ニューラルネットワークを用いたDNA突然変異によるスプライシング異常の解析

ニューラルネットワークを用いたDNA突然変異によるスプライシング異常の解析

古谷 博史[1)]　　林田　裕一[2)]

Neural network analysis of abnormal splicing caused by DNA mutation

Hiroshi FURUTANI　　Yuichi HAYASHIDA

Abstract

In eukaryotic organisms, DNA sequences called introns do not contribute to genetic information, and are interspersed within the amino acid-coding regions called exons. Introns in the primary transcripts are cleaved out of the transcripts, and exons are spliced together. This process of splicing is believed to play an important role in the evolution of organisms. A computer program is presented to predict the positions of the splice junction by a three-layered neural network method. We report the results of its application to the analysis of two genetic diseases, hemophilia and thalassemia.

Key Words:
Neural network, Gene, Splicing, Genetic disease, Hemophilia, Thalassemia

1　はじめに

真核生物では、多くの遺伝子においてアミノ酸配列などの遺伝情報をコードする領域（エクソン）の間にアミノ酸として翻訳されない領域（イントロン）が存在し、核内で直接の転写産物（mRNA 前駆体）から除去される。この現象—スプライシング—は、生物の進化や遺伝情報の発現などにおいて重要な意味を持つと考えられているが、その内容についてはまだ不明な点が多く、エクソンとイントロンの接合部位の特定は難しい [1]。

本研究では、入力層と出力層の間に隠れ素子からなる中間層を組み込んだ3層のニューラルネットワークを、スプライシング接合部位の認識問題に適用した [2]。ニューラルネットワークの学習には、バックプロパゲーション法（誤差逆伝播法）を用いた [3]。得られたニューラルネットワークを血友病 B と β サラセミア患者の遺伝子に適用し、突然変異による異常なスプライシングとの関係を調べた。

[1)]情報システム工学科教授
[2)]情報システム工学科学生

2　ニューラルネットワーク

以下の3層の階層型ネットワークを考える。

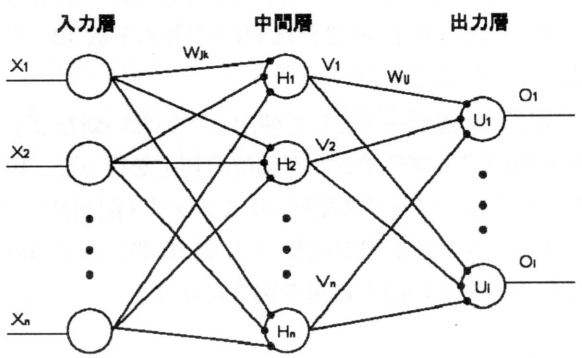

図1　ネットワーク構成図

中間層の結合荷重は W_{jk} とし、これは入力層 k 番目のユニットから、中間層 j 番目のユニットへの結合荷重を表している。また出力層の結合荷重は W_{ij} とし、これは中間層 j 番目のユニットから、出力層 i 番目のユニットへの結合荷重を表す。バックプロパゲーション法では、入力データを入れて出てきた出力を教師信号（望ましい出力）と比較し、違っているときは結合荷重を変更していく。求められた修正量を次の学習時の結合荷重に反映させていき、誤差 E ができるだけ0

に近い値になるまで学習を繰り返す。

3　スプライシング部位の推定

スプライシングはmRNA前駆体で起こる現象であるが、実際の解析はＤＮＡの塩基配列に対して行った。DNAはアデニン(A)、グアニン(G)、チミン(T)、シトシン(C)の4塩基で構成されている。これらをニューラルネットワークの学習データとするため、0と1で符号化する場合、ハミング距離を等しくするよう1塩基につき4ビットを与え、以下のように符号化した。

G→0001　A→0010　C→0100　T→1000

エクソンとイントロンの接合部位には、コンセンサス配列があることが知られており、イントロンの上流と下流にあるＧＴとＡＧの配列はごく少数の例外を除いてほとんど全ての接合部に見られる（ＧＴ・ＡＧ則）。今回はイントロンの上流（以後GT側）10bp(base pair; 塩基対)、下流（以後AG側）15bpとし、スプライス配列を200個と非スプライス配列800個を抽出し、符号化したものをニューラルネットワークの学習データとして入力した。また、教師信号はスプライス配列に対しては1、非スプライス配列に対しては0となるものとし、入力素子40または60、中間素子数40、出力素子1で学習させた。

