転写制御領域の構築原理解明

東京大学医科学研究所ヒトゲノム解析センター
中井　謙太

Clarification of The Principle Governing The Architecture of Transcriptional Regulatory Regions

Kenta Nakai
Human Genome Center, Institute of Medical Science, University of Tokyo

Although hundreds of complete genome sequences have been determined so far, we do not understand their 'grammar' nor the principle very well and thus we cannot interpret how the regulation of gene expression is encoded. Such 'grammatical' knowledge would be practically important for the design of artificial promoters. Therefore, we have challenged to clarify such a principle from the following three aspects: (1) evolutionary analyses, (2) mathematical modeling of promoters governing co-expression, and (3) analyses of transfected/artificial promoters. The results of these analyses give further evidences that the regulatory information is primarily specified with one-dimensional distribution of *cis*-elements in spite of the importance of epigenetic effects.

1. はじめに

本研究の目的は、いろいろな生物種のゲノムにおいて、遺伝子の転写制御情報がどのように書き込まれているのかを理解することであった。具体的には、与えられた遺伝子の周囲の塩基配列から、その遺伝子が受ける制御の内容（どんな外部刺激に応答するか、どんな組織や発生時期に発現するか等）を読み取ることを目指してきた。このような研究を通して、遺伝情報の理解という原理的な興味を追求するだけでなく、人工プロモーター設計の指針になる理論モデルをつくることを目標にして、３つの観点から研究を進めてきたが、副産物的な成果を含め、多くの成果が得られた。

2. 研究開発の成果

2.1 プロモーター比較に基づく進化的視点

ヒトとマウスのオーソログ遺伝子の上流領域と下流のアミノ酸配列の進化的保存度を比較してみたところ、両者の間にはほとんど相関がみられないことを見いだした[1]。同様に、ヒトとマウスの選択的プロモーターの進化的保存を網羅的に調べたところ、多くの選択的プロモーターの保存の度合いは低く、A/T塩基や繰り返し配列に富むなどの特徴をもつことがわかった。そこで、ヒトのプロモーターには大きく２種類があり、後者は進化の過程で比較的簡単に生成消滅していることを提唱した[2, 3]。

プロモーターの起源に関する研究も行った。逆転写酵素によって、mRNA配列がゲノムDNAに組み込まれるレトロトランスポジションでは、もとのmRNAの転写を制御したシグナルはmRNAの

中には含まれないため、組み込まれた配列は偽遺伝子になるものと考えられていた。しかし、DBTSS[4, 5]をみても明らかなように、多くの遺伝子では転写開始点は一つに固定されたものではなく、複数の転写開始点をもつ。従って、下流に存在する制御領域がmRNA配列中に含まれれば、転移後、その制御領域が新しいプロモーターとして機能する可能性が考えられるが、実際に半数以上の例ではコアプロモーター領域の大半が一緒に転移していることがわかった[6]（**図１**）。

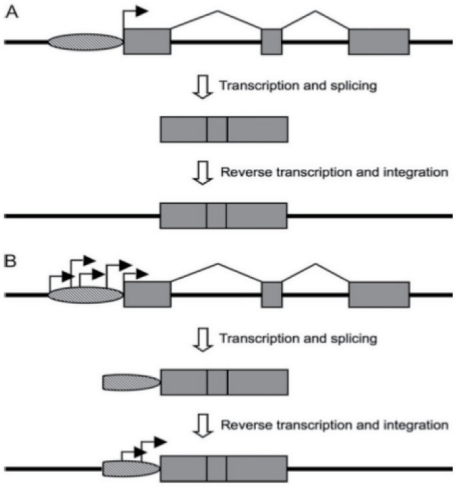

図１　新規プロモーター生成モデル

2.2 共発現遺伝子制御領域のモデル化による数理的視点

ChIP-chip法などで得られる転写因子結合部位（TFBS）のゲノムワイドな情報から、TFがプロモーターに結合する様相をグラフィカルモデルで推定した。この方法により、TFとプロモーターの相互作用様式（TF間直接的相互作用、TF間間接的相互作用、単独）を同定することができる。今回、大規模な転写制御情報を扱うために、ランダムサンプリング法を導入した独自の枠組みを開発した[7]（**図２**）。本手法を出芽酵母の685個のプロモーター領域に関する97種類のTFBSデータに適用したところ、2体間TFの相関だけに着目していた従来法では、偽陽性だけではなく、間接的な相互作用を直接的として検出している例が多く見つかった。さらに、従来法では検出できなかった新しい直接的な相互作用も発見し、そのいくつかは既に個別の生化学実験による検証が報告されていた[7, 8]。

