様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１５　日現在

機関番号：　１３３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３０３２８

研究課題名（和文）ディープラーニングを用いた大規模配列データからの階層的特徴抽出

研究課題名（英文）Hierarchical Feature Extraction from Large Sequence Data by Deep Learning

研究代表者
佐藤　賢二（SATOU, Kenji）

金沢大学・電子情報学系・教授

研究者番号：１０２１５７８３

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：塩基配列や、それから転写・翻訳されてできるアミノ酸配列を解析する際、分子生物学の知識に大きく依存した従来的な方法では、ある程度予想可能な事実しか発見できないという限界があった。本研究ではディープラーニングを含む各種の機械学習手法を用いることで、大規模な配列データから道の階層的な特徴を抽出することが可能であることを明らかにした。

研究成果の概要（英文）：In the field of biological sequence analysis including DNA and amino acid sequence analysis, traditional methods are highly dependent on the knowledge specific to molecular biology, so their ability is limited to the discovery of features easily predicted from domain-specific knowledge. In this study, it is shown that by using various machine learning algorithms including deep learning, it is possible to extract novel and hierarchical features from large sequence data.

研究分野：　生命情報学

キーワード：　テキスト分類　畳み込みニューラルネットワーク　スプライス部位　プロモータ　単語埋め込み　次世代シーケンサ　ゲノム配列決定

１．研究開始当初の背景

1990 年代におけるヒトゲノム計画の遂行に伴い、DNA や RNA の塩基配列決定技術は大きく進歩したが、2008 年に次世代シーケンサ（Next Generation Sequencer: NGS）が登場したことにより、塩基配列決定の速度とコストは、ムーアの法則を超えるほどの劇的な改善を見せた。30 億塩基対のヒトゲノムを決定するのに１日程度の時間と実質 10 万円程度の費用しかかからない現代では、各種生物の配列データも爆発的な速度で増えており、生命科学におけるビッグデータ解析の最有力候補となっている。

塩基配列や、それから転写・翻訳されてできるアミノ酸配列を解析する際、従来は分子生物学の知識に大きく依存した方法が取られていた。例えば、原核生物や真核生物の遺伝子構造に基づいて長大なゲノム配列から遺伝子領域を予測したり、各種生物から採取した相同タンパク質のアミノ酸配列を比較することにより、そのタンパク質における進化的な保存領域（モチーフ）を抽出し、タンパク質の機能と関連づけることが行われた。しかしながら、このような領域知識に基づいた解析方法には、生物学的にある程度予想可能な事実は発見できても、これまでの常識にとらわれない全く新しい事実は発見できない、という限界がある。

一方、機械学習の分野では近年、ニューラルネットワークが進化したことにより、ディープラーニング（深層学習）という一連のアルゴリズムが注目され、数十年に一度とも言われる大成功を収めている。画像認識や薬剤開発のコンペティションで他を圧倒する高性能を記録し、今や Microsoft や Apple、Google などの巨大企業が画像や音声の認識にディープラーニングをこぞって採用していることからも、この技術があらゆる分野に革新をもたらしつつあることが分かる。

ディープラーニングには、大きく分けて２つの側面がある。１つは、従来法よりも高性能な予測アルゴリズムとしてのディープラーニングである。多層ニューラルネットワークをベースに、事前学習やドロップアウト、プーリング、畳み込みなどのテクニックを組み合わせることにより、ディープラーニングは、これまで主流だったサポートベクターマシン等を大幅に上回る予測精度を達成することができた。例えば、1 万種類以上のカテゴリに分類された 1000 万枚の画像を学習して新規画像のカテゴリを予測する ImageNet Large Scale Visual Recognition Challenge 2012 では、ディープラーニングを用いたトロント大学の Hinton らのグループが、２位のチームよりも 40%ほどエラー率を低減し、大差で優勝している。同様に、新規薬剤開発のコンペティションである Merck Molecular Activity Challenge でも、Hinton らのグループが優勝している。