次に、学習データとして使用していない新たなデータを用いてスプライシング部位の判定を行った。用いたデータはスプライス配列・非スプライス配列共に100個ずつ（各々ＧＴ側50個、ＡＧ側50個）の計200個である。各々の出力結果を図2に示す。

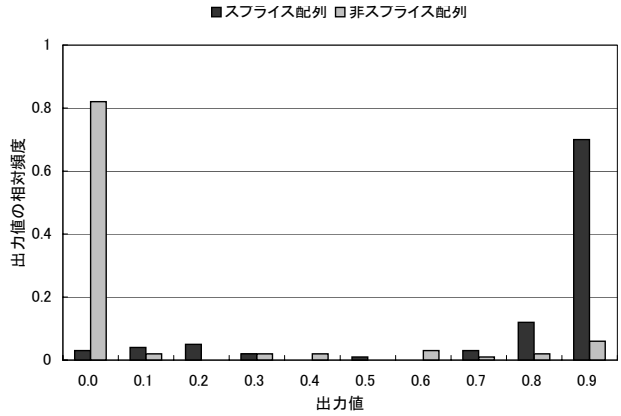

図2　出力の相対頻度

4　血友病における突然変異

4.1　血友病

血友病は先天性血液凝固障害の一つである。その他にも様々な先天性血液凝固障害があるが、そのうち先天性血液凝固第VIII因子障害（血友病Ａ）と先天性血液凝固IX因子障害（血友病Ｂ）を総称して、血友病と呼ぶ。

血友病は伴性遺伝する病気で、性染色体であるX染色体上にある血液凝固因子の第VIII因子、第IX因子をコードする遺伝子に変異が入ることによって引き起こされる。劣性遺伝子であるため、X染色体が二本ある女性の場合は、もう一方のX染色体に異常がなければ機能が補完されるため、発症する事がない。そのため血友病患者のほとんどが男性であり、女性は全血友病患者の1％以下である。基本的には遺伝病であるが、突然変異により非保因者の女性から血友病の子供が生まれる場合もある。実際には血友病患者の25％前後が、家族歴が不明の突然変異と言われている。ただし母親の保因/非保因の検査、判断が難しく、最終的にはDNA検査を要する。

4.2　血友病Bにおける突然変異

血友病Ｂデータベース[4]を元に突然変異とスプライスとの関係をニューラルネットワークを用いて検証した。まず、ＧＴ側スプライシング部位付近の突然変異例とその出力分布を表3.1と図3.1に示す。なお、表の()内の数字は配列中の何番目の塩基が突然変異したかを表しており、Commentsはデータベース中に記入されているものである。

入力番号	Mutation	Sequence		Comments	Output
1	突然変異前 ⇒	TACAG	GTTTG		0.9524
2	(-5) T→G	GACAG	GTTTG	S	0.6311
3	(-3) C→T	TATAG	GTTTG		0.9936
4	(-2) A→G	TACGG	GTTTG	D?	0.3673
5	(-1) G→A	TACAA	GTTTG	D?	0.0569
6	(-1) G→C	TACAC	GTTTG	D?	0.0971
7	(1) -GTTT	TACAG	GTTTC		0.0190
8	(5) G→A	TACAG	GTTTA	D	0.2955
9	(5) G→C	TACAG	GTTTC	D	0.0190
10	(5) G→T	TACAG	GTTTT	D	0.1284

