

生物種間の配列類似性を利用した血友病第 IX 因子遺伝子の解析

末吉健二^{†1} 汪巍^{†1} 青木謙二^{†2} 坂本真人^{†3} 古谷博史^{†3}

血友病 B は第 IX 因子遺伝子の突然変異によって引き起こされる。我々は、第 IX 因子遺伝子の突然変異のデータベースを用いて研究を行った。本研究では、生物種間のアミノ酸の配列類似性を使用し、重回帰分析とサポートベクターマシンを用いて突然変異の解析を行った。この結果、アミノ酸の配列類似性と第 IX 因子活性の関連が示唆された。

Analysis of Factor IX in hemophilia using sequence similarity across species

KENJI SUEYOSHI^{†1} GI OU^{†1} KENJI AOKI^{†2}
MAKOTO SAKAMOTO^{†3} HIROSHI FURUTANI^{†3}

Hemophilia B is caused by mutation in factor IX gene. We study the factor IX missense mutation database. We analyze the mutations using sequence similarity across species, and apply the multiple regression method and support vector machine. The results show the importance of similarities of amino acid sequences.

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子の病因が分子レベルで明らかにされつつある。その中で、血友病もその原因となる遺伝子の異常が多数報告されている。そのため、血友病は患者の遺伝子異常が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が可能となった[1]。

生物の進化の過程において生物の遺伝子には変化のあるものと変化の無いものが存在しており、生物の遺伝子に変化が無い部分には進化の過程において重要な構造及び機能的役割を持つことが考えられる[2]。

本報告では、ホモ・サピエンスとそれ以外の生物 10 種の第 IX 因子遺伝子の配列類似性を用いて凝固因子活性の関連を調べた。また、重回帰分析、サポートベクターマシンを用いて患者データの解析を行った。

2. 理論

ホモ・サピエンスの第 IX 因子遺伝子と他の動物の第 IX 因子遺伝子の配列類似性を比較するため、BLAST[3]を用いて、シーケンスアライメントを行った。10 種類の生物とホモ・サピエンスの第 IX 因子遺伝子のアミノ酸配列と比較し、類似したものの個数を類似スコア(AV)とした。

解析には、(1)分子体積、(2)疎水性、(3)極性要求、(4)等電点の 4 つのパラメータに AV を加えた計 5 つのパラメータを用いた。

(1)は大きさ、(2)と(3)は極性、(4)は電荷に関係した量である。各パラメータにおいて、アミノ酸 i と j の間の距離 D_{ij} を次の式で定義する。

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

ここで f_i, f_j は各パラメータのアミノ酸 i 及び j における値を表す。

5 つのパラメータを用いて、第 IX 因子活性の予測値を求め、重相関係数を求めた。予測値を求める式は次の式で表わされる。

$$E_{ij} = \alpha|a_i - a_j| + \beta|b_i - b_j| + \gamma|c_i - c_j| + \delta|d_i - d_j| + \zeta AV$$

ここで、 i, j はアミノ酸の種類を表し、 a は分子体積、 b は疎水性、 c は極性要求、 d は等電点、 AV は類似スコアを表し、係数 $\alpha, \beta, \gamma, \delta, \zeta$ は重回帰分析により求める。

最後に、サポートベクターマシンを使って、重症と軽症の分類を行った。

サポートベクターマシンは、ソフトマージンとカーネルトリックを用いた。このとき、サポートベクターマシンの特徴空間での双対問題は次式で表される。

$$\max \quad Q(\lambda) = \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i,j=1}^n \lambda_i \lambda_j y_i y_j K(x_i, x_j)$$
$$\text{s.t.} \quad \sum_{i=1}^n y_i \lambda_i = 0, \quad 0 \leq \lambda_i \leq C \quad (i = 1 \dots n)$$

ここで y_i, y_j は正解クラスラベルであり、活性度 1%未滿を重症とし、クラス -1 を重症、+1 を軽症とする。K はカーネルで、ここでは RBF カーネルを使用している。これを用いて λ を求め、決定関数を次式で決定する。

$$D(x) = \sum_{i \in S} \lambda_i y_i K(x_i, x_j) + b$$

活性度の実測値及びサポートベクターマシンで予測値を用

^{†1} 宮崎大学工学研究科
Graduate School of Engineering, University of Miyazaki
^{†2} 宮崎大学情報基盤センター
Information Technology Center, University of Miyazaki
^{†3} 宮崎大学工学部
Faculty of Engineering, University of Miyazaki

いて感度と特異度を導いた。求めた感度及び特異度から ROC 曲線を使って最適なカットオフポイントを選んだ。

3. 数値計算

解析には fixhome[4]の 1497 人のデータを用いて解析を行った。はじめに AV で単回帰分析を行った結果、相関係数は 0.1907 となり、4つのパラメータの中で最も高い相関係数 0.1727 よりも高かった。

