$$p(\boldsymbol{x}) = \frac{\exp(g(\boldsymbol{x}))}{1 + \exp(g(\boldsymbol{x}))} \tag{8}$$

ここで $g(\boldsymbol{x})$ は $p(\boldsymbol{x})$ のロジット（logit）, 対数オッズ（log odd）などと呼ばれ, 式 (9) で表される．

$$g(\boldsymbol{x}) = \log\Big(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\Big) = \text{logit } p(\boldsymbol{x}) \tag{9}$$

式 (9) を変形すると式 (10) となり, 重回帰モデルのような式が現れる．

$$\text{logit } p(\boldsymbol{x}) = \log\Big(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\Big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{10}$$

本研究では，説明変数 $\boldsymbol{x}$ として分子体積（Mv），疎水性（Hy），極性要求（Po），等電点（Is），アミノ酸残基間接触エネルギー（LCE, RCE, ACE）それぞれのアミノ酸間距離

を用いる．

2.5 サポートベクターマシン

サポートベクターマシン（SVM）とは，学習サンプルからサポートベクトル（SV）を抽出して識別器を構成し，それに従いサンプルを分類する教師あり識別手法である．SVM はサンプルを分類するために，各サンプル点との距離が最大となる分離平面をマージン最大化という考え方に基づいて求める．SVM の学習はラグランジュ未定乗数法を用いることにより，最適化問題の一種である 2 次計画問題が定式化される．

n 個の学習サンプル x_i $(i = 1, \cdots, n)$ を 2 つのクラス C_1 と C_2 に分類する場合，識別境界を決定する識別関数は式 (11),(12) で定義される．

$$f(\boldsymbol{x}) = \mathrm{sign}(g(\boldsymbol{x})) \tag{11}$$

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \cdot \boldsymbol{x} + h \tag{12}$$

ここで，$\boldsymbol{w}$ は学習サンプル $\boldsymbol{x}$ に対する重みベクトル，h は閾値であり，sign は式 (13) で表される符号関数である．

$$\mathrm{sign}(u) = \begin{cases} 1 & if \quad u > 0 \\ -1 & if \quad u \leq 0 \end{cases} \tag{13}$$

また正解クラスラベル y_i $(i = 1, \cdots, n)$ は式 (14) で定められる．

$$y_i = \begin{cases} 1 & if \quad x_i \in C_1 \\ -1 & if \quad x_i \in C_2 \end{cases} \tag{14}$$

SVM の学習は $\forall_i\ f(\boldsymbol{x}_i) = y_i$ を満たす $\boldsymbol{w}$ を求めることである．よって学習後は式 (15) を満たす．

$$\forall_i\ y_i(\boldsymbol{w}^t \cdot \boldsymbol{x}_i - j) \geq 1 \tag{15}$$

式 (15) は，$H_1 : \boldsymbol{w}^T \cdot \boldsymbol{x}_i + h = 1$ と $H_2 : \boldsymbol{w}^T \cdot \boldsymbol{x}_i - h = -1$ の 2 つの超平面で学習サンプルが完全に分離されており，2 つの超平面の間には学習サンプルが一つも存在しないことを示している．このとき，両クラスの識別平面 H_1，H_2 間の距離はマージンと呼ばれ，その距離は $2/ \parallel \boldsymbol{w} \parallel$ となる．そして，識別関数の汎化能力を高くするよう，$\boldsymbol{w}$ と h を調整する．そのためにはマージンを最大化するとよいので，$\parallel \boldsymbol{w} \parallel$ を最小化するような $\boldsymbol{w}$ および h を求めるということになる．

本研究では，ロジスティック回帰と同様に学習サンプル $\boldsymbol{x}$ として分子体積（Mv），疎水性（Hy），極性要求（Po），等電点（Is），アミノ酸残基間接触エネルギー（LCE, RCE, ACE）それぞれのアミノ酸間距離を用いる．

ここで，問題を扱いやすくするために式 (15) を制約条件として式 (16) で示した目的関数を最小化する二次計画問題を考える．

$$L(\boldsymbol{w}) = \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 \tag{16}$$

