PRESS RELEASE

東京大学医科学研究所
The Institute of Medical Science, The University of Tokyo

2026 年 6 月 23 日
東京大学医科学研究所

RNA スプライシング予測のための深層学習モデル「SpliceSelectNet」の開発

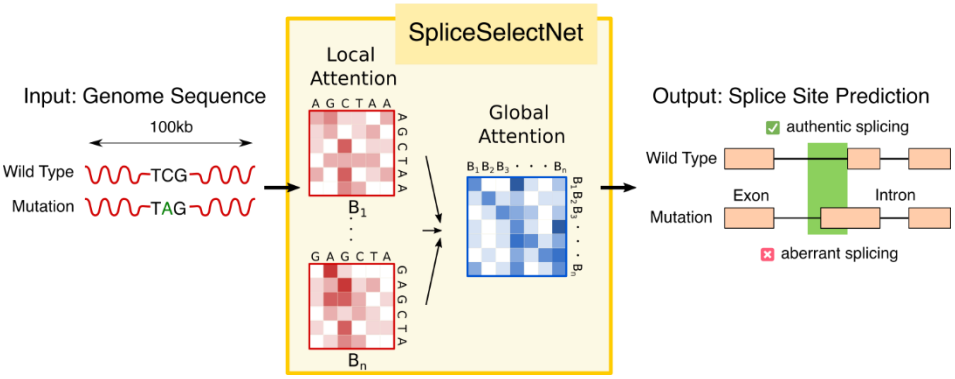
――最大 10 万塩基対の長距離ゲノム配列を考慮したスプライス部位および異常スプライシングの予測――

発表のポイント

◆最大 10 万塩基対（100kb）の長大な DNA 配列からスプライス部位を予測し、疾患に関わる異常スプライシングを検出できる階層型 Transformer ベースの深層学習モデル「SpliceSelectNet（SSNet）」を開発しました。

◆局所的および大域的なアテンション機構を統合した構造により、従来モデルでは計算コストの面で困難であった 1 塩基単位の解像度を保ちながら遠隔の制御領域の影響を効率よく捉えることに成功しました。また、DNA 配列モデルとしてアテンションスコアを可視化することで予測根拠を提示し、生物学的な解釈を可能にしました。

◆がんや遺伝性疾患の原因となる異常スプライシングの発生予測に貢献することが期待されます。また、本モデルの予測技術は、病原性変異の評価や個別化医療・ゲノム創薬への応用が期待されます。



深層学習モデル「SpliceSelectNet」の概要

概要

東京大学医科学研究所機能解析イン・シリコ分野の中井謙太教授と同大学大学院情報理工学系研究科大学院生の宮地佑奈は、深層学習を用いた RNA スプライシング予測において、最大 10 万塩基対にも及ぶ長距離のゲノム配列の相互作用を考慮に入れた、新しい予測モデル「SpliceSelectNet」を構築しました。

従来のスプライシング予測モデルは高い予測精度を示すものの、ヒト遺伝子のイントロンは数キロ塩基長を超えることも珍しくないため、長距離に及ぶ制御配列の影響を効率的に捉えることや、AI の予測根拠を生物学的に解釈することが困難でした。本研究では、階層型 Transformer（注 1）と呼ばれるアーキテクチャを導入することでこれらの課題に取り組みました。その結果、計算効率を保ちながら長距離の相互作用をモデル化し、予測の根拠となる機能的な配列重要度の可視化を実現しました。この成果は、遠隔の変異が引き起こす異常スプライシングの特定や、RNA 制御メカニズムの理解、ひいてはゲノム医療の発展に寄与することが期待されます。

本研究成果は、2026 年 6 月 22 日付で、国際学術誌「*Nucleic Acids Research*」オンライン版で公開されました。

発表内容

人間の遺伝子からタンパク質が作られる過程において、RNA スプライシングは生命維持に不可欠なステップです。この過程に異常が生じると、がんや筋ジストロフィーなどの重篤な疾患を引き起こし得ることが知られています。近年、SpliceAI をはじめとする深層学習を用いたスプライス部位（注 2）の予測モデルが登場しましたが、従来の手法には限界がありました。スプライシングは、スプライス部位の周辺だけでなく、数万塩基対も離れた遠隔の制御領域による影響を受けますが、AI にこのような「長距離依存性」を学習させようとすると計算コストやノイズが膨大になります。また、深層学習特有の「ブラックボックス化」により、AI がなぜその予測を出したのかという生物学的な根拠を解釈することが困難でした。

