論文内容の要旨

論文題目：DNA 一次配列からのプロモーター活性予測

氏名：入　江　　拓　磨

遺伝子発現制御は多くの生命現象において重要な制御段階である．遺伝子発現制御は転写・翻訳など多段階の制御ステップで構成されるが，中でも転写開始の制御は最初のステップであるため主要な制御段階であると言える．転写開始の制御は遺伝子近傍のプロモーターと呼ばれるゲノムの制御領域によって担われている．現在ゲノム規模でプロモーター領域の配列解析をすることが可能であり，プロモーターの配列情報を用いて，細胞の転写制御システムの全体像を明らかにすることが期待できる．実際にプロモーター配列情報を用いて転写制御の数学的なモデル化の方法が幾つか提案されている．従来の転写制御モデルの構築には，マイクロアレイなどを用いた解析が主に用いられてきた．そのため多くの予測モデルは転写応答性，すなわち遺伝子ごとの変化・相対値の予測モデルで，プロモーター自体の強度(転写の絶対量)を予測・モデル化する研究はほとんど存在しなかった．また mRNA レベルの発現は様々な制御ステップを経た産物である．すなわちプロモーター配列以外にもゲノムの CpG のメチル化，クロマチン構造，mRNA 合成効率，分解効率など様々な制御の影響の総体であるため，プロモーター活性の絶対量を予測することは困難であった．したがってプロモーターの DNA 配列情報とそこに内在している転写活性化能の関係の解析には別の実験的アプローチを取る必要があると考えた．本研究では，体系的ルシフェラーゼアッセイの情報を用いることで，DNA 一次配列情報からプロモーター活性を予測が可能であるか検証し，プロモーター活性予測モデルの構築を試みた．さらに構築したプロモーター活性予測モデルを用いてヒトゲノムにおけるプロモーター活性予測値の分布と転写に関わっていると考えられる転写開始点，RNA ポリメラーゼ II の結合位置，ヌクレオソーム構造の情報を用い，DNA 一次配列のプロモーター活性と mRNA の発現量，転写に関係する情報との比較を行った．

転写活性化能の測定には定量的ルシフェラーゼレポーターアッセイの手法を用いた．HEK293 由来の完全長 cDNA の 5’端情報により決定されたプロモーター領域 451 種類，lncRNA のプロモーター領域 35 種類，非プロモーター領域 248 種類を用い，HEK293 細胞内における転写活性の測定を

行った．DNA 配列に存在する転写因子結合配列(transcription factor binding site,以下 TFBS)のモチーフ探索には position weight matrix (PWM)として TRANSFAC2008.3 を用いたマトリックス検索を行った．モチーフの候補及び閾値には vertebrate_non_redundant_minFP.prf を用い，167 種類の TFBS を解析対象とした．

プロモーター活性予測モデルには，プロモーター活性が各 TFBS のスコアの和とした次のような線形和モデルとした．

$$\log(Y) = \sum AX$$

Yを DNA 断片のプロモーター活性，A を TFBS の数(または DNA-転写因子の親和性のスコア)，Xを TFBS のプロモーター活性への寄与のスコアとした．各 TFBS を説明変数，DNA 断片のプロモーター活性を目的変数としたモデルとした．重回帰分析の手法で Xの推定値を計算し，プロモーター活性予測値を得た．モデルの評価はプロモーター活性の実験値と予測値の相関係数(Pearson's correlation coefficient)値を用いた．重回帰分析の結果，プロモーター活性の実験値と予測値の相関係数が r=0.82 となり，ある程度の相関係数が得られた．また実験値の 5 倍以内の範囲で予測できたものが全体の約 75%であった．次にモデルの改善が可能であるか検討した．(1) TFBS のマトリックス検索時のスコア (2)TFBS の存在位置 (3)変数選択について検討した．PWM は位置特異的な塩基の出現確率をスコア化したもので，そのスコアが高いほどコンセンサス配列に近くなるため，スコアが高いほど転写因子と DNA との親和性が高くなると考えられる．そこで PWM のスコアの利用を検討した．TFBS の存在位置については，幾つかの TFBS は転写開始点付近に存在するが知られており，転写開始点付近に存在している TFBS に機能的なものが存在する確率が高いと考えた．また機能的ではないと考えられる TFBS についても計算から除外することで予測精度の向上が可能であるか検討した．得られたモデルに対し赤池情報量規準を用いることで，プロモーター活性の説明に寄与の大きい TFBS の選択を行った．最適なマトリックススコア及び領域の条件を適応した結果，実測値と予測値の相関係数が r=0.87（図 1 左），約 86%のクローンにおいて，予測値が実測値の 5 倍以内に予測できた（図 1 右）．この得られたモデルに対して 10 分割交差検定によるモデルの汎化性の評価を行ったところ，予測値と実測値の高い相関が得られ(r = 0.83)，過学習の影響も小さく，未知データに対しても有効であるモデルを構築することができた．

