様式 C -19

科学研究費補助金研究成果報告書

平成 21 年　4 月　6 日現在

研究種目：若手研究（B）
研究期間：2007 -2008
課題番号：19700273
研究課題名（和文）　全生物種のトランスクリプトーム解析を加速する HiCEP データ高速解析手法の開発
研究課題名（英文）　Development of a method for high -throughput analysis of HiCEP data

研究代表者
門田　幸二（KADOTA KOJI）
東京大学・大学院農学生命科学研究科・特任助教
研究者番号：60392221

研究成果の概要：HiCEP 法は、マイクロアレイなど他の技術では不可能であった未知遺伝子を含む転写物の網羅的単離およびプロファイリングが全ての生物で可能な唯一の実験技術であるが、一次元電気泳動（1-DE）パターン比較に基づく発現プロファイル解析法であるため、データ解析に多くの手間と時間がかかっていた。本研究課題によって開発された手法の適用により、データ解析の大部分において自動化が達成できた。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	500,000	0	500,000
２００８年度	500,000	150,000	650,000
年度			
年度			
年度			
総　計	1,000,000	150,000	1,150,000

研究分野：総合領域
科研費の分科・細目：情報学・生体生命情報学
キーワード：バイオインフォマティクス

１．研究開始当初の背景

HiCEP 法は、マイクロアレイなど他の技術では不可能であった未知遺伝子を含む転写物の網羅的単離およびプロファイリングが全ての生物で可能な唯一の実験技術である。その技術はほぼ確立しており、農学・環境・医学など様々な分野で進められている主に発現変動遺伝子（転写物）の同定を目的としたポストゲノム研究を推進するための日本発の基盤技術としての普及が期待されている。しかしながら HiCEP 法は、実験データ取得後の解析が労働集約型の一次元電気泳動（1-DE）パターン比較に基づく発現プロファイル解析に属するため、他の 1-DE 解析技術である増幅制限酵素断片長多型（AFLP）や Differential Display（DD）などと同様、データ解析の高速化が大きな課題となっていた。

２．研究の目的

本研究では、これまで労働集約型で行われてきた HiCEP データ解析の完全自動化（高速化）を目指すという研究の全体構想の中で、下記項目を具体的な目的として研究を行った。

(1) 比較するサンプル間での同一ピーク認識（アラインメント）精度向上。
フラグメント長を補正し、その後にピークア

ラインメントアルゴリズムを適用することで同一ピーク認識精度の向上を目指した。

(2) 正しくアラインメントされたピーク群の正規化および発現変動ピーク同定。
HiCEP 法は 1-DE パターン比較に基づく発現プロファイル解析法と同様、最終的な判断が視覚的な評価に頼らざるを得ないことから、専用の発現変動ピーク同定手法の開発を目指し、視覚による評価との不一致のない合理的な発現変動ピークのランキング手法の開発を行った。

３．研究の方法

上記二つの目的に対して、以下に示す方法（GOGOT 法）を採用した。

(1) 比較するサンプル間での同一ピーク認識（アラインメント）精度向上。
この目的を達成するために、GOGOT 法の要素技術として、GOGOTnormL の開発を行った。具体的には、補正を行うターゲットサンプルに対して、一定のピーク数を含む波形パターンを window 幅とした moving-window 法を採用した。また、補正のためのリファレンスパターンは、「各ピークに対して、そのフラグメント長が正しく見積もられていることを客観的に示す指標 Q」を用い、window 幅中のピークの Q 値の平均が最も高いサンプルを採用した。

(2) 正しくアラインメントされたピーク群の正規化および発現変動ピーク同定。
本研究では、クラスタリングに基づくピークアラインメント法を採用した。具体的には、比較する全サンプル中のピークのフラグメント長のデータをマージし、ピーク間のフラグメント長の差を距離としてクラスタリングを行った。クラスタリング時に、全て別サンプル由来同一ピークからなる密なクラスターが形成された段階でそのクラスターを計算から除外することでアラインメント時の矛盾の問題を解決できる。また、用いるデータは補正後のフラグメント長であり、同一ピーク間で大幅に距離が離れることは極めてまれであるため、このアラインメント法が有効であると考えた。

