ディープラーニング手法を用いた一細胞エンハンサー検出法の開発

村上 賢
大阪大学 蛋白質研究所

1. はじめに

ゲノムには約２万種類の蛋白質を記録する遺伝子領域と、その遺伝子の発現量を調節する約 90 万のシス調節領域（cRE）が含まれている。この二つはどちらも同じ４種類の塩基で構成されている。しかし、遺伝子領域の３塩基ずつのコドンがアミノ酸配列を決定するという明確なルールとは対照的に、cRE が含まれる Non-coding 領域とよばれるゲノム領域は非常に複雑なルールでコントロールされている。一つの遺伝子の遺伝子発現の制御には複数の cRE が関与することが多いが、これらの領域は常に転写因子などの蛋白質がアクセス可能なわけではなく、状況によりクロマチン状態が変化することで、蛋白質がアクセス可能なアクティブな状態と、クロマチンが折りたたまれ、転写因子などが認識できないインアクティブな状態を遷移する。つまり、細胞はコンテクストによりノンコーディング領域のアクティブなゲノム配列そのものを変化させることにより、細胞は遺伝子の発現を調節している。さらにこのクロマチン状態は確率的に変動しているため、アクティブなゲノム配列には可塑性がある。そのため、細胞集団は常に不均一な細胞状態を含んでおり、ワディントンランドスケープに代表されるように(1)、このような不均一性が細胞種の遷移では重要な役割を持つ。さらに病理的な状況では腫瘍細胞が持つエピジェネティックな不均一性が腫瘍の治療抵抗性などと関わることが知られている(2)。このような理由からヒトゲノムのＤＮＡ配列が明らかになり数十年が経過していても、細胞がどのように遺伝子発現を調節しているかをゲノム配列から読み取ることはいまだ困難である。そこで、細胞集団内において、それぞれの細胞のアクティブなゲノム配列を決定し、それぞれのアクティブな領域がどのように遺伝子発現に関与するか、つまり cRE として活性を持つかを一細胞レベルで決定することが、細胞の遺伝子発現制御を理解する上で重要である。

近年このような複雑な規則により決定される現象に対して、人間が明示的なルールを与えずに予測を得る方法として、ディープニューラルネットワークが広く用いられるようになっている。中でも、Attention 機構と呼ばれる技術は、入力されるデータの順序などに依存せず柔軟に学習ができる手法として自然言語処理などで目覚ましい成果をあげている(3)。そこで我々はこの論文で Attention 機構を用いて塩基配列情報、距離情報、クロマチンアクセシビリティ情報を柔軟に統合し、遺伝子ディープニューラルネットワークを用いて、単一モデルで cRE 活性を一細胞レベルで決定する手法を提案する。

2. フレームワークの概要

一細胞レベルで cRE の活性を決定するために我々は scRNA-seq と scATAC-seq を同時に取得可能な scATAC-seq+GEX データを用いた。活性化 cRE はクロマチンがオープン状態になることが知られている。そのため活性化した cRE は scATAC-seq のピーク領域に含まれていると考えられる。しかし scATAC-seq で検出されるピーク領域のすべてが cRE ではない。そこで scATAC-seq で検出されるピーク領域のうちどれが cRE でどれがそうではないのかを見分ける必要がある。そこで cRE は遺伝子発現を制御する領域なので、cRE の活性が変化するとターゲット遺伝子の遺伝子発現が変化するという性質を用いた。さらに cRE の活性とクロマチンアクセシビリティは一般に正に相関することが知られている(4)。そのため、cRE の活性が細胞間で変化する際、クロマチンアクセシビリティとターゲットとなる遺伝子発現は同時に変化する。これらのことから、cRE 領域のクロマチンアクセシビリティ情報と遺伝子発現の相関の強さからある領域が cRE として機能しているかどうかが予測可能であると考えられる。さらに、我々は cRE