S:Signal peptide　D:Donner splice

表3.1　血友病Ｂ突然変異（ＧＴ側）の一部

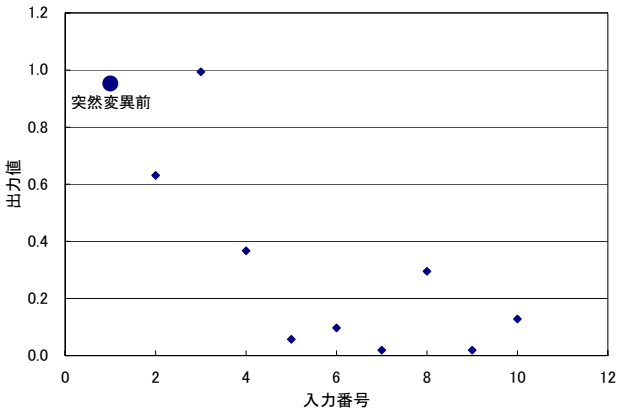

図 3.1　出力分布図

このスプライス部位では、-1 番目のＧと 5 番目のＧが突然変異を起こしたときにニューラルネットワークの出力が大きく変化した。

次に、ＡＧ側スプライシング部位付近の突然変異の一部を示す。

入力番号	Mutation	Sequence		Comments	Output
1	突然変異前 ⇒	TTCTTTATAG	ACTGA		0.9995
2	(-12) -TATTCTTTAT	TCTTCTTTAG	ACTGA	A	0.9978
3	(-3) T→G	TTCTTTAGAG	ACTGA	N	0.3505
4	(3) -TG	TTCTTTATAG	ACAAT	F	0.9910
5	参考	TATTCTTTAG	AGACT		0.9205

A: Acceptor splice　N: New AG　F: Frameshift

表 3.2　血友病Ｂ突然変異（ＡＧ側）の一部

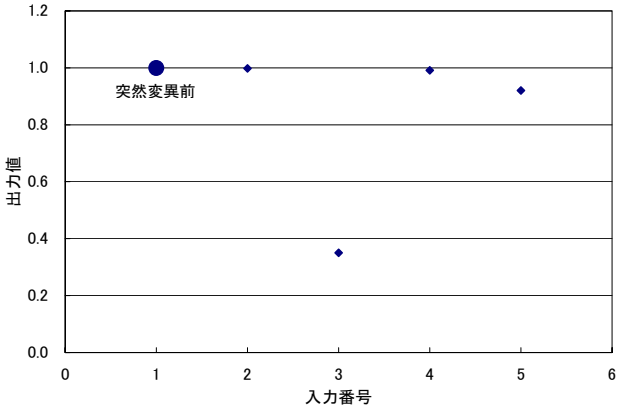

図 3.2　出力分布図

ここでは、-3 番目のＴが突然変異した時に出力の変化が見られた。この時、ＴがＧに変わることで新たなＡＧが生まれる。仮に、新たなＡＧの部分をスプライシング部位と見て出力値を計算すると、表中の参考に示すように１に近い値となる。よって、このＡＧが翻訳時にスプライス部位と判断される可能性があると思われる。

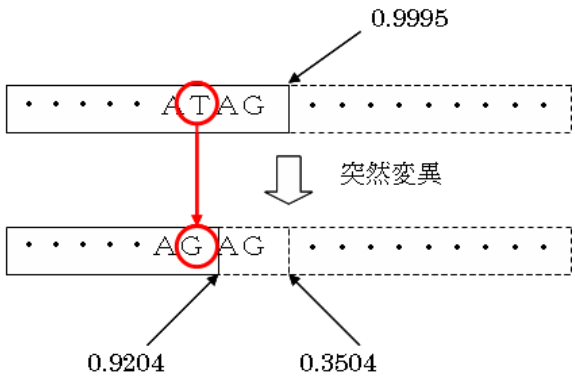

図 3.3　突然変異によるスプライシング異常

このようにして、全てのスプライス部位についてニューラルネットワークを用いた解析を行い、ＧＴ側とＡＧ側のそれぞれの突然変異箇所における、スプライシング異常数についてまとめたものを図 3.4 と図 3.5 に示す。ここでのスプライシング異常とは、突然変異前の値から 0.2 以上変化したものを表している。