アラインメントされたピーク群の正規化は、“比較する全サンプル中にピークが存在し、かつ正しくアラインメントされている転写物群”を用いて総強度正規化法を適用した。便宜的にここで採用された手続きを GOGOTnormH と名づけた。ここまでの手法を適用することで得られる出力が“遺伝子発現行列”となり、一般的なマイクロアレイ解析手法を適用する際の入力データと同じ形式になるのが利点である。

HiCEP データは、最終的に得られる発現変動ピークの波形を目でみて確認する必要がある。入力データ形式は遺伝子発現行列であるので、一般的なマイクロアレイ解析手法の直接的な応用が基本的には可能である。しかしながら、マイクロアレイ解析によく用いられる t 検定に基づく方法を用いた予備的検討の結果、視覚による評価との不一致が数多く見られた。これは、一般にピークの高さが低いほど相対誤差が大きくなることに起因する。このため、全体としてシグナル強度（ピークの高さ）の高い発現変動ピークがより上位にくるような統計量 GOGOTstat を開発した。

４．研究成果

GOGOT 法の開発により、HiCEP データ解析の大部分について自動化が可能となった。ここでは、開発した要素技術の適用により得られた、具体的な成果を示す。

(1) 比較するサンプル間での同一ピーク認識（アラインメント）精度向上
GOGOTnormL 法を適用して電気泳動パターンの補正を行うことで、同一ピークのサンプル間でのずれが大幅に改善された。代表的な改善結果を図 1 に示した。(a)が補正前、そして(b)が補正後の計 10 サンプルの電気泳動パターンを示しているが、同一ピークの視覚的な評価が容易になっているのが分かる。

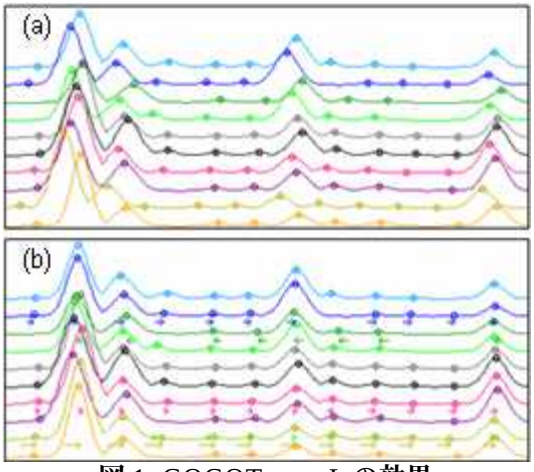


図 1. GOGOTnormL の効果

HiCEP 解析では 10 サンプル程度以上のデータを取り扱う。多サンプル間のピークアラインメントを行う最も一般的な方法は、多重配列比較に基づく方法であるが、多重配列比較時と同様の 2 つの問題（アラインメント時の矛盾や膨大な計算コスト）が起こりうる。これらの問題を一挙に解決するための具体的な工夫として、本研究ではクラスタリングに基づくピークアラインメント法を採用し

た。具体的には、比較する全サンプル中のピークのフラグメント長のデータをマージし、ピーク間のフラグメント長の差を距離としてクラスタリングを行った。クラスタリング時に、全て別サンプル由来同一ピークからなる密なクラスターが形成された段階でそのクラスターを計算から除外することでアラインメント時の矛盾の問題を解決できた。また、用いるデータは補正後のフラグメント長であり、同一ピーク間で大幅に距離が離れることは極めてまれである。このため、距離のカットオフ値を設定することで計算コストの大幅な改善を達成できた。このアイデアは 1-DE パターン間比較における初の試みであったが、GOGOTnormL 法によるフラグメント長補正後の電気泳動パターンに適用することで高速かつ視覚による評価との矛盾がないアラインメントを行うことが可能となった。

図 2 は、GOGOTnormL 法適用前後（図 1 の a と b に相当）の電気泳動パターンについてクラスタリングに基づくピークアラインメント法を適用した結果を示したものである。GOGOTnormL 法適用前のパターンに対するアラインメント結果（図 2a）には黒線で示されるミスアラインメントが含まれるが GOGOTnormL 法適用後のパターンではミスアラインメントがなくなっており、またミスアラインメントでないことの確認も容易となっていることが分かる（図 2b）。

