様式 C-19

科学研究費補助金研究成果報告書

平成　22 年　5 月　14 日現在

研究種目：若手研究（B）
研究期間：　2007 ～2009
課題番号：19700275
研究課題名（和文）発現構造比較による網羅的なトランスクリプトーム解析手法の開発
研究課題名（英文）A method for exhaustive similarity search of genomic positional expression.

研究代表者
瀬尾　茂人（SENO SHIGETO）
大阪大学・大学院情報科学研究科・助教
研究者番号：　30432462

研究成果の概要（和文）：

CAGE（Cap Analysis Gene Expression）法によって得られた発現プロファイルに基づくトランスクリプトーム解析手法に関する研究を行った．このデータを用いることで発現プロファイルとゲノム上の位置を関連付けて解析することができるのが利点である．本研究では，ゲノム上での位置と遺伝子の発現量の示すパターンの類似性を網羅的に探索する手法や，さらにこれを応用し，最尤推定アルゴリズムによって各遺伝子の組織特異的な選択的スプライシングを解析する手法を開発した．

研究成果の概要（英文）：

I carried out analyses of gene expression patterns using the cap analysis gene expression (CAGE) method for exploring systematic views of tanscriptomes. Counting the number of mapped CAGE tags for fixed-length regions allows determination of the genomic expression levels. I developed a novel algorithm which provides a novel view to the genome from the genomic positional expression, and also developed a method based on graph algorithm and statistical approach for discovering tissue-specific regulation of alternative transcripts through a genome-wide analysis.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,000,000	0	1,000,000
2008 年度	1,300,000	390,000	1,690,000
2009 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,200,000	660,000	3,860,000

研究分野：生体生命情報学
科研費の分科・細目：バイオインフォマティクス
キーワード：生物情報学，類似構造探索，トランスクリプトーム解析，発現プロファイル，ゲノム

１．研究開始当初の背景

ヒトやマウスをはじめ，多くの生物の全ゲノム配列が判読されている．しかしながら遺伝子の発現の多様性や制御ネットワークの複雑さは予想されていた以上であり，ゲノム研究の次のステージであるトランスクリプトーム解析の重要性が高まっている．世界各国の研究機関により塩基配列等のゲノム構造に関わる基盤的データが体系的に蓄積整備されつつあるなかで，日本でもゲノム研究は本格的にトランスクリプトーム解析の時代へと突入している．独立行政法人理化学研究所を主体とした国際コンソーシアム「FANTOM3」プロジェクトや，平成 16 年度から文部科学省によって開始されたゲノムネットワークコンソーシアムである．申請者も参加しているこれらのプロジェクトでは，今後のポストゲノム研究の発展を目指して，国際レベルにある研究ポテンシャルを活用しつつ，遺伝子の発現調節機能やタンパク質等の生体分子間の相互作用の網羅的な解析に基づき，生命活動を成立させているネットワークを明らかにすることを目的としている．

申請者は上述の「FANTOM3」プロジェクトにおけるサテライト研究の一つとして，発現プロファイルとゲノム上の位置を関連付けた上でその構造の類似性を網羅的に探索する手法を提案し，PLoS Genetics 誌で報告した．遺伝子の転写制御・発現調節メカニズムがいかに多様であるとはいえ，その転写制御・発現調節メカニズムが類似した遺伝子は存在し，このような遺伝子の発現プロファイル間にはなんらかの形で共通するパターンが見られる可能性があるからである．この研究の成果としては，真核生物においても協調して働く遺伝子がゲノム上で近傍に存在することを示唆するものや，アンチセンス RNA が転写干渉現象に似た仕組みによって転写の制御を行うという仮説を裏付ける結果が得られた．これらが本申請課題の背景となる．

２．研究の目的

本研究の目的は，異種生物間での比較によって，普遍的な現象と特定の生物にのみ存在する現象を分離・同定することであり，そこから転写干渉現象やゲノムインプリンティング等，現在未だ謎の多い転写制御・発現調節メカニズムを理解するための有益な知見を得ることである．本研究では，従来の発現量のみに注目した発現プロファイルに対し，各転写開始点とその発現量を対応付けたトランスクリプトーム解析の新規手法，「発現構造」を導入し研究を推進する．

(1)発現構造比較によるトランスクリプトーム解析方法の開発

本研究における特色・独創的な点の最も大きなものは「発現構造」を用いることにある．発現構造は申請者の提案した構造であり，その網羅的な類似性検索は行われていなかった．従来の発現プロファイルは各遺伝子の発現量のみの集積データである．つまり各遺伝子が「どれだけ」発現しているかを知ることが出来る．これに加えて，発現構造は遺伝子の各転写開始点からの発現量をゲノムにマッピングすることで，「どのように」発現しているかも知ることが出来る．図１は発現構造の例である．