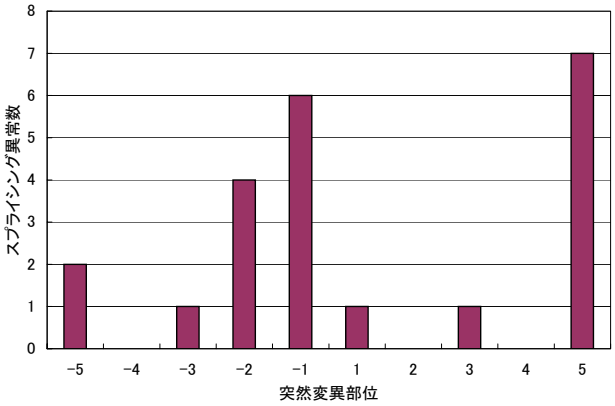

図 3.4　ＧＴ側のスプライシング異常数

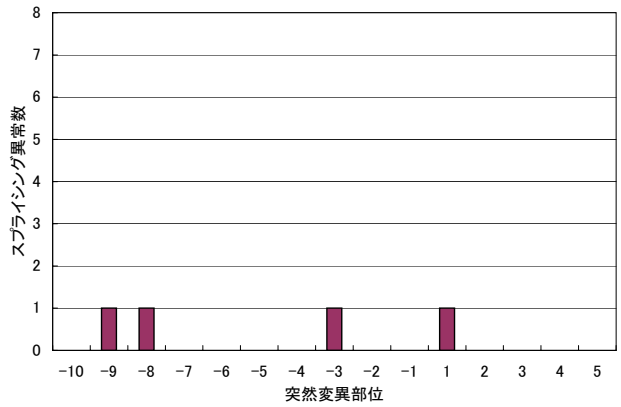

図 3.5　ＡＧ側のスプライシング異常数

このようにＧＴ側でのスプライシング異常では-1、-2、5 番目の塩基が突然変異した場合にスプライシング異常を起こす場合が多く、反対にＡＧ側では特定の部位のスプライシング異常は見られないという結果が得られた。

5　βサラセミアにおける突然変異

5.1　サラセミア

サラセミアは遺伝性貧血症であり、地中海沿岸、アフリカ、インド、中国、東南アジアに広く見られ、α鎖に異常が生じるαサラセミアと、β鎖に異常が生じるβサラセミアの 2 種類に大別される [5]。その一種であるβサラセミアは、ヘモグロビンを構成するβ鎖の合成障害をきたすβグロビン遺伝子の異状による症候群である。サラセミアはいずれも同様の症状を示すが、重症度はさまざまで、軽症型βサラセミアでは軽度の貧血だけで症状はない。しかし、重症型βサラセミアでは重度の貧血症状が生じ、黄疸、皮膚潰瘍、胆石、脾腫がみられることがある。また、骨髄の活動が過剰になることで頭部と顔面の骨が厚く大きくなるとともに、腕と脚の長骨が弱くなり骨折しやすくなる。

5.2　βサラセミアにおける突然変異

Kazazian[6] によると、βサラセミアではＧＴ側に 5 個、ＡＧ側に 2 個、その他イントロン内部に 5 個、エクソン内部に 3 個の計 15 個の突然変異点が実験的に知られていることが分かっている。これらのデータを元に、突然変異とスプライスとの関係をニューラルネットワークを用いて検証した。

まずイントロン (IVS) １のＧＴ側の突然変異について検証した結果を図 4.1 に示す。表中の Type は、これまでの研究を通してそれぞれの突然変異がスプライシング異常を示すかどうかを表しており、+ が陽性、0 が陰性、? が不明であることを示す。

入力番号	Mutant Class	Sequence		Type	Output
1	突然変異前 ⇒	GGCAG	GTTGG		0.7959
2	(−3) C→T	GGTAG	GTTGG	?	0.9164
3	(−1) G→C	GGCAC	GTTGG	?	0.0071
4	(5) G→C	GGCAG	GTTGC	+	0.0028
5	(5) G→T	GGCAG	GTTGT	+	0.0301
6	(5) G→A	GGCAG	GTTGA	+	0.1586