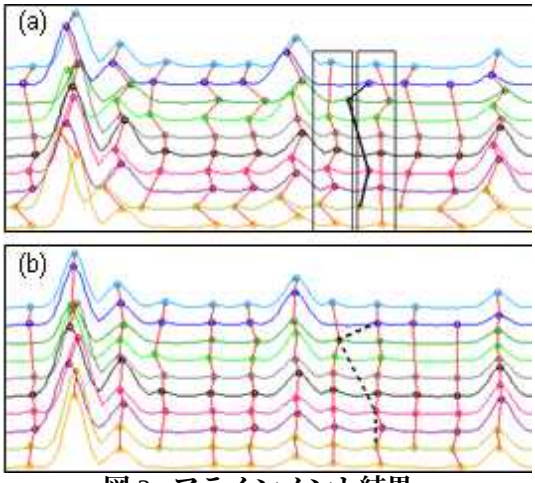


図 2. アラインメント結果

(2) 正しくアラインメントされたピーク群の正規化および発現変動ピーク同定。

ここで用いた正規化法は、マイクロアレイデータの正規化で用いられている仮定「比較するサンプル間で大部分の遺伝子発現は変化しない」と基本的に同じアイデアである。しかし、HiCEP データは数千－数万遺伝子からなるマイクロアレイと異なり、1 つの波形パターンにつき約 150 ピーク程度しかなく、その数はピークとして同定する閾値次第で変動するという特有の問題が存在する。このため、“ 比較する全サンプル中にピークが存在し、かつ正しくアラインメントされている転写物群 ” を用いて総強度正規化法（この方法を便宜的に GOGOTnormH 法と称した）を適用した（図 3）。正規化前（図 3a）に比べて正規化後（図 3b）のパターンのほうが technical replicates 間のバラツキが抑えられていることが分かる。

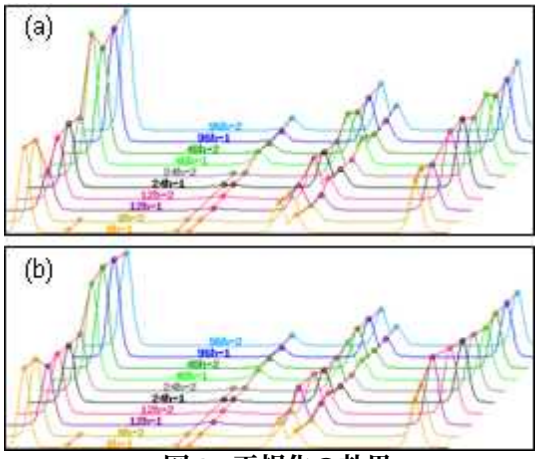


図 3. 正規化の効果

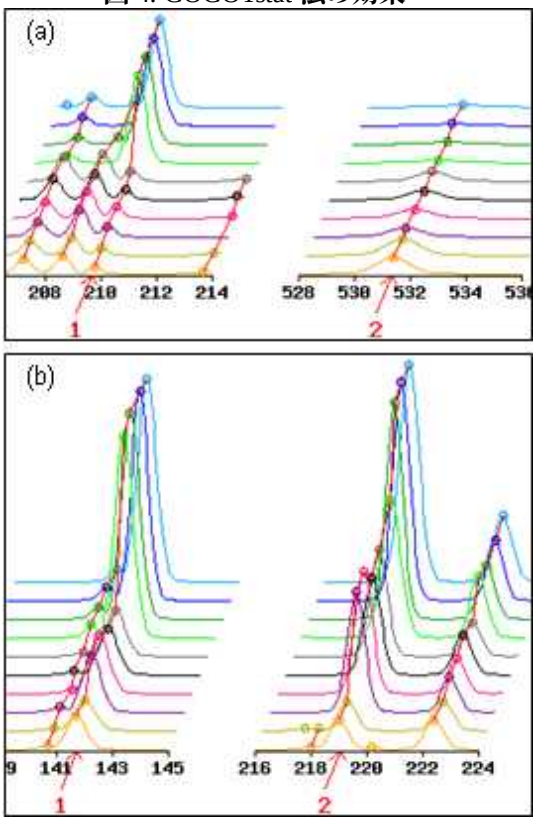
ここまでの手法（GOGOT 法；Kadota et al., Algorithms Mol. Biol., 2007）の適用によって、HiCEP の 1-DE パターンを入力として、HiCEP の遺伝子発現行列を高精度に得ることが可能となった。これは、マイクロアレイ解析分野で数多く開発されている遺伝子発現行列からのデータマイニング手法を HiCEP データにも適用可能であるということを意味する。したがって、GOGOT 法開発と平行して HiCEP データに適用可能な様々なマイクロアレイ解析手法の比較検討も行った。平成 19 年度は、組織特異的発現遺伝子を検出する二つのマイクロアレイ解析手法について比較検討を行い ROKU 法（Kadota et al., BMC Bioinformatics, 2006）が有用であることを確認することができた（Kadota et al., Gene Regulation Systems Biol., 2007）。

