

論文内容の要旨

論文題目: Anatomy of Transcription Start Site Distribution
(転写開始点の分布の解析)

氏名: モハammed バドルル アーサン

1.概要

超並列処理配列決定法による超高速次世代型シーケンサーが登場し、生物の特定な細胞や組織における殆どの転写産物を網羅的に明らかにできる時代がやってきた。次世代型シーケンサーは低コストで高速に大量の塩基配列を決定できることから DNA 塩基配列解析に革命をもたらした。遺伝子の転写時に基礎転写因子が転写開始点の周りに集中して遺伝子の転写を誘導している。よって、転写産物の研究には転写開始点の情報はとても重要である。近年、転写開始点の解析に東京大学医学研究科の橋本博士らによる開発された 5'SAGE(Serial Analysis of Gene Expression)法 はとても重要な役割を果たしてきている。5'SAGE を使って mRNA(メッセンジャーRNA)の転写開始点から下流に数十 DNA 塩基を切断 (以下 5'end tag と呼ぶ) する。5'SAGE 法を少し改良し、次世代型シーケンサーの一つである Illumina/Solexa に適用できる新方法も橋本博士らによって開発された。新方法では転写開始点を含んだ遺伝子全体の発現量を網羅的に、しかも幅広いダイナミックレンジで測定できるようになった。今まで曖昧にされてきた低頻度の発現遺伝子の変化も正確に測定可能になった。技術革命により、いままでと違った現象を詳細に観測できるようになってきた。たとえば、遺伝子の転写開始点はゲノム上にひとつ決まると単純に考えられてきたが、網羅的なデータの取得によって遺伝子の転写開始点は多数存在すること、明らかになりつつある。我々はこの手法をもちいて、キイロショウジョウバエ (*Drosophila melanogaster*) の 7つのサンプル (胚、幼虫、未成熟/成熟・雄/雌、培養細胞 S2) から 5'

末端配列を収集し、2,500万タグのゲノム上での位置を決めた(3塩基までのミスマッチを許容)。このデータは以下のような分析結果から、ショウジョウバエの分野において貴重な資源になると思われる。(1) 転写開始点は唯一つではなくある程度の幅をもって分布し、異なるサンプル間で分布は類似している場合もあれば、異なっている場合もある。転写開始点の分布を定式化して解析する。(2) 一方、サンプル間で分布が著しく異なり、選択的 promoter を示唆する例も多い。(3) 基礎転写因子(TATA, INR, DPE, MTE 等)が結合する promoter と転写開始点の分布との関係やこれらの転写因子が結合しない promoter を明確にすることを上記のサンプルデータから同定することができた。

2. 結果

2.1 転写開始点の分布

転写開始点の上流のゲノム配列(promoter配列と呼ぶ)が転写に影響することが多いから、転写開始点の分布状態が上流の promoter の状況を表すと考えられる。哺乳類(ヒトとマウス)の転写開始点の解析では、転写開始点の分布を“広い”や“狭い”種類に分類されている。哺乳類では分類された転写開始点の分布の種類ごとに promoter 配列の特徴や基礎転写因子との関連は調べられた。“広い”転写開始点の分布の場合は promoter 配列には CpG の割合が高いが“狭い”転写開始点の分布の場合は promoter 配列には TATA-box モチーフの存在率が高いと報告された。しかし、転写開始点の分布を定量的に比較解析することは本分類による方法では困難である。よって、転写開始点の分布の状態を表す指標が必要である。最近、ショウジョウバエの転写開始点の分布を Shape Index(SI)を使って、定量化されている。SI 指標では、 $SI > -0.1$ の転写回転の分布を“狭い” promoter、それ以外の場合は“広い” promoter として分類される。しかし、SI 指標では転写開始点の分布に関係する重要な特徴を考慮されてないため、実際に違っている転写開始点の分布に同じ指標数を与えてしまう。よって、SI 指数を使って転写開始点の分布を定量的に場合分けしても違う転写開始点分布を同じ種類として分類されてしまう。これまでの方法では転写開始点の分布を効率的に表現できないため、本研究では Aggregated Index(集約指標)を提案した。Aggregated Index(AI)では、各転写開始点とその転写量の分散や転写開始点のゲノム上の空間的な情報を考慮して転写開始の分布を定量化した。本指標は“0”から“1”までの値を持つ。“1”に近いほど、転写開始点の分布は一点に集中していて、“0”に近いほど、転写開始点の分布が広く離れていることになる。AI は今までに使用されている他の方法より転写開始点の分布を効率的に定量化することができた。AI を使って以下の解析ができた。