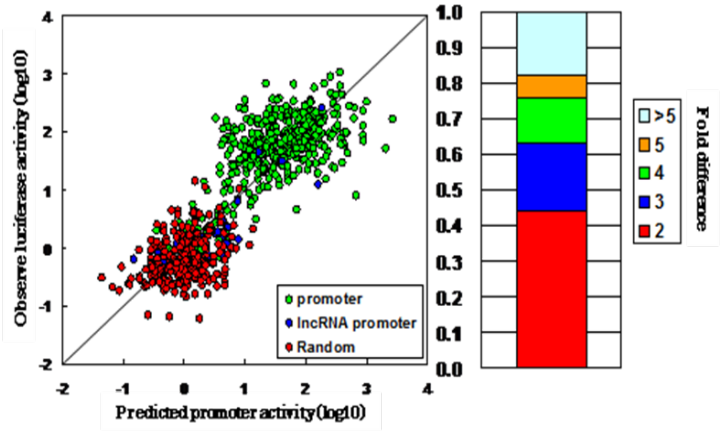

図1：プロモーター活性予測モデルの精度
（左）プロモーター活性予測値（x軸）と実測値（y軸）との相関．（右）予測された値の実測値からの範囲

ルシフェラーゼアッセイの系に対して精度の高いモデルを構築することができた．次に mRNA レベルの発現を *in vivo* の転写活性化能情報として用いてプロモーター活性予測モデルとの比較を行っ

47-077607：入江　拓磨

た．転写の情報としてオリゴキャップ法と Illumina GA を組み合わせた TSS-seq 法による転写開始点情報を用いた．18,686 種類の Refseq 遺伝子の上流 1 kb の領域をプロモーターとし，その領域のプロモーター活性予測値とマップされた転写開始点の頻度情報と相関を調べた．その結果，これらの相関は極めて低かった．次に，HEK293 細胞で転写されている遺伝子とそうではない遺伝子の区別が可能か，モデルの定性的な予測精度について評価を行った．充分に転写活性があると考えられる 5ppm (parts per million)以上の転写開始点が得られたプロモーター領域を HEK293 細胞において”active”な領域(4,749 領域)とし，TSS が観測されなかったプロモーター領域を”silent”な領域(8,315 領域)とした．それぞれの領域のプロモーター活性予測値を算出した結果，前者のセットに対して与えられた予測スコアの分布と後者のセットに対して与えられた予測スコアの分布は，重なりは大きいものの，有意な差異が認められた（図 2 棒グラフ）(P<1×10^{-100};Wilcox test)．高いスコアを与えたプロモーターは HEK293 細胞において有意な転写活性を示すプロモーターであり，本研究によるモデルにより HEK293 で発現しているプロモーターとそうでないプロモーターに定性的な差を検出することが可能であった．このことは，このモデルが単なるプロモーター領域の予測ではなく，HEK293 中での転写が行われているプロモーターの予測が可能であることを示している．

しかしながら，転写が確認できない(TSS タグ数 = 0)ものの高い予測値(> 1)を得たプロモーター領域も存在していた．HEK293 細胞内の RNA polymerase II(Pol II)の結合箇所を ChIP (Chromatin immunoprecipitation) Seq 法により調べたところ，予測スコア>1，TSS タグ 0 の 3,600 領域中約 39%