本研究チームは、画像認識等で用いられる階層的なネットワーク構造に着想を得て、DNA 配列モデルに局所的および大域的アテンション機構（注 3）を統合的に導入した「SpliceSelectNet」を開発しました。この構造により、計算効率を損なうことなく、モデルの受容野を従来の 5,000 塩基長から最大 10 万塩基長にまで拡張することに成功しました（図 1）。

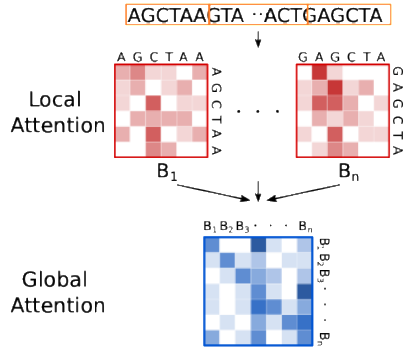


図 1：SpliceSelectNet の階層型アテンション機構

入力された長大な DNA 配列を局所的・大域的なブロックに分けて処理することで、計算量を抑えつつ遠隔の配列領域との相互作用を効率的に学習するモデルの構造。

本モデルが持つ、遠く離れた配列がスプライシングに与える影響を捉える能力を検証するため、非常に長いイントロンを持つデュシェンヌ型筋ジストロフィー（DMD）の原因遺伝子を対象とした実験を行いました。具体的には、コンピュータ上で人工的に変異を導入する in silico（注 4）シミュレーションとして、カノニカル（注 5）なドナー部位から最大 1 万塩基離れた位置まで、段階的に人工的なおとり配列（注 6）を挿入し、ドナー部位のスプライス予測確率がどのように変化するかを評価しました。

その結果、従来の主要な CNN（注 7）ベースのモデルでは、おとり配列が約 4,400 塩基以上離れると感度を失って影響を全く捉えられなくなり、予測値が変化しなくなるのに対し、SpliceSelectNet は 8,000 塩基や 1 万塩基といった超遠隔領域におとり配列を置いた場合でも、その効果を的確に捉えて予測値が変化することが示されました。これにより、従来のツールでは見落とされてきた、カノニカルなスプライス部位から遥か遠くに位置するイントロン深部の変異の病原性を評価できる、新たな道が拓かれました（図 2）。

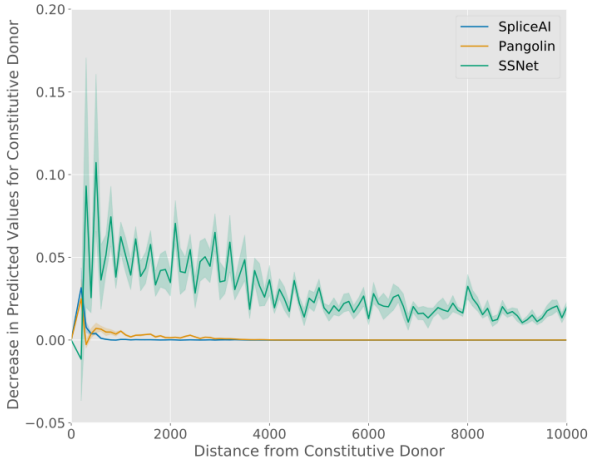


図 2：ジストロフィン遺伝子における長距離スプライシング制御の予測

ジストロフィン遺伝子に対しコンピュータ上で人工的な変異を導入したときの実験結果。従来モデルでは計算範囲の限界で捉えられなかった、数十 kb 離れた「超遠隔領域」の変異がスプライシングに与える影響を、SpliceSelectNet が的確に捉えられていることを示している。

さらに、本モデルの大きな特徴は、AI が DNA 配列の「どこ」に注目して予測を行ったか（アテンションスコア）を可視化できる点です。AI の注目箇所が実際の生物学的な機能部位と一致するかを検証するため、スプライシングを促進する機能を持つ ESE（注 8）を対象に実験を行いました。マウスの免疫グロブリン M（IgM）遺伝子を用いて、コンピュータ上で ESE の配列を別の塩基（N）に置き換えてマスキングする実験を行ったところ、スプライス部位の予測確率が大きく減少しました。さらに、アテンションスコアを可視化すると、ESE 領域に対してモデルが強い「注目」を向けていたことが確認されました。これにより、提案モデルが生物学的に重要な制御領域を予測の根拠として捉えていることが示されました（図 3）。