もう一つの重要な側面は、データの特徴抽出器としてのディープラーニングである。ディープラーニングの事前学習では、Autoencoder や Restricted Bolzmann Machine を多層で組み合わせることにより、入力データに最も近い層では具体的だが断片的な特徴を抽出し、層が深くなるにつれて抽象的だが全体的な特徴を抽出できることが分かっている。有名な例としては、Google の研究者による画像認識実験の結果、特に指定しないにも関わらず、猫の顔や人の顔を表現した抽象的かつ全体的な特徴を抽出できたという報告がある。

２．研究の目的

本研究の目的は、ディープラーニングを応用することにより、塩基配列やアミノ酸配列から全く新しい階層的な特徴を抽出することにある。申請者はこれまで、サポートベクターマシン等を用いて塩基配列やアミノ酸配列を高精度に分類する研究を続けており、その過程で領域知識や確率統計に基づく様々な特徴抽出法を試してきた。このような準備状況と、上述した配列ビッグデータの爆発的な増大、およびディープラーニングアルゴリズムの登場により、本研究の主要なアイデアを着想するに至った。

３．研究の方法

本研究ではまず、ディープラーニングを用いて DNA 配列やアミノ酸配列の分類と予測を行っている事例について、文献調査を行った。位置特異的スコア行列（PSSM）など、従来的な配列分類予測で使われている特徴を多数組み合わせた例がいくつか報告されているが、ディープラーニングに特化した全く新しいエンコーディング方法を提案した例は見られなかった。次に、数種類の配列分類問題に対してディープラーニングを応用し、サポートベクターマシン等の分類器との性能比較を行った。一方、大規模配列データへの対応としては、グラフィックプロセッサ（GPU）を使った高速化の効果を確認した。近年ではディープラーニング用ソフトウェアの多くが GPU を用いた高速化に対応しているが、実験ではトロント大学で開発された DeepNet を用いた。

次に、DNA 配列を直接エンコーディングして深層学習に入力することを試みた。多くの試行錯誤を経て、深層学習の一種である畳み込みニューラルネットワーク（CNN）をテキスト分類に用いた Johnson らの研究に着目し、これを DNA 配列分類に応用した。one-hot vector を用いて配列を２値のマトリクス（バイナリイメージ）としてエンコーディングすることにより、連続性の情報を損なうことなく文字列を数値ベクトルに変換し、深層学習への入力とした。学習アルゴリズムとしては

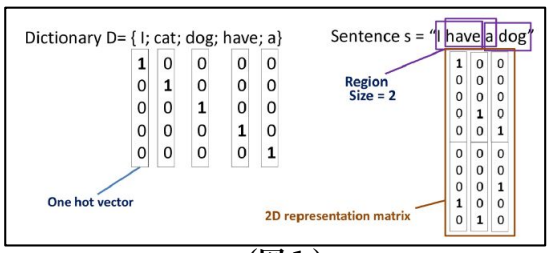
畳み込みニューラルネットワークを用い、最適なハイパーパラメータの組み合わせを探索し、交差検定により予測精度を求めた。一方、エンコーディング手法探索の過程で、ニューラルネットワークを用いた単語埋め込み（Word Embedding）も検討した。

最後に、全く新しいDNA配列に対する提案手法の特徴抽出能力を評価するために、能登地方に自生するキノコの一種を対象として、次世代シーケンサによる配列決定を行った。対象生物の発生時期である１０月以降にサンプル採取を行い、予備調査として、２種類の下部について長さ１００塩基程度のショートリードを読み取り、ゲノム配列の構築を試みた。さらに、ロングリードの配列決定およびRNA-seqによる遺伝子領域決定を行い、これらを総合してドラフトゲノム配列を構築した。並行して、大規模配列データからの新たな特徴抽出法について研究を行った。