さらに，不等式制約条件である式 (15) を直接扱うのは困難なので，式 (15) と式 (16) を双対問題に変形する．双対問題とは，目的関数である式 (15) と制約条件である式 (16) という主問題に，双対変数とも呼ばれるラグランジュ未定乗数を導入することにより得られる補問題である．ラグランジュの未定乗数 $\lambda_i (\geq 0, i = 1, \cdots, n)$ を要素とするベクトル $\boldsymbol{\lambda}$ を導入すると，目的関数である式 (15) を書き換え，ラグランジュ関数 L_p は式 (17) となる．

$$L_p(\boldsymbol{w}, h, \boldsymbol{\lambda}) = \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 - \sum_{i=1}^{n} \lambda_i \{y_i(\boldsymbol{w}^T \cdot \boldsymbol{x}_i - h) - 1\} \tag{17}$$

式 (17) を $\boldsymbol{w}$ および h で偏微分して 0 とおくと，式 (18) と式 (19) の関係を得る．

$$\frac{\partial L_p}{\partial w} = \boldsymbol{w} - \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{x}_i = 0 \tag{18}$$

$$\frac{\partial L_p}{\partial h} = \sum_{i=1}^{n} \lambda_i y_i = 0 \tag{19}$$

ここで式 (18) から式 (20) が得られ，$\boldsymbol{\lambda}$ が求まれば $\boldsymbol{w}$ が求められる．

$$\boldsymbol{w} = \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{x}_i \tag{20}$$

式 (17) に式 (19) と式 (20) を代入すると式 (21) に変形できる．

$$L_p(\boldsymbol{w}, h, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 = L_D(\boldsymbol{\lambda}) \tag{21}$$

よって目的関数である式 (22) と制約条件である式 (23) が得られ，$L_D(\boldsymbol{\lambda})$ を最大化する双対問題が定義できる．

$$L_D(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{22}$$

$$s.t. \quad \forall_i \ \lambda_i \geq 0, \sum_{i=1}^{n} \lambda_i y_i = 0 \tag{23}$$

式 (22) を最大化する λ_i を λ_i^* とおく．ここで，式 (20) を見ると，$\lambda_i = 0$ となるような学習サンプル $\boldsymbol{x}_i$ は $\boldsymbol{w}$ の決定には関与していないことが分かる．つまり，$\lambda_i^* > 0$ となる学習サンプル $\boldsymbol{x}_i$ によって識別関数は決定される．このような学習サンプルをサポートベクトル（SV）と呼び，これらは 2 つの超平面 H_1, H_2 のどちらかの上に存在している．これがサポートベクターマシンによる分類である．

パラメータ $\boldsymbol{w}$ の最適値 $\boldsymbol{w}^*$ は式 (24) で得られる．

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \lambda_i^* y_i \boldsymbol{x}_i \tag{24}$$

また，SV が 2 つの超平面 H_1, H_2 のどちらかの上に存在しているという関係より式 (25) が成り立つ．

$$\forall_i \ \lambda_i^* \{ y_i (\boldsymbol{w}^{*T} \cdot \boldsymbol{x} - h) - 1 \} = 0 \tag{25}$$

よってパラメータ h の最適値 h^* は，任意の $\boldsymbol{x}_s (s \in SV)$ より式 (26) として求められる．

$$h^* = y_s - \boldsymbol{w}^{*T} \cdot \boldsymbol{x}_s \tag{26}$$

以上より，最適な識別関数は式 (27) となる．

$$f(\boldsymbol{x}) = \mathrm{sign}(\sum_{i \in SV} \lambda_i y_i \boldsymbol{x}_i^T \cdot \boldsymbol{x} + h^*) \tag{27}$$

2 つのクラスのサンプルが入り混じって線形分離できない場合, 式 (15) を満たす $\boldsymbol{w}$ は存在しない．その場合は, 学習サンプルの誤識別をある程度許容するように制約をパラメータ α_i(i=1,⋯,n) を用いて緩めるソフトマージン法と呼ばれる手法を用いる．ソフトマージン法では, マージン $1/\parallel \boldsymbol{w} \parallel$ を最大としながら, いくつかのサンプルが超平面 H_1 あるいは H_2 を越えて反対側に存在することを許す．