の予測精度をあげるため二つの情報を加えた。一つは塩基配列情報を用いた。一般に cRE には転写因子などの蛋白質が特異的な塩基配列を認識して結合し転写複合体を形成することで、転写を促進する。そのためそれぞれの細胞種ごとに cRE には塩基配列の特徴があることが知られており、塩基配列情報からエンハンサー活性やクロマチンアクセシビリティが予測可能であることが知られている(5)。それだけではなく、エンハンサーの活性はエンハンサーそれ自身の塩基配列だけではなく、ターゲットとなる遺伝子のプロモーターの性質にも影響されることが報告されている(6)。そのため、cRE の塩基配列とターゲット遺伝子のプロモーターの塩基配列の組み合わせ情報を用いることが、それぞれの遺伝子を制御する cRE 領域を予測するのに重要であると考えられる。最後の要素として遺伝子と cRE 領域の距離も重要である。cRE と遺伝子プロモーターとの相互作用はゲノムの 3D 構造が変化し cRE と遺伝子プロモーターが近接することが重要であると報告されている(7)。そのため、遺伝子と相互作用する cRE 領域はターゲット遺伝子に近い領域により多く存在する可能性が高い。

これらの情報を柔軟に統合して cRE 領域を決定するために我々はディープラーニング手法を用いた。まず、cRE が一般的に遺伝子を制御可能な遺伝子近傍領域の ATAC-seq ピーク領域のクロマチンアクセシビリティ、ピーク領域の塩基配列、ピークから遺伝子 TSS までの距離情報を用いて遺伝子発現を一細胞レベルで予測するようなニューラルネットワークを構築する。そしてこのニューラルネットワークを学習させたのちに、それぞれのピーク領域の遺伝子発現への寄与度を計算し、先行研究と同様に、寄与度が高い領域をエンハンサー領域として決定した。遺伝子発現への寄与度が高いほどエンハンサー活性が高い傾向があることは過去の研究でも報告されている傾向である(8)。

我々はこのフレームワークは 1 回分の scATAC-seq+GEX データからこのデータに含まれる細胞集団で機能する cRE 領域を決定する。まず一般的なＱＣを行ったのち、プロモーターピークを持つ遺伝子のみを抽出する。プロモーターピークは遺伝子の TSS ±500bp 以内にある ATAC ピークを定義される。そしてプロモーターピークを持つすべての遺伝子に関して、遺伝子近傍領域（TSS±3*10^5）を定義し、この領域に含まれるすべての ATAC-seq ピーク領域がそれぞれの遺伝子の cRE 候補領域となる。１つの学習データは、１つの遺伝子からなり、プロモーターの塩基配列、cRE 候補領域の塩基配列、TSS からの距離、プロモーター・cRE 候補領域の一細胞ごと

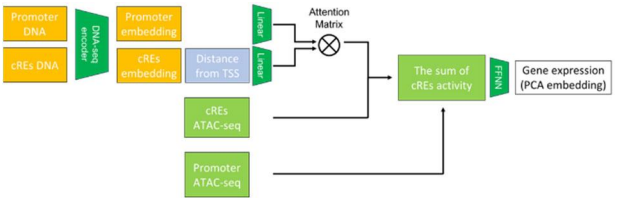

図 1:モデル構造図

の ATAC-seq カウントがネットワークに対する入力となり、これらの情報から一細胞ごとの遺伝子発現を予測する。われわれが構築したニューラルネットワークは二つの部分からなっている（図 1)。一つが DNA シークエンスエンコーダで、もう一つがアテンションブロックである。ニューラルネットワークの処理はまず DNA シークエンスエンコーダが塩基配列情報を処理するところからはじめる。DNA シークエンスエンコーダは 8 層からなる CNN である（図 2)。このネットワークは 1 つのピークの 1350 塩基の塩基配列情報を入力として受け取り、それを 32 次元の特徴ベクトルに圧縮する。まずこのフレームワークでは遺伝子周囲の cRE 候補領域と、プロモーター領域の塩基配列を、ピーク 1 つずつ DNA シークエンスエンコーダに入力し 32 次元に圧縮する。