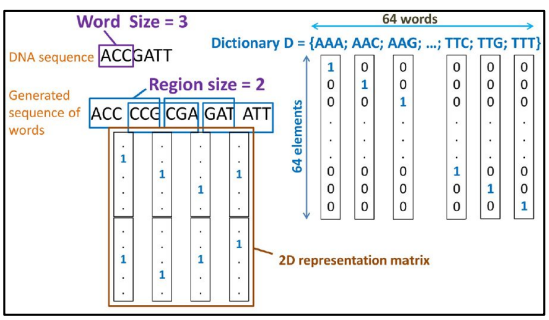
４．研究成果

数種類の配列分類問題に対してディープラーニングを応用し、サポートベクターマシン等の分類器との性能比較を行った。特徴として数種類の固定長部分配列頻度を用いた結果、エクソン・イントロンの境界予測問題とヒストンアセチル化の予測問題については、サポートベクターマシンと同等程度の分類精度を得ることができた。また、ディープラーニングの有力な要素技術の１つであるドロップアウトについても網羅的な調査を行い、その効果を確認した。ドロップアウトにより必ずしも精度が向上するわけではなかったが、その場合でも学習の収束速度が向上することが確認された。

次に、深層学習の一種である畳み込みニューラルネットワーク（CNN）を用いてDNA配列を行った。DNA配列を一種のテキストと見做すことにより、図１および図２のようにone-hot vectorを用いてDNA配列を直接エンコーディングし、CNNに入力した。



（図１）



（図２）

予測精度を検証した結果、１０種類のヒストン修飾データセットと、UCI Machine Learning Repositoryで公開されている２種類のデータセット（SpliceおよびPromoter）について、従来法を上回る精度を達成することができた（表１）。また。学習済みのネットワークの各階層から、特徴情報を抽出することができた。

（表１）

データセット	従来法の最高予測精度	本手法の平均予測精度
H3	86.47	88.99
H4	87.32	88.09
H3K9ac	75.08	78.84
H3K14ac	73.28	78.09
H4ac	72.06	77.4
H3K4me1	69.71	74.2
H3K4me2	68.97	71.5
H3K4me3	68.57	74.69
H3K36me3	75.19	79.26
H3K79me3	80.58	83
Splice	94.7	96.18
Promoter	96.23	99.06

一方、エンコーディング手法探索の過程で、ニューラルネットワークを用いた単語埋め込み（Word Embedding）も検討した結果、テキストマイニングを用いたドラッグリポジショニングが可能であることを示した。

最後に、能登地方に自生するキノコの一種を対象として、次世代シーケンサによる配列決定を行った。ショートリードからのゲノム配列構築では、リード長が短いことから多くのゲノム領域で欠損が見られたが、不完全ながらも数千個の遺伝子を同定することができた。また、同様のショートリードが公開されている近縁種のキノコと比較を行った。さらに、ロングリードの配列決定およびRNA-seqによる遺伝子領域決定を行い、これらを総合してドラフトゲノム配列を構築した。並行して、大規模配列データからの新たな特徴抽出法について研究を行った結果、数量的特徴とカテゴリカルな特徴を組み合わせることで配列分類の予測精度を改善できることを明らかにした。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① Ngo,D.L., Yamamoto, N., Tran,V.A., Nguyen,N.G., Phan,D., Lumbanraja,F.R., Kubo,M., Satou,K., (Application of Word Embedding to Drug Repositioning), Journal of Biomedical Science and Engineering, 査読有, Vol.9, No.1, 2016, 7-16

② Nguyen,N.G., Tran,V.A., Ngo,D.L., Phan,D., Lumbanraja,F.R., Faisal,M.R., Abapihi,B., Kubo,M., Satou,K., (DNA Sequence Classification by Convolutional Neural Network), Journal of Biomedical Science and Engineering, 査読有, Vol.9, No.5, 2016, 280-286

③ Phan,D., Nguyen,N.G., Lumbanraja,F.R., Faisal,M.R., Abapihi,B., Purnama,B., Delimayanti,M.K., Kubo,M., Satou,K., (Combined Use of k-Mer Numerical Features and Position-Specific Categorical Features in Fixed-Length DNA Sequence Classification), Journal of Biomedical Science and Engineering, 査読有, Vol.10, No.8, 2017, 390-401

〔学会発表〕（計０件）

６．研究組織

(1)研究代表者

佐藤　賢二（SATOU KENJI）
金沢大学・電子情報学系・教授
研究者番号：10215783