ソフトマージン法により, 学習サンプルが線形分離不可能な場合でも識別関数を決定するパラメータを求めることが可能となる．しかし, ソフトマージン法を用いたとしても, 本質的に非線形で複雑なサンプル集合の問題には識別器を構成できない場合がある．このような問題に対応するための方法として, カーネルトリックと呼ばれる方法がある．カーネルトリックでは学習サンプルを高次元空間に写像するが, 実際には写像された空間での特徴の計算を避けて識別関数を求めることができる．一般的に使用されるカーネルには，線形カーネル，d 次多項式カーネル，RBF カーネル，シグモイドカーネルなどがあり，本研究では，カーネル関数として式 (28) で表される RBF カーネルを用いた．

$$\exp(-\frac{\parallel \boldsymbol{x}_i - \boldsymbol{x}_j \parallel^2}{2\sigma^2}) \tag{28}$$

また，SVM プログラムには T. Joachims らが提供している SVM-Light[14)15)] を用いた．

2.6 患者データベース

解析に使用するデータは実際の血友病 B 患者から得られた情報を P.M. Green らによってデータベースにまとめられた fixhome（Haemophilia B Mutation Database）[16)] からダウンロードした 2924 人の患者データのうち，ミスセンス変異による男性患者 1493 人分のデータを抽出して使用した．このデータベースには，患者 ID，第 IX 因子活性度，ヌクレオチド番号，突然変異の種類，アミノ酸タンパク質中の位置および変異前後のアミノ酸が患者ごとに記されている．

また，ヌクレオチド番号の領域分割に従って，データを領域毎に分け，領域毎の解析に用いた．領域毎のデータ数は Gla 領域，Act 領域，Cat 領域それぞれ，99，241，795 である．

2.7 感度，特異度，ROC 曲線

一般的に，臨床検査においてその有効性を評価する指標として，感度および特異度を用いる．ある疾患の診断に用いる検査を評価する場合，診断対象の各患者を一定の確立した基準に基づき検査の成績とは独立に疾患の有無を決定する．次に，各患者に対して検査を行い，それらの結果をまとめる．感度とは,「陽性と判定されるべきものを正しく陽性と判定する可能性」を意味し，(検査陽性数)/(疾患陽性総数) で，特異度とは,「陰性のものを正しく陰性と判定する可能性」を意味し，(検査陰性数)/(疾患陰性総数) で表される．感度および特異度の両方が高いほど良い判別ができたことになる．

感度および特異度はカットオフポイント（陽性と陰性を判定するための閾値）をどこに取るかによって値が変動する．そこで，最適なカットオフポイントを探るための手法として ROC 曲線がある．ROC 曲線を作成するには, まず真陽性率と偽陽性率を求める．真陽性率は感度と同じものである．偽陽性率は 1 から特異度を引いた値となる．横軸に偽陽性率，縦軸に真陽性率をとってカットオフポイントを変更した場合のそれぞれにおいてプロットする．これが ROC 曲線である．この曲線上で座標 (0,1) から最も距離の近い点がカットオフポイントとして最良のものとなる．さらに，ROC 曲線は複数の検査の識別能力の比較する

Vol.2011-MPS-85 No.7
2011/9/15

ためにも用いられる．診断対象が同じである複数の検査法の ROC 曲線を比べ, その曲線が左上に近いものほど識別能力が高いと判断することができる．

3. 結　　果

3.1 SVM による重症，軽症推定

第 IX 因子活性度が 1% 未満の患者を重症，1% 以上を軽症とみなし，このデータをもとに SVM による活性度の予測を行い，重症，軽症を推定した．第 IX 因子全領域，Gla 領域，Act 領域，Cat 領域における推定結果と実際の患者データ（実測結果）との関係をそれぞれ**表 1**，**表 2**，**表 3**，**表 4** に示す．また，この結果から求めた感度，特異度をそれぞれの領域ごとに**表 5** に示す．カットオフポイントはすべて 0.6 とし，物理化学的パラメータには LCE，RCE，ACE，Mv，Hy，Po，Is の 7 つのパラメータを用いた．

表 1　第 IX 因子全領域における SVM による判別結果
Table 1　Discriminant result by SMV in All domain of factor IX.

