点突然変異によるスプライシング異常の予測

高野 伶美(指導教員:由良 敬)

1 はじめに

ゲノムとは、細胞を構成するタンパク質などの鎖状高分子の設計図が書き込まれている分子である。多くの生物において、この化学的実体は DNA である。DNA は、4 種類の塩基(アデニン(a),チミン(t),シトシン(c),グアニン(g))が組み合わされて構築される鎖状高分子で、ヒトゲノムの場合、その長さは約 30 億塩基に達する。塩基の並び順(塩基配列)や長さは生物種ごとに異なっており、塩基の並び順そのものが、タンパク質の暗号化された設計図となっている。ゲノム塩基配列決定の技術はさまざまな生物種のゲノムを「解読」するプロジェクトを通じて向上し、2003 年には、ヒトゲノムの全塩基配列が決定した。[1]

ゲノム塩基配列にはタンパク質をコードする部分(遺伝子)とタンパク質の発現を制御する部分がある。タンパク質は 20 種類のアミノ酸が組み合わされて構築された鎖状高分子である。4 種類の塩基で 20 種類のアミノ酸をコードするために、連続する 3 つの塩基を 1 組とすることで、4^3=64 種の異なるコードを構成し、20 種のアミノ酸に対応させている。そのため、組み合わせは違っても同じアミノ酸をコードすることもある。

ヒトなどの高等真核生物では、遺伝子は、最終的にアミノ酸配列に翻訳される領域（エクソン）と、アミノ酸配列に翻訳されない領域（イントロン）のまだら構造となっており、イントロンは翻訳過程の直前で切り出され、イントロンを挟むエクソンが結合する（スプライシング）。このスプライシングの仕方は、ひとつの遺伝子に対して複数あることがわかっている。あるスプライシングではエクソンであった領域が、他のスプライシングにおいてはイントロンとして用いられることがあり、その結果ひとつの遺伝子から複数種類のタンパク質を作り出すことができる。こうした機構によって、体内の組織ごとに異なるタンパク質を作り出すことができ、遺伝子数よりもはるかに多くの種類のタンパク質がコードできるようになっている。このスプライシングは、遺伝子内の塩基配列で制御されていると考えられており、何らかの事情で塩基が変化すると、スプライシングに変化がおこると考えられている。この変化は、タンパク質にも変化を及ぼし、タンパク質の機能に影響をもたらし、疾患にいたらしめる場合がある[2]。

塩基配列の変異が原因となる疾患として知られるものの中に、アルツハイマー病がある。アルツハイマー病は、認知症と診断される疾患の中で、最も症例が多いとされており、神経細胞間の情報伝達が阻害されたり神経が死滅していくことで、記憶力や思考能力が低下し、最終的には日常生活における単純な作業を行う能力さえも失われていく疾患である。

アルツハイマー病には家族性と孤発性とがあり、家族性は全体の約 1%である。家族性アルツハイマー病の原因となるゲノム変異があった場合、発症率はほぼ 100%である。このような、発症を決定付ける変異をもつ遺伝子を原因遺伝子とよぶ。全体の約 99%を占める孤発性アルツハイマー病には、原因遺伝子のように単体で発症を左右する遺伝子は無く、発症の確率を高める遺伝子(危険遺伝子)が複数存在する[3]。

今までに報告された危険遺伝子は、ALZgene データベース[4]に登録されている。ここに登録されているゲノム変異の中には、遺伝子として読み出されない領域やイントロンなど、アミノ酸をコードしない領域での変異も多く見られた。遺伝子外の配列に変異があれば遺伝子の発現制御に変化がおき、エクソンに変化があれば、アミノ酸配列が変わる可能性がある。イントロンに変異があればスプライシングに変化がおきる可能性がある。もしスプライシングに異常が起きれば、アミノ酸配列の長さが変わり、合成されるタンパク質に大きな影響を与えることが考えられる。そこで本研究では、イントロンでの変異に焦点をあて、実際にアルツハイマー病患者に見られたイントロンでの一塩基置換がスプライシング異常を起こすのか否かを予測する手法を開発し、実際のデータに適用した。

