TSUBAME 共同利用　平成 29 年度　学術利用　成果報告書

利用課題名　深層学習によるゲノム上の遺伝子発現制御情報抽出
英文：Extracting regulatory code on genome by deep learning

木立　尚孝
Hisanori Kiryu

東京大学　新領域創成科学研究科　メディカル情報生命専攻
Dept. of Computational Biology and Medical Sciences, GSFS, Univ. of Tokyo
https://sites.google.com/edu.k.u-tokyo.ac.jp/kiryulab

邦文抄録（300 字程度）

細胞中の遺伝子の発現量は主にシスエレメントと呼ばれるゲノム上の DNA 配列によって制御されている。本研究では、 ゲノム中のシスエレメントを精度よく検出するために、遺伝子周辺のシスエレメントを含むと予想される DNA 配列を入力としその遺伝子の発現パターンを出力とする畳み込みニューラルネットワーク(CNN)モデルを作成した。本手法を線虫（*C. elegans*）の遺伝子発現量データ（single cell RNA-seq）に対して適用したところ、CNN のパラメタの一部である畳み込みフィルタに転写因子の結合モチーフを含む、シスエレメントと予想される配列特徴が学習された。

英文抄録（100 words 程度）

Gene expression in a cell is controlled mainly by DNA sequences on genome called cis-regulatory elements. In this study, we developed a convolutional neural network (CNN) model to predict an expression pattern of a gene from DNA sequence around the gene, which may harbor cis-regulatory elements. We have applied our model to a gene expression data (single cell RNA-seq) of worm (*C. elegans*). The CNN model learned sequence features including transcriptional factor binding motifs, which may be cis-regulatory elements.

Keywords: deep learning, convolutional neural network, single cell RNA-seq, cis-regulatory elements, worm

背景と目的

細胞中の遺伝子の発現量は、主にシスエレメントと呼ばれるゲノム上の DNA 配列によって制御されている。このシスエレメントを遺伝子の発現量データを用いて検出する試みがなされてきた。また、近年、画像認識等の分野で成果を上げた畳み込みニューラルネットワーク（CNN）が生物の配列解析にも用いられるようになってきた。CNN は予測を行う上で有用な特徴を入力の中から自動的に抽出するという特徴を持っている。

本研究では、ゲノム中のシスエレメントを精度よく検出するために、遺伝子周辺の DNA 配列を入力とし、その遺伝子がどの細胞で発現するかという発現パターンを出力とする CNN モデルを作成した。発現パターンを予測するように学習を行った CNN モデルは、入力の DNA 配列中からシスエレメントを含む発現パターンの予測に有用な特徴を抽出していると期待出来る。

概要

線虫（*C. elegans*）の遺伝子の転写開始点周辺の DNA 配列を入力とし、その遺伝子の発現パターン（single cell RNA-seq データより作成）を出力とする CNN モデルを作成した。線虫の遺伝子のうち、11865 遺伝子を訓練データ、2966 遺伝子をテストデータとし、訓練データに対し予測誤差を最小化するように学習を行った 。学習後のモデルの畳み込み層において認識されている配列特徴を計算した。

結果および考察

上記のデータを用いて学習を行い、テストデータに対する予測精度を評価した結果、CNN モデルが入力配列から発現パターン予測に有用な特徴を抽出していることが示唆された。学習後のモデルの畳み込み層において認識されている配列特徴を計算したところ、一部が既知の転写因子結合モチーフなどのシ

スエレメントと一致した（図 1）。

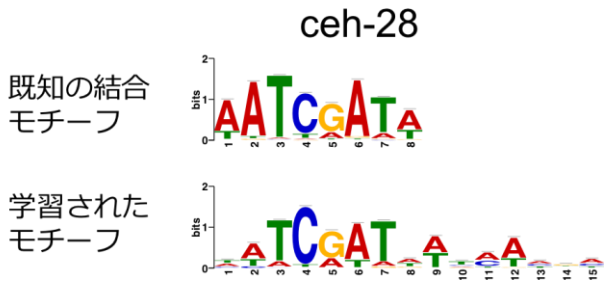

図 1. 転写因子 ceh-28 の既知結合モチーフ（上段）と CNN モデル中の畳み込みフィルタに学習されたモチーフ（下段）。

まとめ、今後の課題

遺伝子周辺 DNA 配列から遺伝子発現パターンを予測する CNN モデルを作成した。このモデルは発現パターン予測に有用な配列特徴の一部を抽出したことが確認された。今後の課題としては、この CNN モデルを用いた新規シスエレメントの発見等が挙げられる。