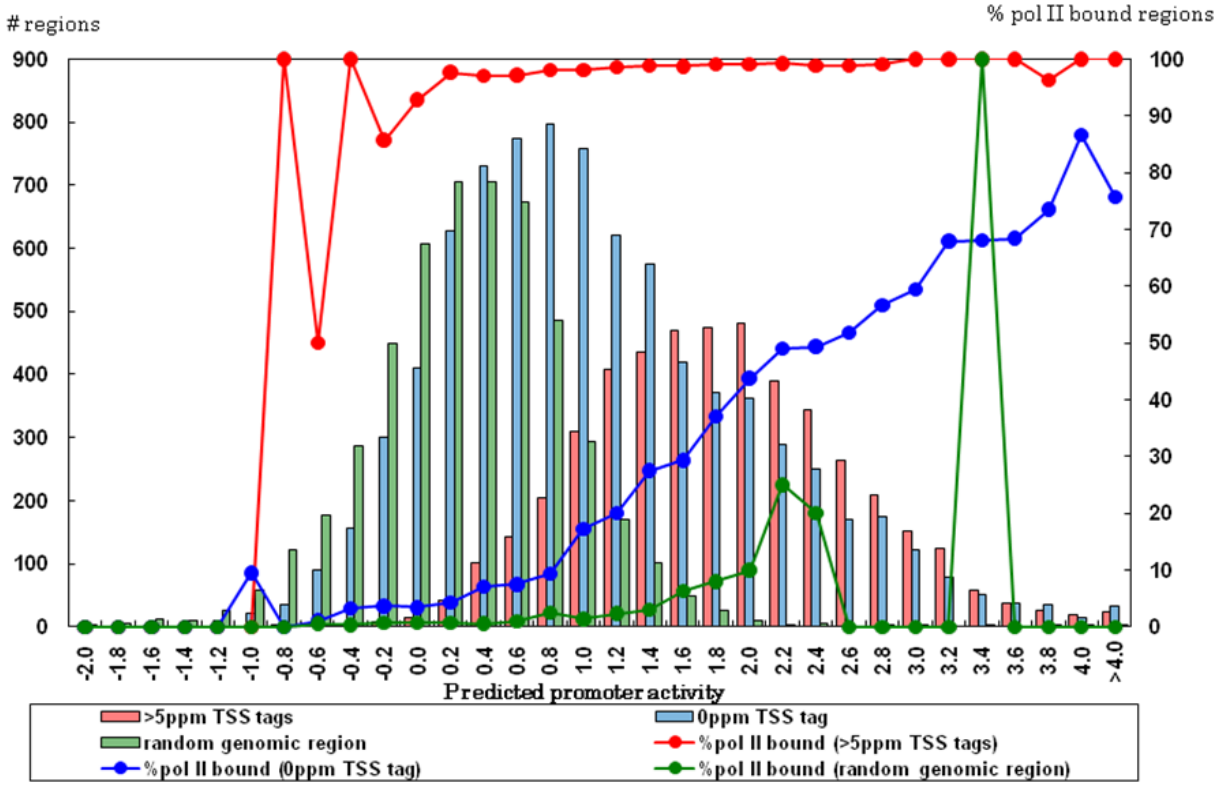

図2：TSS seqとPol II seqを用いた予測モデルの評価

RefSeq遺伝子の5′上流領域のプロモーター活性の予測値スコア（x軸）の分布のヒストグラム（頻度；y軸左側），棒グラフ赤（>5ppm），青(0ppm)，緑(ランダム領域)． 折れ線グラフはChIP Seq(pol II)の結合が確認された領域の割合(y軸右側)．赤(>5 ppm)，青(0ppm)，緑(ランダム領域)