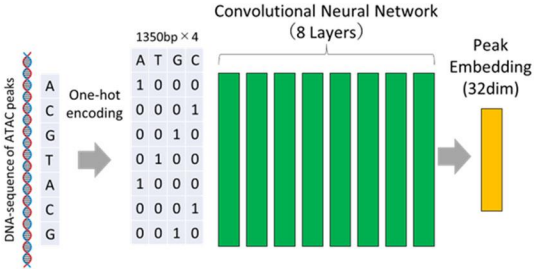

図 2:DNA エンコーダの構造

アテンションブロックは入力として、cRE 候補領域、プロモーター領域の 32 次元の塩基配列圧縮表現、それぞれの cRE 候補領域と TSS との距離、一細

胞ごとの scATAC-seq データを入力として受け取る。cRE 候補領域の数は遺伝子ごとに異なるため、入力のデータ次元は遺伝子ごとに異なる。そこで我々は cRE 候補領域数が少ない場合は、パディングにより入力データを保管し、アテンションを計算した。

アテンションブロックでは、どの cRE 候補領域が遺伝子発現の予測に重要であり、どの領域がそうではないのかを取捨選択する。この cRE 領域のターゲット遺伝子の選択性はプロモーターと cRE 候補領域の塩基配列と組み合わせと、距離によって決定されることが既存研究として報告されている。そこで、我々はプロモーター領域の塩基配列と、cRE 候補領域の塩基配列・距離との間でクロスアテンションを計算することによって重要な領域を抽出した。この重要度はアテンションマトリクスとして計算される。そしてこの重みづけ（アテンションマトリクス）に基づいて、ATAC-seq の情報を足し算する。このようにして得られた ATAC-seq から抽出された特徴量から、遺伝子発現を予測する。モデルの学習は学習を安定させるため、遺伝子発現値の代わりに PCA 圧縮座標を用いて行われる。ニューラルネットワークモデルの学習が終了したのち、我々のフレームワークは Contribution score を計算することで cRE の活性を一細胞レベルで出力する。cRE の活性の計算のために我々は DEEPLift スコアを用いる。DEEPLift スコアはニューラルネットワークの Contribution スコアを計算するためのツールとして開発されている手法である。

3. 性能評価

我々はまず 10x 社から公開されている健常ドナーの顆粒球除去後ヒト末梢血細胞(PBMC)をベンチマーク用のデータセットとして用い、cRE の予測精度を評価した。低発現遺伝子のフィルタリングと、プロモーターピークを持たない遺伝子の除外を行ったのち、我々は 6853 個の遺伝子を解析に用いた。代表的な遺伝子の cRE 活性予測結果を示す。CD3D は T 細胞特異的な発現を持つ遺伝子であるが、興味深いことに、個々の cRE 活性でみると CD4 T 細胞特異的な cRE や CD8 T 細胞特異的な cRE が存在し、細胞種によって活性化される cRE-遺伝子ペアが異なることがわかった(図 3) 。

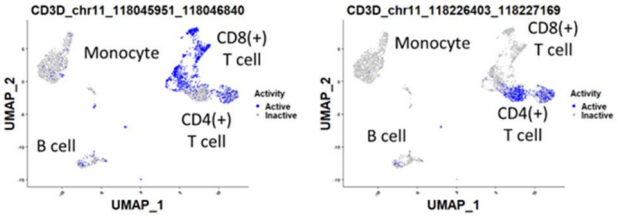

図 3:一細胞レベルで検出された
エンハンサー活性

次に我々は、cRE 予測精度の定量的なベンチマークを行った。ここでは FANTOM5 database(CAGE-seq)(9)と PCHiC-seq(10)の二つの指標を用いて予測をおこなった。そしてモデルが予測する cRE の活性値が、どの程度 CAGE-seq や PCHiC-seq の予測と合致しているかを評価した。そして我々のフレームワークの精度を、同一条件でトレーニングした既存手法(8, 11, 12, 13)と比較した。結果として我々の手法は CAGE-seq の予測タスクでも PCHiC-seq の予測タスクでも既存手法より高い AUROC で cRE を予測することが可能であった(図 4、5)。