2.2 選択的プロモーター

ショウジョウバエの既知の遺伝子の転写開始点の近くにいくつかの転写開始点分布が存在することがあることから、これの転写開始点の分布のことを選択的 promoter として考え

られる。選択的プロモータはヒトなどで報告されているがショウジョウバエの場合は個々の遺伝子のレベルでいくつか見つかっている。ショウジョウバエの全ゲノムにどのぐらい選択的プロモータが存在しているかはまだ解明されていない。よって、既知の遺伝子の転写開始点から上下流 1000 塩基以上離れたゲノム領域に AI 値が 0.4 以上の転写開始点分布だけを選択的 promoter として定義する。最後に 300 塩基離れているクラスタだけを残して、選択的 promoter 持つ遺伝子候補を選ぶ。以上の条件により、ショウジョウバエの全ゲノム内に 894 個の遺伝子が選択的 promoter を持つことが予測された。選択的 promoter を持つ可能性が高い遺伝子の Gene Ontology 調べると、多くの遺伝子が "binding GTP", "phosphorylation", "kinase" などのシグナル伝達に関係していることがわかった。よって、これらの遺伝子が選択的 promoter によって、異なる転写産物を生成し、違ったタンパク質を翻訳し、遺伝子発現の制御の変化を引き起こす可能性が高いと考えられる。

2.3 基礎 Promoter モチーフ

転写開始点の近辺のゲノム領域は転写メカニズムに緊密な関係を持つ。特に、ショウジョウバエの場合、転写開始点の周りの基礎 promoter と呼ばれる領域（転写開始点の上流 50bp から下流の 50bp）に TATA-box, DPE, Inr と MTE モチーフ（転写因子を結合する領域）がよく存在すると報告されている。これらのモチーフに基礎転写因子が結合することで、遺伝子の転写が開始されると言われている。これらの転写因子結合領域を元に転写のモデルも提唱されている。今回の大量転写開始点データにより、決定された転写開始点の分布と基礎 promoter 領域に上記のモチーフが存在するかどうかを調べた。驚くべきことに、約 14% の promoter 領域に TATA-box, DPE, Inr と MTE 内どれか一つ以上のモチーフが存在することが分かった。多くの promoter 領域に上記のモチーフは発見されなかった。これは、ショウジョウバエの転写に関連する基礎転写因子の多くはまだ見つかっていないことを示している。これらの基礎転写因子を解明して、転写のメカニズムを再定義する必要がある。発見されたモチーフとそのプロモータの AI 値との関係を見ることにした。上記のどの基礎転写因子があっても AI 値の分布が互いによく類似している。そのほとんどの転写開始点の分布の AI 値は 0.5 以上にあるものが多い。

3. 結論

本論文では初めて、転写開始点分布を定量的に解析するための集約指標を提唱した。本指標を利用して、ショウジョウバエのゲノム中で選択的なプロモータを持つ遺伝子の候補を特定した。ほとんどの転写開始点では、基礎 promoter の中でどの転写因子の組み合わせが結合するか解明されておらず、今後のショウジョウバエゲノム研究の重要な課題として残っている。今後さらに多様なサンプルから大量の転写情報を収集できれば、本研究で提案した手法を用いてさらに大規模な解析を行うことで、遺伝子全体の転写メカニズムが明らかになると期待される。