47-077607：入江　拓磨

の領域おいて Pol II の結合が確認された．さらに予測スコアの値と Pol II の結合の割合にも相関が見られた(図 2，青線)．また，HEK293 細胞の Nuleosome seq のデータを利用し，遺伝子近傍のクロマチン構造を解析した結果，プロモーター活性予測値 >1，TSS>5ppm の領域では，転写活性の高い遺伝子に特徴的な開いたクロマチン構造を取っていた（図 3A)．さらに転写が見られないものの高いプロモーター活性予測値を与えた領域においても開いたクロマチン構造をとる傾向にあった（図 3B)．すなわち高いプロモーター活性を与えた領域は転写の有無に関わらず，開いたクロマチン構造を取り，Pol II が結合していた．これらの例は転写後速やかに分解され TSS が検出できない遺伝子，もしくはクロマチン構造を開き Pol II をリクルートすることはできるが転写伸長を起こす要因に欠けている遺伝子であると示唆される．後者の例は近年解析が進んでいる“transcriptional pausing”の例であると考えられる．

またプロモーター活性予測値のゲノム全体の分布について解析した．ヒトゲノム配列を 1.2kb の幅で分割しそれぞれのプロモーター活性予測値を算出した．その結果，高いプロモーター予測値 lncRNA のクローンの 5’端領域とのオーバーラップする例を確認できた．また RefSeq 遺伝子の 5’端領域と同様に，TSS が存在していないものの高いプロモーター活性予測値を得た領域において開いたクロマチン構造を取る傾向にあることも確認でき，潜在的に高いプロモーター活性をもつ領域がヒトゲノム中に多数存在していることが示唆された．

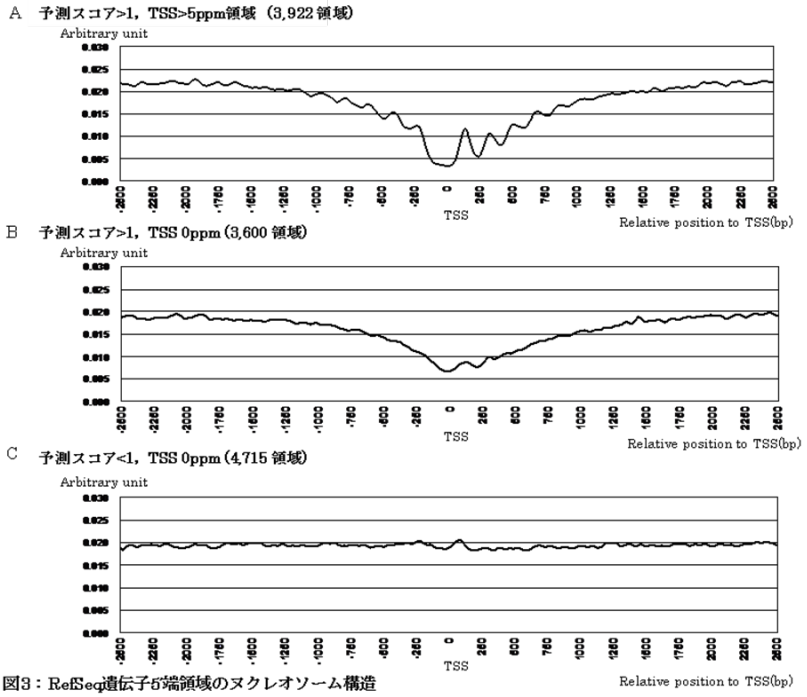

図3：RefSeq遺伝子5'端領域のヌクレオソーム構造

RefSeq遺伝子5'周辺領域のヌクレオソーム構造を示した．転写開始点を基準（0）としてヌクレオソーム占有率(y軸)を計算した．(A)予測値>1，TSS > 5ppm，(B)予測値>1，TSS=0，(C)予測値<1，TSS=0のRefSeq遺伝子

以上の結果から，体系的なルシフェラーゼアッセイの情報を用いることによって，単純なモデルではあるが DNA 一次配列のプロモーター活性予測モデルを構築できたと考えている．またヒトゲノム配列のプロモーター活性予測値の解析から，転写が行われていないが潜在的に高いプロモーター活性を有する領域を大多数見出した．本研究のプロモーター活性予測モデルがゲノム配列とトランスクリプトームを結ぶ転写制御の全体像の理解へ繋がると期待している．

論文目録

Predicting promoter activities of primary human DNA sequences.

Takuma Irie, Sung-Joon Park, Riu Yamashita, Masahide Seki, Tetsushi Yada, Sumio Sugano, Kenta Nakai, Yutaka Suzuki

Nucleic Acids Research. *in press*