表 4.1　IVS-1(GT 側) での突然変異

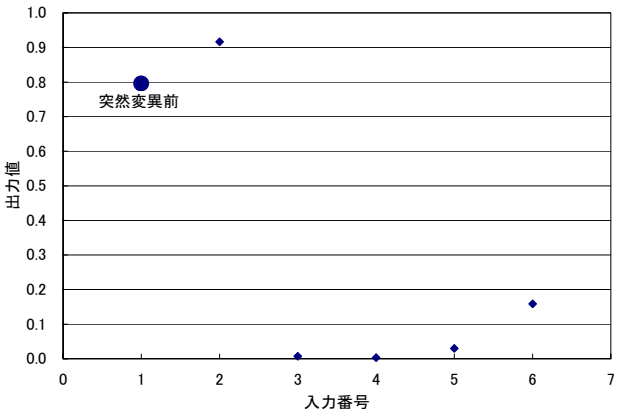

図 4.1　IVS-1(GT 側) での突然変異出力分布

入力番号 2 の突然変異では、出力値はむしろ正常配列の出力値より大きくなり、スプライシングの異常が起こらないことを示している．それに対して，入力番号 3 から 6 の突然変異ではニューラルネットワークの出力に大きな変化が見られた。いずれの変異も出力値は小さな値となり、スプライス部位として機能しなくなることを示唆している．

次に、ＡＧ側について検証した結果を図 4.2 に示す。

入力番号	Mutant Class	Sequence		Type	Output
1	突然変異前 ⇒	CCACCCTTAG	GCTGC		0.7028
2	(−3) T→G	CCACCCTGAG	GCTGC	+	0.0024
3	突然変異前 ⇒	CCTCCCACAG	CTCCT		0.9211
4	(−3) C→A	CCTCCCAAAG	CTCCT	+	0.5272

表 4.2　IVS-1,2(AG 側) での突然変異

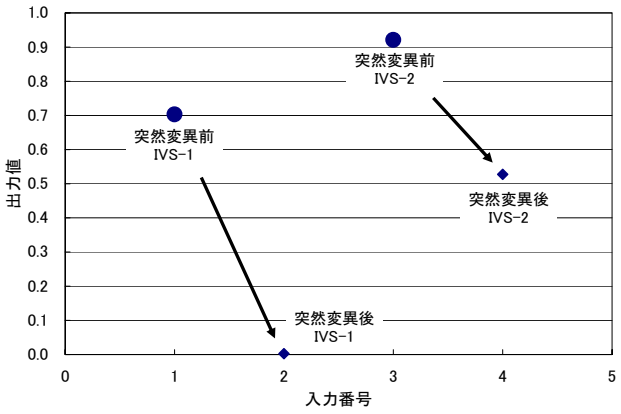

図 4.2　IVS-1,2(AG 側) での突然変異出力分布

論文 [6] によればＡＧ側では２つの突然変異が分かっている。入力番号 2 では正常配列（入力番号 1）に比較して出力値が大きく減少し、スプライシング異常を示唆している。入力番号 4 では、出力値が正常配列（入力番号 3）の半分程度に減少し、これもなんらかの異常の発生を予測している。

次にＧＴ側にもＡＧ側にも含まれない、イントロン内部およびエクソン内部の突然変異箇所についても解析を行った。このような場所でも突然変異によってＧＴやＡＧが生まれ、こうした隠れた (criptic) スプライシング部位がスプライシングに影響を及ぼす可能性がある。

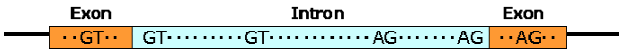

図 4.3　隠れたスプライシング部位

入力番号 1 では突然変異Ｇ→Ａにより新たにＡＧが生まれ、大きな出力値を持つことがわかる。正常配列も大きな出力値を持つが、これはＡＧ配列ではないので意味はない。この結果から、イントロン内部に新しくスプライシング部位が発生したことを示している。

入力番号 6 と 7 はエクソン内部の突然変異を表し、いずれもＧＴの出力値が増大している。この計算結果から、これらの部位ではもともと大きな出力値をもっていたが、突然変異により活性化されたことを示唆している。