当初、GOGOT 法の要素技術として、GOGOTstat という発現変動の度合いでランキングする方法を独自開発し、その方法がマイクロアレイ法などで一般によく用いられる *t* 検定に基づく方法に比べ、より視覚による評価との一致度が高いことを確認していた（図 4）。しかし、その後統計量を計算するための数式が、全体的な発現シグナル強度が高いピークが上位にランキングされる効果が極端に高いものであることが判明したため、数式の改良を行った。

改良したランキング法のよしあしを評価するためには、真の発現変動ピークを含むべ

ンチマークデータセットを用いるのが理想ではあるが、HiCEP でそのようなデータセットで得ることは事実上不可能である。そこで、そのようなベンチマークデータセットを豊富に含むマイクロアレイデータを用いて、新たに開発した発現変動遺伝子（ピーク）ランキング法 WAD（Kadota et al., Algorithms Mol. Biol., 2008）の評価を行った。結果として、感度・特異度を評価基準とした場合に他の手法に比べ高い精度を示すことを確認することができた。

図 4. GOGOTstat 法の効果



以上の研究成果によって、これまで低スループットであった HiCEP データ解析の劇的な高速化に成功した。また、本研究課題で開発した発現変動遺伝子（ピーク）ランキング法は、電気泳動波形比較に基づく HiCEP データのみならず、マイクロアレイデータにも適用可能であることから、今後様々なトランスクリプトーム解析に今回開発した手法が適用されることが期待される。

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　7 件）

① Natsume Y, Kadota K, Satsu H, Shimizu M., Effect of Quercetin on the Gene Expression Profile of the Mouse Intestine, *Bioscience, Biotechnology, and Biochemistry*, **73**(3), 722-725, 2009.

② Kadota K, Nakai Y, Shimizu K., A weighted average difference method for detecting differentially expressed genes from microarray data, *Algorithms for Molecular Biology*, **3**, 8, 2008

③ Shimizu-Ibuka A, Nakai Y, Nakamori K, Morita Y, Nakajima KI, Kadota K, Watanabe H, Okubo S, Terada T, Asakura T, Misaka T, Abe K., Biochemical and Genomic Analysis of Neoculin Compared to Monocot Mannose-Binding Lectins., *Journal of Agricultural and Food Chemistry*, **56**(13), 5338-5344, 2008.

④ Nakai Y, Hashida H., Kadota K, Minami M, Shimizu K, Matsumoto I, Kato H, Abe K., Up-regulation of genes related to the ubiquitin-proteasome system in the brown adipose tissue of 24-h-fasted rats, *Bioscience, Biotechnolgy, and Biochemistry*, **72**(1), 139-148, 2008.

⑤ Kadota K, Araki R, Nakai Y, Abe M., GOGOT: a method for the identification of differentially expressed fragments from cDNA-AFLP data, *Algorithms for Molecular Biology*, **2**, 5, 2007.

⑥ Kadota K, Konishi T, Shimizu K., Evaluation of two outlier-detection-based methods for detecting tissue-selective genes from microarray data, *Gene Regulation and Systems Biology*, **1**, 9-15, 2007.

⑦ Ohkura N, Oishi K, Sakata T, Kadota K, Kasamatsu M, Fukushima N, Kurata A, Tamai Y, Shirai H, Atsumi G, Ishida N, Matsuda J, Horie S., Circadian variations in coagulation and fibrinolytic factors among four different strains of mice, *Chronobiology International*, **24**(4), 651-669, 2007.

