様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 25 年 4 月 30 日現在

機関番号：14301
研究種目：基盤研究(A)
研究期間：2010 年度～2012 年度
課題番号：22240032
研究課題名（和文）
プロモーター配列の設計
研究課題名（英文）
Designing promoter sequences
研究代表者
矢田 哲士（YADA TETSUSHI）
京都大学・大学院情報学研究科・准教授
研究者番号：10322728

研究成果の概要（和文）：
与えられた培養細胞において、与えられた転写強度を示すプロモーターの DNA 塩基配列を体系的に設計する計算手法を開発した。この手法を用いることによって、わずか 2 箇所に塩基置換を導入することで、培養細胞 HEK293T におけるヒト CRE プロモーターの転写強度を～2.5 倍に引き上げる改変プロモーターの設計に成功した。一方で、ヒト IFNB プロモーターのように、転写因子の結合部位（TFBS）が互いに重複している場合には、設計が極めて難しくなることが明らかになった。

研究成果の概要（英文）：
We have developed a computational method of designing DNA base sequences of promoters whose transcriptional activities satisfies a given condition (transcriptional strength in a given cell line). The method has introduced two base substitutions into human CRE promoter, then we have verified that these substitutions successfully increased its transcriptional strength in HEK293T cell line ~2.5 times as high as that of the original promoter. On the other hand, it has become clear that designing promoter sequences is interestingly difficult in the case of existing overlapping TFBSs (transcriptional factor binding sites) such as human IFNB promoter.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	14,100,000	4,230,000	18,330,000
2011 年度	12,200,000	3,660,000	15,860,000
2012 年度	12,200,000	3,660,000	15,860,000
年度			
年度			
総　計	38,500,000	11,550,000	50,050,000

研究分野：バイオインフォマティクス
科研費の分科・細目：情報学・生体生命情報学
キーワード：プロモーター，転写制御，バイオインフォマティクス，ゲノム生物学，合成生物学

１．研究開始当初の背景

完全長 cDNA の配列データをゲノム配列にアラインメントすることで、ヒトでは、ゲノム中のプロモーターの位置に関する大規

模なカタログ化が完了した［Wakaguri et al. 2008］。しかし、ここで明らかになったプロモーター像は一次元的なもので、プロモーターが持つ生物学的な性質（時間や空間における転写活性の強度）は分からないままであった。

一方で、酵母をモデル生物にしたものであるが、プロモーターの配列からその転写の強度を推定しようとする試みが行われた［Bussemaker & Siggia 2001；Beer & Tavazoie 2004；Das et al. 2004；Hguyen & D'haeseleer 2006；Yuan et al. 2007］。これらの試みでは、プロモーター配列の構造を単純化して、幾つかの転写因子結合部位（TFBS）の配列がバックグラウンドの配列にランダムに埋め込まれたものとしてプロモーターの配列を捉えている。そして、各々の TFBS を説明変数に、プロモーターの転写強度を目的変数にした回帰式を導出している。回帰式には、線形回帰やロジスティック回帰をはじめとするさまざまなタイプのものが提案されている。しかし、これらの試みは、マイクロアレイによる遺伝子発現データをプロモーターの転写強度のデータに見立てて行なわれた。つまり、異なる条件下での遺伝子発現の相対的な変化をプロモーターの転写強度としていたため、絶対的な転写強度の推定になっていないとの批判があった。

２．研究の目的

ここでは、プロモーターの DNA 塩基配列を体系的に設計する計算手法を開発する。プロモーターの塩基配列の設計とは、入力として与えられた転写活性の条件（例えば、ある培養細胞における転写の強度）を満たすプロモーターの配列を出力する問題である（図 1 の右向きの矢印）。

遺伝子の転写制御を司るプロモーターの塩基配列を自在に設計することができれば、その学術的な価値や産業的な価値は計り知れない。プロモーター配列の設計法を確立することができれば、さまざまな遺伝子の機能を解析する基礎研究において、また、遺伝子治療をはじめとする応用研究において、強力な基盤技術を提供することができる。そして何より、この方法論の確立は、プロモーター配列の構築原理の理解に深く結びついている。この研究は、ゲノム上での位置が同定されることで明らかになったプロモーターの一次元的な像に、転写活性の強度の軸を時間や空間とともに新たに与え、それらと塩基配列との関係性を詳らかにしようとする挑戦的で独創性の高い試みである。