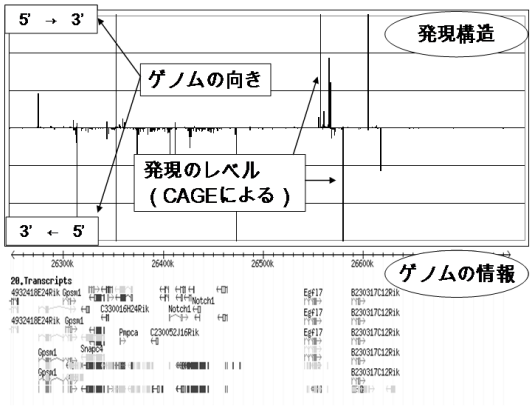

図１：発現構造

グラフ中央より上下に伸びる縦の棒が各転写開始点における転写のレベルを表している．ゲノムは２重鎖になっているため，上側が順方向の，下側が逆方向に対応している．発現構造は独立行政法人理化学研究所の開発した CAGE(Cap Analysis Gene Expression)のデータ（図２）から構成され，これは発現量と正確な転写開始点が得られる，塩基配列決定技術を基本としたトランスクリプトーム解析の手法である．この発現構造間の類似性探索，異種生物間での分類比較による共通パターンの抽出により，転写干渉現象やゲノ

ムインプリンティングなど，現在未だ謎の多い転写制御・発現調節メカニズムに関しても有益な知見を得られる可能性がある．

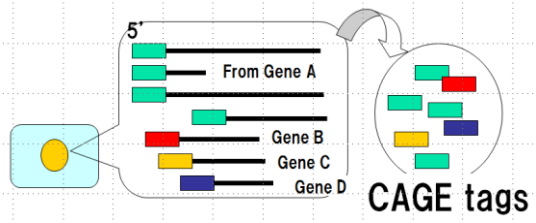

図 2：CAGE (Cap analysis gene expression)

(2)発現構造を軸とした既知の知見・情報との統合方法の開発

ポストゲノム研究においては ncRNA や Sense-Antisense 鎖，タンパク質間相互作用の研究など，新しい仮説を検証するための新規の測定技術が続々と開発されている．そのため，これらの新規データと従来の手法によって蓄積されてきた大量のデータ・知見を結びつける方法の開発が急務である．本研究は，CAGE データからの情報科学的なアプローチによるデータマイニングを縦軸とする一方で，ゲノム配列へのマッピングを通じた既知の知見・過去に蓄積されたデータとの統合という横軸的な側面を持つため，これらの媒介となることが期待される．特に，本研究成果は，今後大きな発展が望まれるタイリングアレイ技術とは密接に関連することが予想され，比較検証により相互の発展が見込まれる．さらに，CAGE 自体が新しい技術であり，現在はヒトとマウスに対して適用したデータのみが存在する．転写開始点の解析はトランスクリプトーム解析において重要な位置を占めており，転写開始点と発現量を同時に調べることができる CAGE の有用性は非常に高い．そのため，今後 CAGE によるさまざまな生物種に対する実験が行われることが予想でき，本研究成果を新規データに適用することにより，迅速かつ詳細な解析を行うことが出来ると期待される．加えて，より多様な生物種間で共通する，より普遍的なパターンが発見できれば，生命現象の解明の一助となると考える．

３．研究の方法

本研究では前述の目的を達成するため，発現構造比較によるトランスクリプトーム解析方法として(1)ランレングス符号化と suffix tree による高速な発現構造比較手法の開発と，発現構造を軸とした既知の知見・情報との統合方法のために(2)CAGE による選択的スプライシングの解析手法の開発，(3)発現構造による転写制御の組織特異性比較手法の開発を行った．

(1)suffix treeを用いた高速な網羅的発現構造比較手法の開発

CAGEを用いたゲノム上の発現構造の網羅的類似性探索手法の開発を行った。この手法は，次の２つの段階に分けられる．まず転写開始点間の距離に注目し，ゲノム上でCAGEタグがマッピングされた部分とそうでない部分をランレングス法により符号化する．そして，この符号化された系列に対して接尾辞木（suffix tree）を利用することで，ゲノム上における共通発現パターンを高速に抽出する．提案手法の評価実験として，マウスのCAGEライブラリから得られたデータに適用し，考察を行った．

(2)CAGEを用いた選択的スプライシングの解析手法の開発

選択的スプライシングは同じ遺伝子から複数のバリエーションのタンパク質（アイソフォーム）が生成される現象のことで，この現象は生物の複雑性が増すにつれて頻繁に見られるようになり，マウスでは65%以上の遺伝子で観測されている．また，発生段階や組織など環境に応じて，時間的・空間的に選択的スプライシングを制御することによってアイソフォームを作り分けている例が知られており，発現の多様性や制御ネットワークのメカニズムを解明する上で重要であると考えられる．本手法では，まずESTやcDNAなどの情報を用いて，各遺伝子のスプライシングの構造をグラフによってモデル化する．次にCAGEの情報を付加して最尤推定アルゴリズムを用いることで，各遺伝子の組織特異的な選択的スプライシングの推定を行った（図3）．またこの手法をマウスゲノムに対して適用した．