		推定		
		軽症	重症	計
実測	軽症	746	200	946
	重症	105	442	547
	計	851	642	1493

表 2　第 IX 因子 Gla 領域における SVM による判別結果
Table 2　Discriminant result by SMV in Gla domain of factor IX.

		推定		
		軽症	重症	計
実測	軽症	50	13	63
	重症	0	36	36
	計	50	49	99

第 IX 因子の全領域，Gla 領域，Act 領域，Cat 領域における感度および特異度はそれぞれ, (感度：80.80%，特異度：78.85%) , (感度：100.00%，特異度：79.36%) , (感度：90.97%，特異度：94.85%) , (感度：72.00%，特異度：87.70%) であった．どの領域においても，感度は 72% 以上，特異度は 78% 以上の非常に高い割合を示している．

表 3　第 IX 因子 Act 領域における SVM による判別結果
Table 3　Discriminant result by SMV in Act domain of factor IX.

		推定		
		軽症	重症	計
実測	軽症	92	5	97
	重症	13	131	144
	計	105	136	241

表 4　第 IX 因子 Cat 領域における SVM による判別結果
Table 4　Discriminant result by SMV in Cat domain of factor IX.

		推定		
		軽症	重症	計
実測	軽症	478	67	545
	重症	70	180	250
	計	548	247	795

表 5　SVM による判別の感度，特異度
Table 5　Sensitivity and specificity of SVM.

領域	感度	特異度
全領域	80.80%	78.85%
Gla 領域	100.00%	79.36%
Act 領域	90.97%	94.85%
Cat 領域	72.00%	87.70%

3.2 ロジスティック回帰による重症・軽症推定

SVM と同様に，第 IX 因子活性度が 1% 未満の患者を重症，1% 以上を軽症とみなし，このデータをもとにロジスティック回帰による重症，軽症推定を行った．第 IX 因子全領域，Gla 領域，Act 領域，Cat 領域それぞれにおける推定結果と実測結果との関係をそれぞれ**表 6**，**表 7**，**表 8**，**表 9** に示す．また，この結果から求めた感度，特異度をそれぞれの領域ごとに**表 10** に示す．カットオフポイントは 0.6 とし，パラメータには LCE，RCE，ACE，Mv，Hy，Po，Is の 7 つを用いた．

第 IX 因子の全領域，Gla 領域，Act 領域，Cat 領域における感度および特異度はそれぞれ, (感度：53.02%，特異度：72.52%) , (感度：58.33%，特異度：69.84%) , (感度：74.23%，特異度：56.94%) , (感度：46.00%，特異度：81.47%) であった．

表 6 第 IX 因子全体におけるロジスティック回帰による判別結果
Table 6 Discriminant result by Logistic Regression in All domain of factor IX.

		推定		
		軽症	重症	計
実測	軽症	686	260	946
	重症	257	290	547
	計	943	550	1493

表 7 第 IX 因子 Gla 領域におけるロジスティック回帰による判別結果
Table 7 Discriminant result by Logistic Regression in Gla domain of factor IX.

		推定		
		軽症	重症	計
実測	軽症	44	19	63
	重症	15	21	36
	計	59	40	99

表 8 第 IX 因子 Act 領域におけるロジスティック回帰による判別結果
Table 8 Discriminant result by Logistic Regression in Act domain of factor IX.

		推定		
		軽症	重症	計
実測	軽症	82	62	97
	重症	25	72	144
	計	107	134	241

表 9 第 IX 因子 Cat 領域におけるロジスティック回帰による判別結果
Table 9 Discriminant result by Logistic Regression in Cat domain of factor IX.

		推定		
		軽症	重症	計
実測	軽症	444	101	545
	重症	135	115	250
	計	579	216	795

表 5 と表 10 に示した値をそれぞれ比較すると，表 10 に示した値の方が小さい．このことから，SVM による判別推定結果の方がロジスティック回帰による判別推定結果よりもよいことがわかる．

表 10 ロジスティック回帰による判別の感度，特異度
Table 10 Sensitivity and specificity of Logistic Regression.