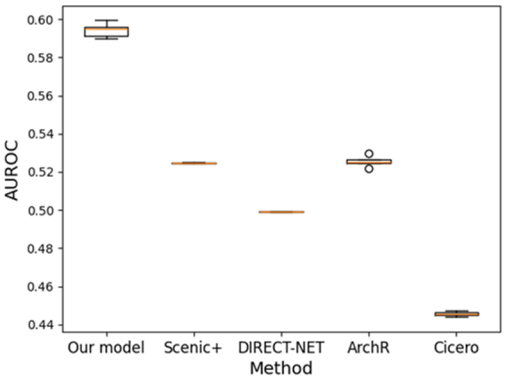

図 4:CAGE-seq データによる性能評価

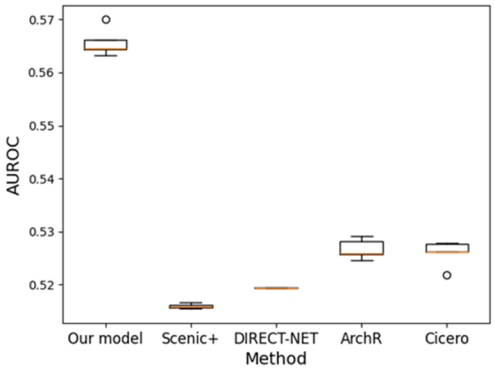

図 5:PCHiC データによる性能評価

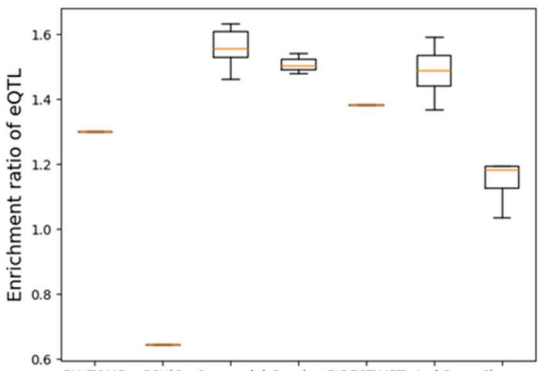

図 6:eQTL の濃縮率評価

次に我々は予測された cRE 領域がどの程度正確であるかをさらに検証するために eQTL がどの程度予測領域に濃縮するかを用いて検証を行った。ここでは GTEx portal(14)に登録されている PBMC の eQTL データを用いてそれぞれのツールで予測された cRE 領域に eQTL がどの程度濃縮するか計算した。結果として我々のモデルは既存モデルよりも有意に eQTL を含む領域を cRE として検出する傾向があった(図 6)。興味深いことに、我々のモデルが予測した cRE 領域は、FANTOM5 データベースや、PCHiC で予測された cRE 領域と比較しても eQTL の濃縮率が高いことがわかった。

4. おわりに

本研究で我々は、Attention 機構を組み込んだディープラーニング手法により、cRE の塩基配列、距離情報、クロマチンアクセシビリティ、遺伝子発現データを用いて一細胞レベルでエンハンサーの活性を決定する手法を構築した。我々の手法は遺伝子間で共通のモデルを用いて転写制御のルールを学習することを特徴としており、既存手法よりも高い精度で cRE の活性を決定することが可能であった。本手法は一検体のシングルセルデータで学習可能であるように設計されているため、臨床検体を用いたエンハンサー活性の解析に適していると考えられる。

5. 参考文献

(1) C Waddington The Strategy of the Genes Allen &Unwin, London, (1957).
(2) M Bi et al. Nature Cell Biology **22** 701–715 (2020)
(3) A Vaswani et al. arXiv:1706.03762 (2017)
(4) Christoph Neumayr et al. Nature **606** 406–413 (2022)
(5) Žiga Avsec et al. Nature Methods **18** 1196–1203 (2021)
(6) D Bergman et al. Nature. **607** 176-184 (2022)
(7) Gil Ron et al. Nature Communications **8** 2237 (2017)
(8) C González-Blas et al. Nature Methods **20** 1355–1367 (2023)
(9) R Andersson et al. Nature **507** 455–461 (2014)
(10) Biola M. Javierre Cell. **167** 1369-1384 (2016)
(11) L Zhang et al.Sci Adv. **22** eabl7393 (2022)
(12) Jeffrey M Granja et al. Nat Genet. **53** 403-411 (2021)
(13) H Pliner et al. Mol Cell. **71** 858-871 (2018)
(14) The GTEx Consortium Nat Genet. **45** 580–585 (2013)