次に4つのパラメータと5つのパラメータを使って重回帰分析を行った。4つのパラメータと AV を加えた5つのパラメータを用いた重回帰分析では、重相関係数が 0.2925 で4パラメータの 0.2416 より高い結果が得られた。

次に、5つのパラメータを用いて、サポートベクターマシンを使った重症・軽症の分類の結果を活性度 1%未満の場合を表 1 に、活性度 5%未満の場合を表 2 に示す。また、図 1 に活性度 1%未満を重症とした場合、図 2 に活性度 5%未満を重症とした場合の第IX因子全体におけるサポートベクターマシンの ROC 曲線の結果を示す。縦軸が真陽性率、横軸が偽陽性率の図である。

表 1：活性度 1%未満

		予測		
		軽症	重症	計
実測	軽症	794	144	938
	重症	139	420	559
	計	933	564	1497

表 2：活性度 5%未満

		予測		
		軽症	重症	計
実測	軽症	258	208	466
	重症	53	978	1031
	計	311	1186	1497

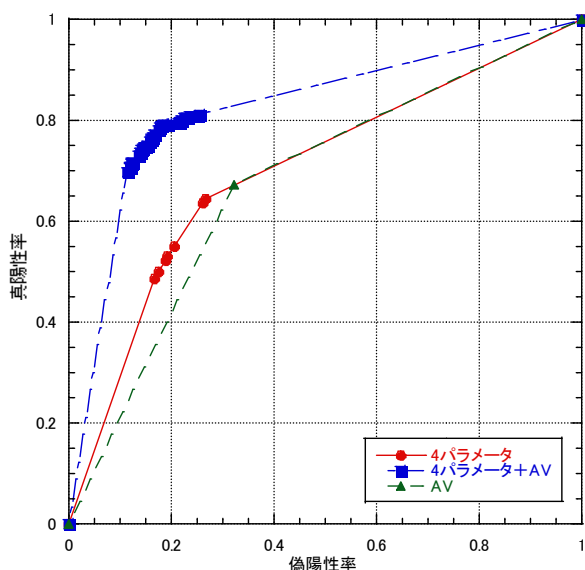


図 1：活性度 1%未満

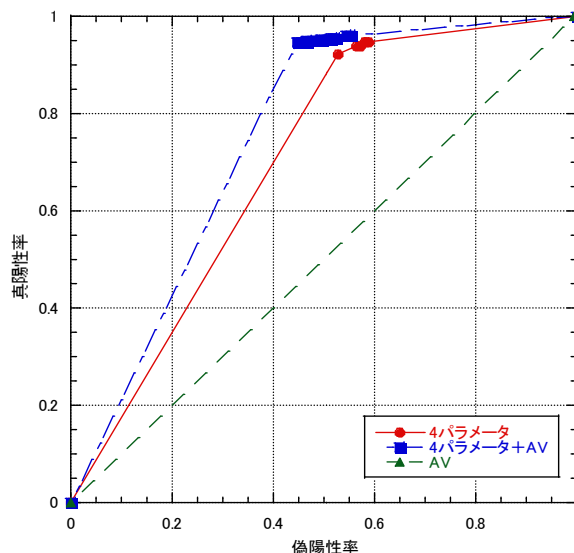


図 2：活性度 5%未満

このとき、活性度 1%未満のカットオフポイントは 0.35 で、活性度 5%未満のカットオフポイントは 0.01 である。表 1 から活性度 1%未満の感度は 75.13%、特異度は 84.65% であった。4つのパラメータを使ったとき、感度は 54.92%、特異度は 79.53%であり、AV を加えたほうが感度と特異度が高くなっていることがわかる。表 2 から活性度 5%未満の感度は 94.86%、特異度は 55.36%であった。4つのパラメータを使ったとき、感度は 92.34%、特異度は 47.21%であり、AV を加えたほうが感度と特異度が高くなっていることがわかる。

4. おわりに

アミノ酸の配列類似性を用いた重回帰、サポートベクターマシンによる解析を行った。その結果、配列類似性が血友病第 IX 因子遺伝子活性度と関連することが示唆された。本報告では、類似スコア (AV) を用いて研究を行ったが、今後は配列類似性と立体構造を組み合わせるさらなる検証を重ねていきたい。

参考文献

- 1) 古谷博史, “血友病 B における第 IX 因子アミノ酸置換と活性の相関分析” 医療情報学 Vol.13 ,No.4 ,pp.211-220(1994)
- 2) Ewy Mathe, Magali Olivier, Shunsuke Kato, Chikashi Ishioka, Pierre Hainaut and Sean V. Tavtigian, “Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods”, Nucleic Acids Research, Vol.34, No.5, 1317-1325(2006)
- 3) BLAST: Basic Local Alignment Search Tool, <http://blast.ncbi.nlm.nih.gov/>
- 4) fixfome, <http://www.kcl.ac.uk/ip/petergreen/haemBdatabase.html>