領域	感度	特異度
全領域	53.02%	72.52%
Gla 領域	58.33%	69.84%
Act 領域	74.23%	56.94%
Cat 領域	46.00%	81.47%

3.3 判別性能の比較

カットオフポイント 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1 それぞれに対して SVM およびロジスティック回帰により，重症，軽症推定を行い，ROC 曲線を求めた．**図 1** に SVM とロジスティック回帰の ROC 曲線を重ね書きした．図 1 の（a），(b)，(c)，(d）はそれぞれ，第 IX 因子の全領域，Gal 領域，Act 領域，Cat 領域を使って推定を行った結果の ROC 曲線である．図 1 は横軸に偽陽性率，縦軸に真陽性率をとり，◆が SVM の結果を，■がロジスティック回帰の結果を示している．

SVM とロジスティック回帰の RCO 曲線を比較すると，どの領域における結果でも SVM の ROC 曲線の方が図のより左上に位置しており，この結果から SVM の方がロジスティック回帰よりも優れた判別性能を持つことがわかる．

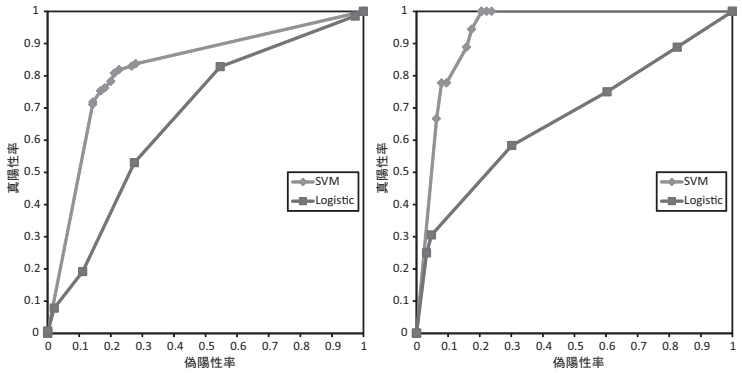
また，同様に SVM においてパラメータを LCE，RCE，ACE，Mv，Hy，Po，Is の 7 つを使った場合と，LCE，RCE，ACE を除く 4 つを使った場合の重症，軽症推定を行い，ROC 曲線を求めた．**図 2** にパラメータが 7 つの場合と 4 つの場合の ROC 曲線を重ね書きした．図 2 の（a），(b)，(c)，(d）はそれぞれ，第 IX 因子の全領域，Gal 領域，Act 領域，Cat 領域を使って推定を行った結果の ROC 曲線である．図 2 は横軸に偽陽性率，縦軸に真陽性率をとり，◆が 7 つのパラメータを使用した時の結果を，■が 4 つのパラメータを使用した時の結果を示している．

SVM とロジスティック回帰の RCO 曲線を比較すると，Act 領域（図 2(c)）では明確ではないものの，どの領域における結果でも 7 つのパラメータを用いた予測の方が図のより左上に位置しており，この結果らアミノ酸間残基エネルギー（LCE，RCE，ACE）をパラメータに加えた方がより優れた判別性能を持つことがわかる．

4. 考　　察

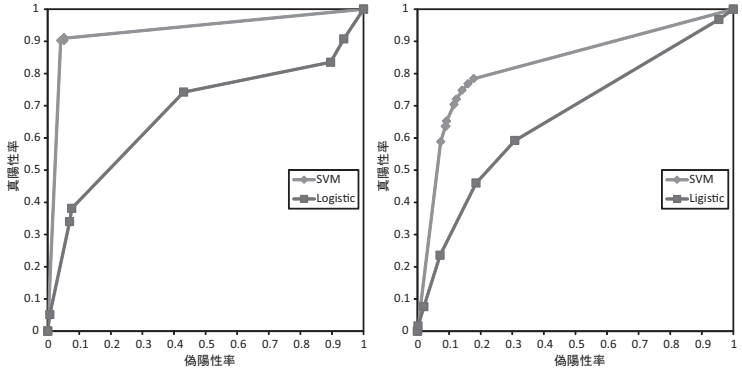
ROC 曲線の比較結果より，サポートベクターマシンを用いた重症度推定の方がロジス