2 手法

2.1 翻訳領域既知遺伝子からのスプライシング情報の取得

塩基配列データベース GenBank[5]から、霊長類のゲノムのうち、遺伝子領域とそのスプライシング箇所が分かっているもののデータを取得した。結果、122,222,256 塩基からなる 4,240 本の遺伝子データを得た。

2.2 隠れマルコフモデル(HMM)の構築

2.1 で取得したデータから、エクソン・イントロンの属性ごとの各塩基数と、状態の遷移数を、属性も区別して集計した。この結果を表 1、表 2 に示す。

表 2 から分かるように、属性の遷移が同じ遷移でも、塩基の組み合わせによって、出現数に偏りがあった。例えば、同じエクソンからイントロンの変異でも、g から c の変異が 85 回しか見られなかったのに対し、g から g への変異は 10,043 回見られている。これは、スプライシングされる点における塩基の存在確率が、イントロンにおいてもエクソンにおいても偏りがあるためである。イントロンとエクソンの境界には、多くの遺伝子において一致する配列があり、これをコンセンサス配列とよぶ。このコンセンサス配列は、エクソン側に 3 塩基、イントロン側に 6 塩基ほどあるとされているが、中でも高度に保存されているのが、エクソンからイントロンに変わる点でのエクソンの ag という並びと、イントロンの gt という並びである。つまり、配列中に aggt という並びが現れた時、2 つの g の間でスプライシングが起こることが極めて多いとされている。また、この塩基配列は左から読んでいくが、読み始める側の端を 5'末端、読み終わる方の端を 3'末端とよぶ。(図 3)

本研究で対象とするデータは、塩基置換の起こった遺伝子の塩基配列であり、予測したいものは配列中での各塩基の属性である。塩基配列は連続した文字列であり、その属性を分類する、という問題が形態素解析に似ていると感じたことから、形態素解析にも用いられている隠れマルコフモデルを用いることにした。コンセンサス配列の存在もあり、スプライシング点では塩基の存在確率の偏りが大きかった。分類したい状態数は、エクソンとイントロンの 2 つであるが、この 2 状態における塩基の存在確率と状態間の遷移確率を設定するだけでは、スプライシング点のみに顕著に現れる偏りを表現できないため、図 4 のようにモデルを構築した。状態数は、エクソンとイントロンからそれぞれ 3'と 5'を分離させた 6 状態である。それぞれの状態における 4 塩基の出現頻度を存在確率、状態間で遷移が起こる確率を遷移確率とよぶ。

まず、各状態での存在確率を考えた。塩基ごとに状態間の比を出し、それを状態内での存在確率の和が 1 になるように調節することで、塩基配列全体における各塩基の出現

数の差をならした。次に、遷移確率を考えた。各状態間での遷移数を全遷移数で割った頻度を遷移確率としたものを遷移確率 A とし、遷移確率 B の下でエクソン・イントロンの連続する期待値がそれぞれの状態の平均長と等しくなるように B を定めた。A、B それぞれの確率についてビタビアルゴリズム[6]を用い、2.1 の遺伝子配列全てについて状態遷移を予測し、実際のデータとの差を調べた。その結果、A と B の予測結果には違いが出たため、A、B の E から E への遷移確率の差、約 $4 \cdot 10^{-3}$ と、I から I への遷移確率の差、約 $5 \cdot 10^{-4}$ により予測に差が生じることが分かった。そこで、E から E の遷移の確率を、A、B での確率を内包する 0.9900 から 0.9990 の範囲で、刻み幅 10^{-3} で動かし、同時に I から I らの遷移を 0.9990 から 0.9999 の範囲で、刻み幅 10^{-4} で動かしながら予測を行い、実際の構造との一致度が最高となった確率の組み合わせを配列ごとに集計した。この内、一致度が 80%以上であった配列における確率の組み合わせの中で最も多かったものを遷移確率 C とする。

2.3 ALZgene からのデータへの適用

ALZgene では、変異のデータごとに、健常者群と患者群において何人に見られた変異か、また、正常な遺伝子との出現割合はどれくらいか、というデータから、その変異と発病の関連性は統計的に有意かを予測してある。そこで、統計的に有意だとされているもののうち、イントロンに存在する一塩基置換のデータを検索し、その変異を内包する領域既知の遺伝子を調べた。結果、39 本の遺伝子についてのデータが得られた。

	a	g	c	t
exon	2276145	2146319	2112003	2228089
intron	31562114	24103229	23001151	34793106