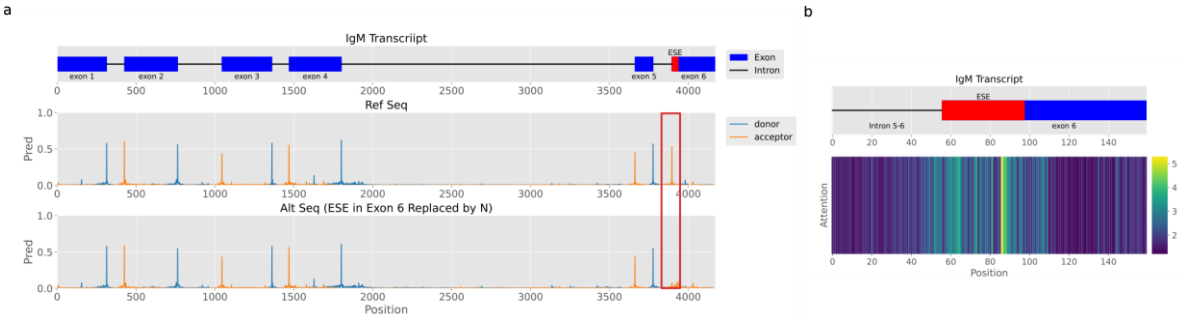


図 3：IgM 遺伝子における AI の予測結果とアテンションスコア

SpliceSelectNet による IgM 遺伝子のスプライシング予測結果（a）と、その際にモデルが注目した配列を示すアテンションスコア（b）。注目箇所が、生物学的に重要な制御領域（ESE）と一致していることがわかる。

本研究によって、これまで見逃されていた遠隔の遺伝子変異がスプライシングに与える影響を高精度に評価できるようになりました。今後は、未解明の病原性変異の特定や、異常スプライシングを標的とした核酸医薬品の開発などへの波及効果が期待されるとともに、DNA 言語モデルや RNA 制御メカニズム研究の基盤技術としての活用が見込まれます。

発表者・研究者等情報

東京大学
医科学研究所　附属ヒトゲノム解析センター　機能解析イン・シリコ分野
中井　謙太　教授
兼：同大学大学院情報理工学系研究科　コンピュータ科学専攻

大学院情報理工学系研究科　コンピュータ科学専攻
宮地　佑奈　博士課程

論文情報

雑誌名：Nucleic Acids Research
題　名：SpliceSelectNet: A Hierarchical Transformer-Based Deep Learning Model for Splice Site Prediction
著者名：Yuna Miyachi,Kenta Nakai*（*責任著者）
DOI：10.1093/nar/gkag625
URL：https://academic.oup.com/nar/article/54/12/gkag625/8713015

研究助成

本研究は、JST SPRING「JPMJSP2108」の支援により実施されました。

用語解説

（注 1）　Transformer
自然言語処理などの分野で革命をもたらした深層学習のアーキテクチャ。文章中の離れた単語同士の関係性を捉えるのが得意であり、近年では DNA 配列などのゲノムデータの解析にも応用されている。

（注 2）　スプライス部位
遺伝子（DNA）から転写された「前駆体 mRNA」において、タンパク質の合成に必要な部分（エキソン）を繋ぎ合わせ、不要な部分（イントロン）を切り落とす反応（RNA スプライシング）が起こる境界の配列のこと。

（注 3）　アテンション機構
AI が予測を行う際に、入力データ（本研究では DNA 配列）の中で「どの部分に注目（アテンション）すべきか」を学習・重み付けする仕組み。

（注 4）in silico
「コンピュータ（半導体チップの主成分であるシリコン）の中で」という意味の学術用語。生体内（in vivo）や試験管内（in vitro）での実験に対比して使われ、コンピュータシミュレーションを用いた解析手法を指す。

（注 5）カノニカル
「標準的な」「典型的な」という意味の用語。スプライス部位の中でも、ほとんどの遺伝子で共通して使われている最も基本的かつ普遍的な塩基配列（イントロンの始まりの「GT」、終わりの「AG」など）を指す。

（注 6）おとり配列
本物のスプライス部位とよく似た構造を持ち、スプライシングに不可欠な因子を奪い合うような競争関係を作り出すために人工的に設計された偽物の配列。

（注 7）CNN（Convolutional Neural Network）
画像認識の分野で広く使われてきた畳み込みニューラルネットワークという機械学習の手法のこと。データの局所的なパターンを見つけるのが得意である。

（注 8） ESE（Exonic Splicing Enhancer）
エキソン内に存在し、スプライシング反応を促進するように働く特定の塩基配列（エキソン性スプライシングエンハンサー）のこと。この領域に変異が生じると、正常なスプライシングが阻害されることが多い。

問合せ先

<研究内容について>
東京大学医科学研究所 附属ヒトゲノム解析センター　機能解析イン・シリコ分野
教授　中井　謙太（なかい　けんた）
https://www.ims.u-tokyo.ac.jp/imsut/jp/lab/hgclink/section06.html

<機関窓口>
東京大学医科学研究所 プロジェクトコーディネーター室（広報）
https://www.ims.u-tokyo.ac.jp/