Vol.2011-MPS-85 No.7
2011/9/15



(a) 全領域

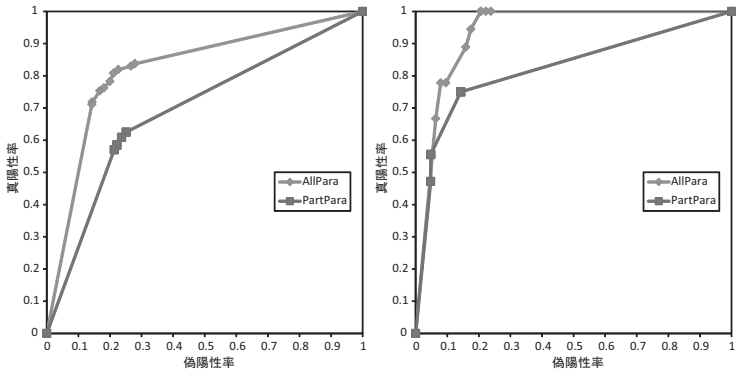
(b) Gla 領域



(c) Act 領域

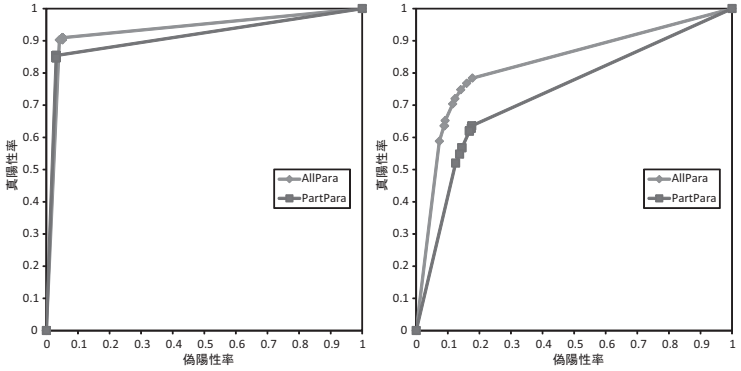
(d) Cat 領域

図 1 ROC 曲線による比較（SVM vs ロジスティック回帰）
Fig. 1 Comparison SVM with Logistic Regression.



(a) All

(b) Gla



(c) Act

(d) Cat

図 2 ROC 曲線による比較（全パラメータ vs 一部パラメータ）
Fig. 2 Comparison all parameters with partial parameters.

ティック回帰よりも優れていることが分かった．図 1 に示した SVM の ROC 曲線を見ると，座標 (0, 1) に近い狭い範囲に点が集まる傾向が見られる．これは，感度と特異度のバランスがよく，感度を上げると特異度も上がる傾向にあることを示している．これに対し，ロジスティック回帰の ROC 曲線は，全体に分布しており，感度が上がると特異度が下がり，逆に，特異度が上がると感度が下がる傾向にあることを示している．ロジスティック回帰の場合，最適な感度と特異度を得る範囲が狭いことがわかる．このような特性から，SVM の方が感度および特異度が高くなることが考えられる．

また，SVM において 7 つのパラメータを使った方が 4 つのパラメータを使った方よりも重症度推定が優れていることが分かった．LCE，RCE，ACE と第 IX 因子の活性度との相関を求めると，相関が高いことが分かっている．つまり，LCE，RCE，ACE は活性度に強く影響を与えており，そのためパラメータとして用いることにより，より優れた推定結果を求めることができると考えられる．

5. おわりに

本研究では血液凝固第 IX 因子の活性度を 1% 未満を重症とし，それ以外を軽症とした血友病 B 患者の重症，軽症の分類推定をサポートベクターマシンにより行った. 分類を行うために，活性度と関連のあるパラメータとして分子体積，疎水性，極性要求，等電点に加え，アミノ酸残基間接触エネルギー (上流，下流，平均値) の 7 つを使用した．また，ロジスティック回帰による重症，軽症の分類推定を行い，サポートベクターマシンの結果と比較した．この結果，サポートベクターマシンによる重症，軽症推定の方が，ロジスティック回帰による推定より優れていることが分かった．パラメータを 7 つ使用した場合と 4 つを使用した場合では，7 つを使用した方が優れた分類結果が得られた．今後の課題として，どのパラメータがどこの領域に最も影響を与えているのかを細分化して検討することが必要である．