３．研究の方法

ここでは、プロモーター配列の設計問題に取り組む前に、プロモーターの転写強度の推定問題を考える。転写強度の推定問題は、ある培養細胞におけるプロモーターの転写強度をその塩基配列から推定する問題である（図 1 の左向きの矢印）。プロモーターの塩基配列からその転写強度を推定するモデルを構築することができれば、プロモーター配列の設計問題は、このモデルを用いた探索問題と見なすことができる。

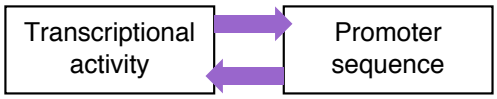

図１：プロモーター配列の設計問題とプロモーターの転写強度の推定問題の関係

転写強度の推定モデルを構築するための基盤データとして、ここでは、プロモーターの転写活性の絶対量を計測する。さらに、プロモーターが受ける転写制御を塩基レベルの解像度で明らかにするために、ランダムな突然変異を導入したプロモーターライブラリを作成する。

４．研究成果

（１）基盤データの生成

転写強度の推定モデルを構築する基盤データを生成するために、ハイスループットなプロモーター解析を行なう実験系を構築した。プロモーター解析は、プロモーターの転写活性の測定と塩基配列の決定から成り、これらの情報の取得を次世代型シークエンサーIllumina HiSeq2000 を利用して行なった。実験にはプラスミド中のレポーター遺伝子の 3'UTR 中にランダムな 12 塩基を挿入したライブラリーを用い、ランダム 12 塩基をいわゆるバーコードタグとしてプロモーターとの対応付けに利用した。幾つかの野生型のヒトプロモーターについて、error prone PCR によって 1～4%のランダムな突然変異を導入したプロモーターライブラリーを作成し、ヒト培養細胞における転写活性を測定した。レポーター遺伝子の cDNA からランダム 12 塩基配列を決定し、その塩基配列のリード数を計測することで転写活性の指標とした。その結果、いずれの変異プロモーターにおいても転写活性が正負に 5 倍程度の範囲で変化することを確認した。プロモーター配列の決定では、paired-end sequencing により、read1 でプロモーターの塩基配列を決定し、read2 でランダム 12 塩基の配列を決定した。そして、ランダム 12 塩基配列を用いて read1 の配列情報をグループ化することで、プロモーター中の突然変異の位置を推定する。

このプロモーター解析では、数万種類のプロモーターに関する転写活性の情報と塩基配列の情報を一度に取得することができる。また、突然変異の導入が限定的（数%のランダムな突然変異）であるため、ここで取得されるプロモーターの多くが共通の転写制御を受けると期待することができる。

（２）転写強度の推定モデル

ここでは、転写強度の推定モデルとして、線形回帰モデルを採用した［Jonsson et al. 1993］。目的変数はプロモーターの転写強度、説明変数はプロモーターの各位置における各塩基である。回帰モデルの汎用性を高めるために、ここでは、BOLASSO（bootstrap LASSO）［Baraud et al. 2003］を導入することで説明変数を選択した。

転写強度の推定モデルに線形回帰モデルを採用することで、プロモーター配列の設計はし易くなるが、転写強度の記述力は乏しくなる。一方で、非線形回帰モデルを導入することの効果は、極めて限定的であることがから報告されている［Melnikov et al. 2012；Patwardhan et al. 2012］。そこで、ここでは、複数の線形回帰モデルを構築することで記述力の乏しさを補うことを考えた。具体的には、セントロイドになるプロモーター配列（初めは野生型のプロモーター配列）を選び、その配列から一定の配列距離以内にあるプロモーター配列を使って線形回帰モデルを構築する。次のセントロイドは、それまでの線形回帰モデルの構築に使われなかったプロモーター配列の中から、一定の配列距離以内にあるプロモーター配列の数ができるだけ多いものを選ぶ。ここで、線形回帰モデルの記述力を損わないように、できるだけ多くのプロモーター配列から線形回帰モデルを構築するように配列距離の測り方を工夫している。すなわち、BOLASSO によって選択された位置の塩基だけで配列距離を計測している。

転写強度の推定モデルの性能を評価するために、ここでは、ヒト CRE プロモーターに～10%程度のランダムな突然変異を導入したプロモーターの転写強度のデータを利用した［Melnikov et al. 2012］。配列数は 26,438 配列、配列長は 87bp である。このプロモーターの下流に TATA ボックスを繋げ、培養細胞 HEK293T で転写強度を計測した。