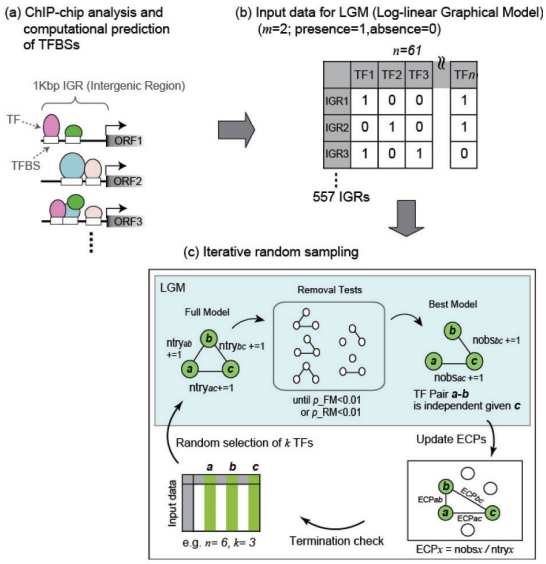

図２　転写因子間相互作用様式の推定

組織特異的発現遺伝子のプロモーター領域の数理モデル化の研究としては、まず、ホヤと線虫における筋肉特異的遺伝子のプロモーター領域をマルコフ連鎖によってモデル化し、交差検定法などで、従来法以上の予測精度が得られることを確認した[9]。ホヤについては、得られたモデルをゲノムの全遺伝子上流配列に適用することで新たな筋肉特異的発現遺伝子を探索した。上位にあるものをいくつか実験的に確認していただいたところ、４つのうち３つについては筋肉での発現が確認された。しかし、他の二つの組織については、全く実験で確認できなかった（日下部、私信）。そこで、別の表現法として、たとえば「エレメントAの上流－150～－100bpの位置にエレメントBが存在する」などのルールを比較的少数探索的に組み合わせることで、線虫の筋肉特異的プロモーターのモデル化を行い、さらに高い予測成績を得た[10]。さらに、同様の方法をマイクロアレイ実験などで得られた大規模なヒト・マウスの組織特異的発現遺伝子群のデータに適用した結果、組織によっ

てモデル化の難易度に差があった[11]。得られたモデルは基本的にヒト・マウス間でも保存され、また関連した組織の解釈にも適用可能な傾向が見られた。

2.3 プロモーター活性の網羅的データ解析に基づく工学的視点

HEK293細胞に対して、以前網羅的なin vitro実験によって得たプロモーター活性データの他に、TSS Seq解析を行い、転写開始点タグを1000万収集し、転写開始頻度としてのプロモーター活性のデータを得た[12, 13]。このデータに基づき、プロモーター活性の予測問題とプロモーター配列の設計問題を研究した。ここでは、TFBSの組みとしてプロモーターを捉えた。

前者の問題について、TFBSを独立変数としてプロモーター活性を説明する回帰式を導出した。そこでまず、TRANSFACのプロファイル検索により、プロモーターに潜むTFBSを同定した。こうして得た回帰式で、プロモーター活性の実測値と予測値の間の定性的な相関は確認できた。一方、HEK293細胞では発現が認められないが予測値は高い領域では、RNAポリメラーゼIIのChIP Seq解析から、RNAポリメラーゼIIの結合が数多く観察された。これにより、実際の細胞での転写開始反応では、さらに検討すべき要素のあることが示唆された。

