©2024 筑波大学生物学類

機械学習によるメダカの遺伝子発現量の予測

古川　愛海（筑波大学 生物学類）　　指導教員：尾崎　遼（筑波大学　医学医療系）

【背景・目的】

遺伝子が生体内において適切に機能を発現するために、時空間的な転写制御は不可欠である。特に真核生物においては、プロモーターやエンハンサーなどの転写制御領域によって遺伝子発現調節が行われる。遺伝子の転写制御領域と遺伝子発現に関する分野は情報科学を用いた研究がされており、深層学習や機械学習を用いた転写制御領域の予測モデルや、転写因子結合モチーフの研究開発が行われている。転写制御の先行研究は、ヒトやマウスなどのモデル生物に対しては盛んに行われている[1]のに対し、非モデル生物においてそのような研究はほとんど行われておらず、ゲノム配列やオミクスデータに関するデータベースが整備されていない。

そこで、非モデル生物のゲノム配列から転写制御領域を予測するモデルの構築を見据え、本研究では、非モデル生物の１種であるメダカの転写制御領域から遺伝子発現量を予測するモデルをメダカの転写制御領域の遺伝子配列から遺伝子発現量を予測する機械学習モデルを構築することを目的とした。

【材料】

メダカの転写制御領域配列(入力値)

メダカ(*oryzias latipes*)の転写開始点の上流 1000bp と下流 1000bp の合計 2000bp の塩基配列を転写制御領域の配列とし、遺伝子ごとに USCS Table Browser からダウンロードした。

メダカの遺伝子発現量(予測値)

メダカの胚発生のステージごとの RNAseq データを用いた[2]。データの分散を減らすために、全てのデータに 1 を足し、10 で対数をとったステージ 11 とステージ 40 の数値を実験に用いた。

対象とした遺伝子

転写制御領域と発現量のデータの両者が揃っている遺伝子を用いた。また、対象遺伝子から Y 染色体およびミトコンドリアゲノムの遺伝子を除いた。さらに、適切な予測の検証を行うために、全てのサンプルでの発現量が 0 である遺伝子は除いた。最終的に 11203 の遺伝子を対照とした。

【方法】

データセットの作成

対象の遺伝子データのうち、8 割を訓練データ、2 割をテストデータとした。リーキング（テストデータの情報が訓練データに不適切に含まれること）を防ぐために、訓練データセットの遺伝子の中で、テストデータセットの遺伝子領域に重なるものは除いた。その上で、全ての遺伝子において発現量から対象サンプルの訓練データの平均値を引いた数値を実験用の遺伝子発現量とした。

また、下記で用いるモデルには塩基配列をそのまま入力できないため、転写制御配列を長さ 3,4,5 の部分配列の数である 3,4,5-mer のカウントデータ、または、遺伝子配列のベクトル化処理である GENA-LM[3]による 1024 次元の数値ベクトルに変換した。

機械学習モデルによる計算機実験

機械学習モデルには、random forest (RF) と XG boost を用いた。訓練データで発現量の予測を行うことでモデルの訓練を行い、訓練を行ったモデルでテストデータの予測を行い、モデルの精度の評価を行なった。ハイパーパラメータは、n_estimators, random_state, max_depth の 3 点について手動で調整した。結果の評価には、テストデータにおける実際の発現量の値と予測値の差の 2 乗である平均二乗誤差(MSE)と、実際の発現量の値と予測値の相関係数を利用した。MSE が低く、相関係数が高い予測が精度の高い予測であると言える。

【結果・考察】

入力処理とモデルの組み合わせにおいて、最も正確に予測できたのは GENA-LM のベクトル表現と RF の組み合わせであった。ステージ 11 のサンプルに関して相関係数 0.55、MSE 0.024 という精度で予測が可能であった。また、1 つの遺伝子配列入力から 2 つのサンプルの発現量を予測するモデル(multiple-output model)で最も精度が高くなった。

最も精度が高かったモデルを用い、野生型の転写制御配列に一塩基置換を入れた変異配列を入力とした際に遺伝子量の予測がどのように変化するかを調べる in silico mutagenesis 実験を、ステージ 11 で特異的に発現している上位 100 遺伝子に対して実施した。その結果、特に、G や C に置換した際に発現量が上昇するという傾向が見られた。

さらに転写開始点の付近では配列に変異が加わると発現量が下がるという既存知識から、特に転写開始点の付近において変異を加えた配列からの予測値が元の配列からの予測値よりも低下するという仮説を立てた。しかし、予想に反して変異を加えた配列の予測値の変化は転写開始点からの位置と関係ないということがわかった(図 1)。

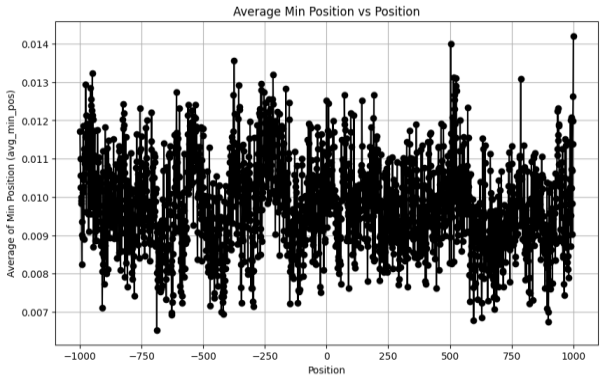

図１　変異を加えた塩基の位置(x 軸の 0 が TSS)と変異による予測値変化の関係

【展望】

変異に対する予測値の変化を求め、転写制御配列の中で遺伝子発現量予測に対して影響力の高い部分配列を同定したい。そのような部分配列は転写調節領域である可能性が高いため、その部位を予測するモデルを作ることによって、メダカの転写調節領域を予測することができるのではないかと期待される。

【参考文献】

[1]Eraslan et l., Net Rev Genet 20, 389-403, 2019
[2] Li Y et al., Genome Res, 30(6) 924-937,2020
[3]Veniamin F et al., bio Rxis, 544592, 2023