入力番号	Mutant Class	Sequence		Type	Output
1	突然変異前 ⇒	TGCCTATTGG	TCTAT		0.8435
	(−2) G→A	TGCCTATTAG	TCTAT	+	0.9812
2	突然変異前 ⇒	ATTGGTCTAT	TTTCC		0.0000
	(−1) T→G	ATTGGTCTAG	TTTCC	0	0.0000
3	突然変異前 ⇒	TGTAACTGAT	GTAAG		0.0000
	(−1) T→G	TGTAACTGAG	GTAAG	0	0.0015
4	突然変異前 ⇒	TACAATCCAG	CTACC		0.0048
	(1) C→G	TACAATCCAG	GTACC	0	0.0203
5	突然変異前 ⇒	CTGGGTTAAG	GCAAT		0.0016
	(2) C→T	CTGGGTTAAG	GTAAT	0	0.0040
6	突然変異前 ⇒	TGGTG	GTGAG		0.5451
	(3) G→A	TGGTG	GTAAG	+	0.8566
7	突然変異前 ⇒	TGGTG	GTGAG		0.5451
	(−2) T→A	TGGAG	GTGAG	+	0.9112
8	突然変異前 ⇒	GTGAG	GGCCC		0.0001
	(2) G→T	GTGAG	GTCCC	0	0.0000

表 4.3　エクソン、イントロン内部の突然変異

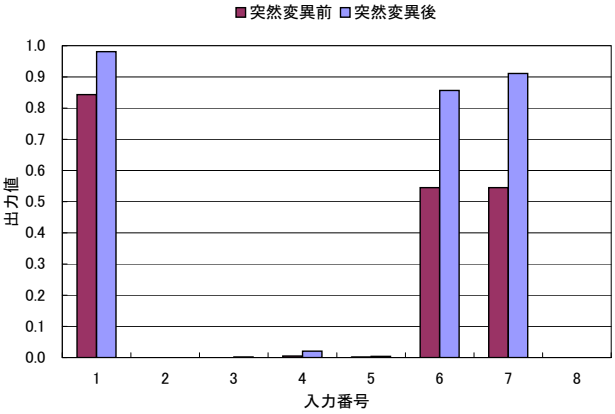

図 4.4　突然変異の前後による出力値

ここでは 8 つの突然変異パターンの中で、イントロン内部に 1 箇所、エクソン内部に 2 箇所の突然変異によるスプライシング異常が見られ、そのどれもが突然変異によってニューラルネットの出力値が上昇していることが分かった。つまり、これらの 3 箇所は突然変異することにより隠れたスプライシング部位として活性化する可能性があるということになる。

6　考察

突然変異におけるスプライシング異常については、血友病Ｂとβサラセミア共に複数のスプライシング異常が見られ、これらの遺伝疾患がスプライシング異常

によって発病に結びつく可能性を示唆した。また、βサラセミアの例のようにスプライス部位から遠く離れた場所でも隠れたスプライス部位が存在し、突然変異によって活性化される。ニューラルネットワークの学習データとして用いた非スプライス配列の中で出力の高い部分があったが、これらの中には隠れたスプライス部位が含まれている可能性があり、今後注意深く検討していく必要がある。

本研究では１つのニューラルネットワークのみを用いたが、複数のニューラルネットを作り、それらの多数決を取ることで、より精度の高いスプライシング部位判定が行えるものと考える。また、スプライシング部位の推定には、隠れマルコフ法や mRNA 前駆体の構造を利用したものなど様々な手法が提案されており[1]、今後はこれらの方法を組み合わせてより信頼できる予測法を開発していきたい。

Reference

[1] W.H. Majoros:“ Methods for Computational Gene Prediction ”,Cambridge University Press, Cambridge, 2007.

[2] 古谷博史, 他:“ ニューラルネットワークによるタンパク質をコードする遺伝子のスプライス部位の推定 ”, 医療情報学 Vol.9,163–173,1989

[3] 中野馨:“ ニューロコンピュータの基礎 ”コロナ社 1990

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

[5] 服巻保幸:“ サラセミア ”, 代謝, Vol.327, 247–252, 1988

[6] H.H.Kazazian.Jr and C.D.Boehm:“ Molecular Basis and Prenatal Diagnosis of β-Thalassemia ”, Blood Vol.72,1107–1116,1988