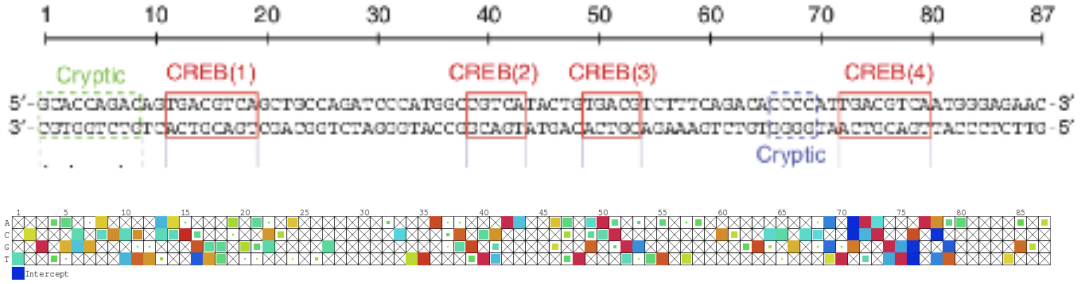

図２：ヒト CRE プロモーターの構造（上）と構築された転写強度の推定モデルの構造（下）。ヒト CRE プロモーターには、CREB の TFBS が 4 つ、cryptic な TFBS が 2 つあることが知られている。転写強度の推定モデルの縦軸は塩基、横軸はプロモーター中の位置を表す。各位置における各塩基のセルが転写強度の推定モデルの説明変数の様子を表わす。×印は、BOLASSO で選択されなかったものを表わす。セルの色塗りされた四角の大きさは、統計的な有意性を表す。赤色は転写への正の寄与を、青色は負の寄与を表す。

図２にヒト CRE プロモーターの構造と構築された転写強度の推定モデルの構造を示す。この転写強度の推定モデルは、野生型をセントロイドにした変異型プロモーター13,663 配列から構築されたもので、155 個の説明変数から成る。既知の TFBS に当たる説明変数がうまく選択されていることが分かる。10 分割クロス検定で明らかになったこの線形回帰モデルの性能は、相関係数 0.833、決定係数 0.694、自由度調整済み決定係数 0.690、平均二乗誤差は 0.436 である（図３）。

（３）プロモーター配列の設計

図２の転写強度の推定モデルを用いて、ヒト CRE プロモーターの改変を試みた。まず、効果的に転写強度を増強すると推察される 2 つの塩基置換をプロモーターに導入したところ、転写強度が～2.5 倍に増強されることが確認できた（図４）。これまでの解析では、変数選択された塩基について、～5%までの改変であれば、信頼性の高いプロモーター配列の設計が可能であった。

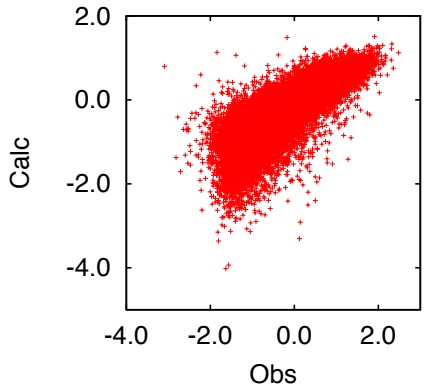

図３：転写強度の推定モデル（図２下）の性能。10 分割クロス検定で評価した。

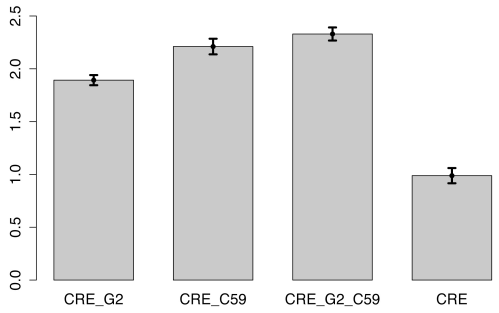

図４：ヒト CRE プロモーターに 2 つの置換を導入した改変プロモーターの転写強度。ここでは、CRE プロモーターの転写強度を 1.0 にしている。CRE_G2 は CRE プロモーターの 2 番目の位置の塩基を G に置換したもの、CRE_C59 は 59 番目の位置の塩基を C に置換したもの、CRE_G2_C59 は 2 番目と 59 番目の位置の塩基をそれぞれ G と C に置換したもの。