また，コンピュータを用いた変異タンパク質の機能を予測する研究として，Ng らのアミノ酸配列情報を基にした研究[17] や Prokop らのタンパク質 3 次元立体構造データを利用した研究手法[18] などが考案されているが，本研究の結果とこれらの結果を比較し，どのような手法が優れているのかを検証することも今後の課題である．さらに，これらの手法を組み合わせることによって，活性度予測性能がさらに向上する可能性も考えられることから，これを今後の課題としたい．

参 考 文 献

1) 古谷博史：血友病 B における第　因子アミノ酸置換と活性の相関分析，医療情報学会論文誌，　Vol.14, No.4, pp.211–220 (1994).
2) Furie, B. and Furie, B.C.: The Molecular Basis of Blood Coagulation, *Cell*, Vol.53, pp.505–518 (1988).
3) Giannelli, F., Green, P., High, K., Sommer, S., Lillicrap, D., Ludwig, M., Olek, K., Reitsma, P., Goossens, M., Yoshioka, A. and Brownlee, G.: Haemophilia B: database of point mutations and short additions – second edition, *Nucleic Acids Research*, Vol.19, pp.2193–2219 (1991).
4) Vapnik, V.N.: *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc. (1995).
5) 赤穂昭太郎：カーネル多変量解析，岩波書店 (2008).
6) Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T. and Müller, K.-R.: Engineering support vector machine kernels that recognize translation initiation sites, *Bioinformatics*, Vol.16, No.9, pp.799–807 (2000).
7) Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Jr., M.A. and Haussler, D.: Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proceedings of the National Academy of Sciences*, Vol.97, No.1, pp.262–267 (2000).
8) Hua, S. and Sun, Z.: Support vector machine approach for protein subcellular localization prediction, *Bioinformatics*, Vol.17, No.8, pp.721–728 (2001).
9) Xie, D., Li, A., Fan, M. W.Z. and Feng, H.: LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST, *Nucleic Acids Research*, Vol.33, pp.w105–w110 (2005).
10) Dror, G., Sorek, R. and Shamir, R.: Accurate identification of alternatively spliced exons using support vector machine, *Bioinformatics*, Vol. 21, No. 7, pp. 897–901 (2005).
11) Miyazawa, S. and Jernigan, R.L.: Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation, *Macromolecules*, Vol.18, No.3, pp.534–552 (1985).
12) Miyazawa, S. and Jernigan, R.L.: Protein stability for single substitution mutants and the extent of local compactness in the denatures state, *Protein Engineering*, Vol.7, No.10, pp.1209–1220 (1994).
13) Yoshitake, S., Schach, B.G., Foster, D.C., Davie, E.W. and Kurachi, K.: Complete Nucleotide Sequence of the Gene for Human Factor IX (Antihemophilic Factor B), *Biochemistry*, Vol.24, No.14, pp.3736–3750 (1985).
14) Joachims, T.: *Making large-Scale SVM Learning Practical*, chapter11, pp.41–56, MIT Press (1999).
15) Joachims, T.: SVM^{light} Suport Vector Machine, Version:6.02 (2008).
16) Green, P., Giannelli, F., Sommer, S., Poon, M.-C., Ludwig, M., Schwaab, R., Reitsma, P., Goossens, M., Yoshioka, A., Figueiredo, M., Tagariello, G. and Brownlee, G.: fixhome: The Haemophilia B Mutation Database – version 13 (2004).
17) Ng, P. C. and Henikoff, S.: Predicting Deleterious Amino Acid Substitutions, *Genome Research*, Vol.11, pp.863–874 (2001).
18) Prokop, M., Damborský, J. and Koča, J.: TRITON: in silico construction of protein mutants and prediction of their activities, *Bioinformatics*, Vol.16, No.9, pp. 845–846 (2000).