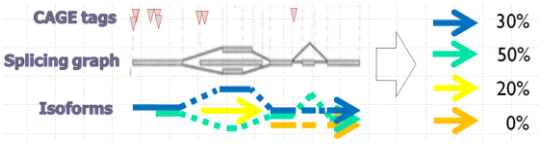

図 3：CAGE による選択的スプライシングの解析

(3) 発現構造による転写制御の組織特異性比較手法の開発

一つの遺伝子には複数個の転写開始点が存在することが知られており，CAGE 法により遺伝子の転写開始点を網羅的に同定することが可能となった．本研究では，発現構造を用いることにより，遺伝子単位の発現量にはほとんど違いが現れない転写制御の違いを解析するための方法を開発した．

発現構造のパターンと転写制御の関連を明らかにするために，発現構造に基づいて遺伝子を分類し，任意の数のクラスタを生成する手法を提案した．大域アライメントを用いて発現構造の距離を定義し，階層型クラスタリングを行う．これにより，発現構造に基づいて分類された遺伝子のクラスタが生成される．発現構造のクラスタと転写制御因子を対応づけることにより，発現構造のパターンに関係する転写制御因子を抽出する．

４．研究成果

CAGE (Cap Analysis Gene Expression) 法によって得られた発現プロファイルに基づくトランスクリプトーム解析手法に関する研究を行った．遺伝子の転写制御・発現調節メカニズムがいかに多様であるとはいえ，転写制御・発現調節メカニズムが類似した遺伝子は存在しており，そのような遺伝子の発現プロファイル間にはなんらかの形で共通するパターンが見られる可能性がある．従来の発現量のみに注目した発現プロファイルに対し，CAGE では発現量に加え正確な転写開始点の情報が得られるため，このデータを用いることで発現プロファイルとゲノム上の位置を関連付けて解析することができるのが利点である．本研究では，ゲノム上での位置と遺伝子の発現量の示すパターンの類似性を網羅的に探索する手法や，さらにこれを応用し，最尤推定アルゴリズムによって各遺伝子の組織特異的な選択的スプライシングを解析する手法，発現構造の持つパターンと転写開始点のパターンを対応付けて解析する手法を開発し，新たな知見を得た．

現在，次世代シークエンサーと呼ばれる高速な配列読み取り技術によって，ゲノム配列や全トランスクリプトームを配列断片として大量に読み取る方法が様々な生物学分野に浸透しており，従来を大幅に上回るペースでデータを産み出している．その膨大なデータは NCBI の Short Read Archive (SRA) や EBI の European Nucleotide Archive (ENA) にアーカイブされつつあり，本研究による「位置」と「量」を同時に解析する方法は，これらのデータを二次解析するための方法として有意義な成果であると考えられる．

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 3 件）

1. Tatsuya Yoshikawa, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, ″Improved prediction method for protein interactions using both structural and functional characteristics of proteins″, IPSJ Transactions on Bioinformatics, 査読有, vol. 3, pp10-23, 2010.
2. Gen Kawamura, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, ″A Combination Method of the Tanimoto Coefficient and Proximity Measure of Random Forest for Compound Activity Prediction″, IPSJ Transactions on Bioinformatics, Vol.49 No.SIG5 pp.46-57, 2008.
3. Yoshiyuki Kido, Shigeto Seno, Susumu Date, Yoichi Takenaka, Hideo Matsuda, ″A Distributed-Processing System for Accelerating Biological Research using Data-Staging″, IPSJ Transactions on Bioinformatics, Vol. 49, No.SIG5, pp.58-64, 2008.

〔学会発表〕（計 15 件）

1. 石川 元一, 瀬尾 茂人, 竹中 要一, 松田 秀雄, An approach for efficient computational analysis of massive RNA-Seq data, 第32回日本分子生物学会年会, December 11, 2009, パシフィコ横浜.
2. Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, ″A method for analysis of tissue-specific alternative transcripts using CAGE tags″, Intelligent Systems for Molecular Biology (ISMB2008), July 21, 2008, Toronto, Canada.
3. Mitsuru Jikeya, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, ″A Clustering Method for Expression Patterns of Transcription Starting Sites″, Intelligent Systems for Molecular Biology (ISMB2008), July 21, 2008, Toronto, Canada.

4. 寺家谷 満, 瀬尾 茂人, 竹中要一, 松田 秀雄, CAGE発現プロファイルを用いた転写開始点の発現パターンのクラスタリング手法, 日本分子生物学会第8回春季シンポジウム, 2008年5月26日, 京王プラザホテル札幌
5. 瀬尾茂人, 竹中要一, 松田秀雄, グラフ構造を用いた組織特異的な選択的スプライシングの解析手法, 日本分子生物学会第30回大会, 2007年12月11日, パシフィコ横浜.

６．研究組織

(1)研究代表者

瀬尾　茂人（SENO SHIGETO）

大阪大学・大学院情報科学研究科・助教

研究者番号：30432462