定量的なプロモーター活性の予測を実現するために、系統フットプリント法と階層的クラスタリングによりTFBSの同定を見直すとともに、HEK293細胞におけるTFBSの機能性を考察した。すなわち、各TFBSの有無がプロモーター活性に与える影響を統計的に評価して、HEK293細胞で機能しているTFBSを同定した。これらのTFBSを独立変数にした回帰式によるプロモーター活性の予測値は、実測値との間に十分強い相関（交差検定で$R = 0.70$）を示した[8]。

後者の問題を、ここでは前者の問題の逆問題として捉えた。すなわち、与えられたプロモーターの活性を予測する手法を用い、望みの活性を示すプロモーター（TFBSの組み）を探索的に生成する。そのために、独自の多目的最適化遺伝的アルゴリズムを開発した。HEK293細胞と人工的に生成したプロモーター活性データを用い、この手法に様々な活性条件を与えたところ、既知のTFBSの組み合わせに加え、活性条件を満たす新しいTFBSの組み合わせが探索された。

2.4 その他の研究

上述の３つの柱の研究の過程で様々な副次的成果が得られた[14-21]。たとえば、代表的なモチーフ表現法である重み行列法において、体系的に条件検討を行うことで、最適疑似度数の性質を調べたところ、いくつか興味深い性質を見いだした[14]。また、ここ２、３年の間で網羅的なヌクレオソームの位置情報や、DNA/ヒストンのメチル化情報等が得られるようになったため、それらを使ったエピジェネティックな効果の研究にも着手した。その結果、ヌクレオソームの位置を予測するプログラムの精度は生物種によってかなり違いがあること[15]、ヒトゲノムではAlu配列がヌクレオソーム配置に少なからず影響していること、複数のヌクレオソーム位置情報を組み合わせて、ヌクレオソームの再配置などの動的性質も読み取れることなど、興味深い知見が得られている。

3. まとめ

上述のように、プロモーター領域とよばれる転写制御領域の構築原理について、さまざまな知見が得られた。これを一般原理にまで昇華できるかどうかはまだわからないが、いわゆるエピジェネティックな効果を考えないで、制御領域をシスエレメント（TFBS）の一次元的な配置によって記載するモデルでも、転写制御を理解する上で十分有望だとは言えるかもしれない。今後、より多く

の網羅的データを取り入れて、より精密なモデル化を進めていくのがやはり王道で、そこにいかにエピジェネティクス的効果を組み込んでいくかが今後の課題であると考えられる。

4. 研究開発実施体制

代表研究者　中井　謙太（東京大学医科学研究所）

研究開発題目

(1) プロモーターの比較解析
　グループリーダー　中井　謙太（東京大学医科学研究所）

(2) プロモーター構造のモデル化
　グループリーダー　矢田　哲士（京都大学大学院情報学研究科）

(3) 人工プロモーターの解析
　グループリーダー　鈴木　穣（東京大学大学院新領域創成科学研究科）

5. 参考文献

[1] Chiba, H., Yamashita, R., Kinoshita, K., and Nakai, K. Weak correlation between sequence conservation in promoter regions and in protein-coding regions of human-mouse orthologous gene pairs, *BMC Genomics*, **9**:152, 2008

[2] Tsuritani, K., Irie, T., Yamashita, R., Wakaguri, H., Kanai, A., Mizushima-Sugano, J., Sugano, S., Nakai, K., and Suzuki, Y., Distinct class of putative "non-conserved" promoters in humans: comparative studies of alternative promoters of human and mouse genes, *Genome Res.*, **17**(7): 1005-1014, 2007

[3] Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C., Huang, T.H., The functional consequences of alternative promoter use in mammalian genomes, *Trends Genet.* 24(4):167-77, 2008 (Review)

[4] Wakaguri, H., Yamashita, R., Suzuki, Y., Sugano, S., and Nakai, K. DBTSS: DataBase of Transcription Start Sites, progress report 2008, *Nucl. Acids Res.*, **36**: D97-D101, 2008

[5] Yamashita, R., Wakaguri, H., Sugano, S., Suzuki, Y., and Nakai, K. DBTSS provides a tissue specific dynamic view of Transcription Start Sites, *Nucl. Acids Res.*, in press.