一方で、ヒト IFNB プロモーターのように、TFBS が互いに重なり合うプロモーターでは、設計が極めて難しくなることが明らかになった。このプロモーターの転写強度の推定モデルは、112 個の説明変数から成る。そしてその性能は、相関係数 0.237 まで下がる。TFBS が互いに重なり合うプロモーターでは、数塩基の置換が転写の制御をがらりと変えてしまうため、転写強度の推定が難しくなると考えられる。一方で、比較的多くの塩基置換を導入しても、それらがうまくバランスすれば、野生型プロモーターに似た転写制御を受ける変異型プロモーターが生成されているようである。

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 5 件）

1. Irie, T., Park, S.-J., Yamashita, R., Seki, M., Yada, T., Sugano, S., Nakai, K., Suzuki, Y.: Predicting promoter activities of primary human DNA sequences, Nucl. Acids Res., 査読有, 39, 2011, e35 DOI:10.1093/nar/gkr173
2. Yada, T., Yoshida, K., Morita, M., Taniguchi, T., Irie, T., Suzuki, Y.: Linear regression models predicting strength of transcriptional activity of promoters, Genome Inform., 査読有, 25, 2011, 53-60 https://www.jstage.jst.go.jp/article/gi/25/1/25_1_53/_pdf
3. Ichinose, N., Yada, T., Gotoh, O.:Large-scale motif discovery using DNA Gray code and equiprobable oligomers, Bioinform., 査読有, 28, 2012, 25-31 10.1093/bioinformatics/btr606
4. Fujiwara, T., Yada, T.:miRNA-target prediction based on transcriptional regulation, BMC Genomics, 査読有, 14(Suppl.2), 2013, S3 10.1186/1471-2164-14-S2-S3
5. 入江拓磨, 鈴木穣：DNA 配列から転写因子結合部位を予測・検索する WEB ツール, 実験医学, 査読無, 29, 2011, 625-629

〔学会発表〕（計 8 件）

1. Yada, T., Yoshida, K., Morita, M., Taniguchi, T., Irie, T., Suzuki, Y.: Linear regression models predicting strength of transcriptional activity of promoters, 日本バイオインフォマティクス学会年会, 2010.12.13-15, 福岡
2. Irie, T., Park, S.-J., Yamashita, R., Seki, M., Yada, T., Sugano, S., Nakai, K., Suzuki, Y.: Predicting promoter activities of primary human DNA sequences, Biochemistry and Molecular Biology BMB2010, 2010.12.7-10, Kobe, Japan
3. Suzuki, Y.:Transcriptome analysis of human genes, Multilevel Systems Biology: Genome, Structure, and Networks, 2011.11.16-17, Osaka, Japan
4. Irie, T., Park, S.-J., Yamashita, R., Seki, M., Yada, T., Sugano, S., Nakai, K., Suzuki, Y.: Predicting promoter activities of primary human DNA sequences, 第34回日本分子生物学

会年会, 2011.12.13-16, 横浜

5. Liu,Y., Ichinose,N., Yada,T.:A statistical model for predicting transcription activities based on linear regression, 日本バイオインフォマティクス学会年会, 2012.10.14-17, 東京
6. 矢田哲士, プロモーター配列の設計：第35回日本分子生物学会年会, 2012.12.11-14, 福岡
7. 入江拓磨,関真秀,菅野純夫,矢田哲士, 鈴木穣:超並列シークエンサーを用いたヒト変異プロモーターのハイスループット解析, 第35回日本分子生物学会年会, 2012.12.11-14, 福岡
8. Fujiwara,T., Yada,T.:miRNA-target prediction based on transcriptional regulation, ISCB-Asia/SCCG, 2012.12.17-19, Shenzhen, China

〔その他〕
ホームページ等
http://www.genome.ist.i.kyoto-u.ac.jp/~ichinose/hegma/

６．研究組織
(1)研究代表者
矢田 哲士（YADA TETSUSHI）
京都大学・大学院情報学研究科・准教授
研究者番号：10322728

(2)研究分担者
鈴木 穣（SUZUKI YUTAKA）
東京大学・大学院新領域創成科学研究科・准教授
研究者番号：403236464