表 1: 取得したデータに含まれていた各塩基数とその属性

b＼a	ea	et	ec	eg
ea	667089	388938	632695	579033
et	510513	649693	629786	433270
ec	447037	507935	619346	532524
eg	646411	677276	227457	582726
ia	612	458	525	438
it	540	714	521	406
ic	380	426	445	344
ig	2405	1985	686	16631
b＼a	ia	it	ic	ig
ea	645	502	540	5777
et	515	760	527	2686
ec	323	413	410	3177
eg	527	610	85	10043
ia	9912873	7323212	7704547	6619390
it	8409746	12018616	8113654	6248886
ic	5292782	6736894	5977410	4992442
ig	7944647	8712086	1203960	6220804

表 2: 取得したデータに含まれていた遷移数(縦軸:前状態, 横軸:後状態)

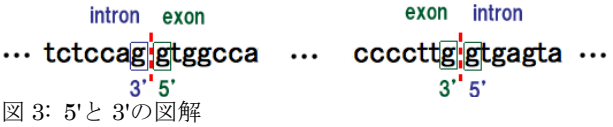

図 3: 5'と 3'の図解

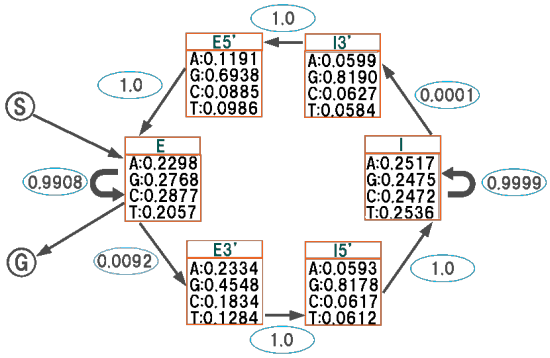

図 4: 隠れマルコフモデルの状態遷移図

3 結果

遷移確率 A、B、C によって求められた状態遷移は、全塩基 122,222,156 に対して、A での状態遷移と実データとの差は 12,955,075、B は 12,055,115、C は 11,266,756 であった。結果に有為な差が見られなかったため、全て用いて 2.3 を行った。

結果、一塩基変異により予測が変わった配列はなかった。

4 考察

一塩基ずつ見るマルコフモデルでは、連続した 2 塩基の関係しか調べることができない。これにより、スプライシング点から離れたところに現れる変異のスプライシングに与える影響は考えなかったために予測ができなかったと考えられる。

5 今後の課題

イントロンのみのデータやエクソンのみのデータである程度の長さがある時、予測の精度が落ちることが分かったので、配列によって出る精度の差を無くしていくことを今後の課題にしたい。

さらに、コンセンサス配列に依存しすぎないことで、既存手法が苦手とすることの多い、コンセンサス配列がないスプライシング部位の推定を可能にしたいと考えたが、2 塩基間の関係を見るだけでは予測は難しいことが分かったため、複数塩基間における関係性を考慮する方法を考えていきたい。

また、今回はビタビアルゴリズムによって求めた、最も尤度の高い配列を予測結果として採用した。しかし、前述したように、スプライシングの仕方はひとつの遺伝子に対して複数通りある。今後は尤度の下がる経路についても、どこまで採用するかを含めて考えていこうと思う。

参考文献

[1] DDBJ, http://www.ddbj.nig.ac.jp/whatsnew/wn030423-e.html
[2] D.Sadava(Eds.), 大学生物学の教科書 第２巻 分子遺伝学 石崎泰樹・丸山敬(訳), ブルーバックス(2010), pp.267-278
[3] C.Medway and K.Morgan, Review: The genetics of Alzheimer's disease; putting flesh on the bones, Neuropathology and Applied Neurobiology(2014), 40, 97-105
[4] ALZgene, http://www.alzgene.org/
[5] GenBank, http://www.ncbi.nlm.nih.gov/genbank/
[6] Robert E. McAuliffe, Durbin–Watson Statistic, 2015