[6] Okamura, K., and Nakai, K. Retrotransposition as a source of new promoters, *Mol. Biol. Evol.*, **25**(6):1231-1238, 2008

[7] Park, S.J., Ichinose, N., and Yada, T. Probabilistic Graphical Modeling for Large-scale Combinatorial Regulation of Transcription Factors, *Proc. the workshop on Knowledge, Language, and Learning in Bioinformatics (KLLBI)*, 72-86, 2008

[8] Park, S.J., Ichinose, N., and Yada, T. Inferring Probabilistic Conditional Independency from Large-scale Combinatorial Regulation of Transcription Factors, *Journal of Software*, in press, 2009

[9] Vandenbon, A., Miyamoto, Y., Takimoto, N., Kusakabe, T., Nakai, K. Markov chain-based promoter structure modeling for tissue-specific expression pattern prediction, *DNA Res.*, **15**(1) :3-11, 2008

[10] Vandenbon, A. and Nakai, K. Using simple rules on presence and positioning of motifs for promoter structure modeling and tissue-specific expression prediction, *Proceedings of the*

19th International Conference (Genome Informatics Series Vol.21):189-199, 2008

[11] Vandenbon, A. and Nakai, K. Modeling tissue-specific structural patterns in human and mouse promoters, *Nucl. Acids Res.*, in press.

[12] Sakakibara, Y., Irie, T., Suzuki, Y., Yamashita, R., Wakaguri, H., Kanai, A., Chiba, J., Takagi, T., Mizushima-Sugano, J., Hashimoto, S., Nakai, K., and Sugano, S., Intrinsic promoter activities of primary DNA sequences in the human genome, *DNA Res.*, **14**(2): 71-77, 2007

[13] Tsuchihara, K., Suzuki, Y., Wakaguri, H., Irie, T., Tanimoto, K., Hashimoto, S., Matsushima, K., Mizushima-Sugano, J., Yamashita, R., Nakai, K., Bentley, D., Esumi, H., and Sugano, S. Massive transcriptional start site analysis of human genes in hypoxia cells, *Nucl. Acids Res.*, **37**: 2249-2263, 2009

[14] Nishida, K., Frith, M., and Nakai, K. Pseudocounts for Transcription Factor Binding Sites, *Nucl. Acids. Res.*, **37**:939-944, 2009

[15] Tanaka, Y., and Nakai, K. An assessment of prediction algorithms for nucleosome positioning, *Proc. 20th Int. Conf. Genome Informatics (GIW2009)*, in press.

[16] Okumura, T., Makiguchi, H., Makita, Y., Yamashita, R., and Nakai, K. Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions, *Nucl. Acids Res.*, **35**:W227-W231, 2007

[17] Uno, Y., Suzuki, Y., Wakaguri, H., Sakamoto, Y., Sano, H., Osada, N., Hashimoto, K., Sugano, S., Inoue, I. Expressed sequence tags from cynomolgus monkey (*Macaca fascicularis*) liver: a systematic identification of drug-metabolizing enzymes, *FEBS Lett.* **582**(2):351-358, 2008

[18] Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information, **36**: D93-D96,2008

[19] Yamashita, R., Suzuki, Y., Takeuchi, N., Wakaguri, H., Ueda, T., Sugano, S., and Nakai, K. Comprehensive detection of human terminal oligo-pyrimidine (TOP) genes and analysis of their characteristics, *Nucl. Acids Res.*, **36**(11):3707-3715, 2008

[20] Murakami, K., Imanishi, T., Gojobori, T., and Nakai, K. Two different classes of co-occurring motif pairs found by a novel visualization method in human promoter regions, *BMC Genomics*, **9**:112, 2008

[21] Sierro, N., Li, S., Suzuki, Y., Yamashita, R., and Nakai, K. Spatial and temporal preferences for trans-splicing in Ciona intestinalis revealed by EST-based gene expression analysis, *Gene*, **430**(1-2):44-49, 2008