〔学会発表〕（計　14 件）

① 成川真隆, 中井雄治, 南道子, 門田幸二, 三坂巧, 阿部啓子, 食餌性亜鉛の欠乏がラット小腸の遺伝子発現に与える影響, 日本農芸化学会2009年度大会, 2009年3月27-29日, 福岡.

② 新谷政己, 高橋裕里香, 徳丸裕樹, 門田幸二, 原啓文, 宮腰昌利, 西田洋巳, 山根久和, 野尻秀昭, pCAR1を有することが宿主である*Pseudomonas*属細菌に鉄欠乏を生じさせる, 第3回日本ゲノム微生物学会年会, 2009年3月5-7日, 東京.

③ 高橋裕里香, 新谷政己, 徳丸裕樹, 門田幸二, 原啓文, 宮腰昌利, 西田洋巳, 山根久和, 野尻秀昭, IncP-7群プラスミドpCAR1を保持する異種*Pseudomonas*属細

菌の形質およびトランスクリプトーム比較，第3回日本ゲノム微生物学会年会, 2009年3月5-7日，東京.

④ 門田幸二，マイクロアレイ解析の話：発現変動遺伝子検出あたりを中心に，日本バイオインフォマティクス学会第5回九州地域部会講習会, 2009年2月26-27日，福岡.

⑤ Kadota K., Nakai Y, Shimizu K., Comparison of methods for detecting differentially expressed genes from microarray data, 2008年日本バイオインフォマティクス学会年会, 2008年12月15-16日，大阪.

⑥ 門田幸二，中井雄治，清水謙多郎，マイクロアレイデータからの発現変動遺伝子検出法WAD，日本分子生物学会第31会年会（BMB2008）, 2008年12月9-12日，神戸.

⑦ 門田幸二，トランスクリプトーム解析手法の開発，アグリバイオインフォマティクス成果報告シンポジウム, 2008年12月8日，東京.

⑧ 中井雄治，門田幸二，ニュートリゲノミクス研究におけるDNAマイクロアレイ解析戦略，日本農芸化学会2008年度大会, 2008年3月29日，名古屋.

⑨ Nakai Y, Hashida H, Kadota K, Minami M, Shimizu K, Matsumoto I, Kato H, Abe K., Genes Related to the Ubiquitin-Proteasome System Are Up-regulated in the Brown Adipose Tissue of 24 h-Fasted Rats, The University of Tokyo International Symposium “Frontier of Microbial and Plant Biotechnology in Environmental and Life Sciences”, 2007年12月5-6日，東京.

⑩ Kadota K, Ye J, Nakai Y, Terada T, Shimizu K., A new method for detecting tissue-selective genes from microarray data and its comparative study, The University of Tokyo International Symposium “Frontier of Microbial and Plant Biotechnology in Environmental and Life Sciences” , 2007 年 12 月 5-6 日，東京.

⑪ Natsume Y, Ito S, Satsu H, Kadota K, Ohsawa K, Shimizu M., The effect of quercetin on ER stress at intestinal epithelia, 3rd International Conference on Polyphenols and Health (ICPH2007), 2007 年 11 月 25-28 日，京都.

⑫ 門田幸二, R でお手軽にアレイ解析，特定領域研究「植物ゲノム障壁」マイクロアレイワークショップ, 2007 年 10 月 22-23 日，東京.

⑬ 門田幸二, HiCEP データのマイクロアレイ様データへの変換：HiCEP もアレイ用解析手法が利用可能です，放射線医学総合研究所第 14 回重粒子医科学センター研究交流会, 2007 年 7 月 19 日，千葉.

⑭ 門田幸二，マイクロアレイ解析手法あれこれ，日本バイオインフォマティクス学会第 2 会機能ゲノミクス研究会, 2007 年 4 月 12 日，東京.

〔図書〕（計　2 件）

① 門田幸二，有意差検定とマーカー遺伝子の意義（他），*藤渕航・堀本勝久/編 実験医学別冊 マイクロアレイデータ統計解析プロトコール*，羊土社, 72-95，2008.

② 門田幸二，書評（統計解析環境 R によるバイオインフォマティクスデータ解析）”, *バイオサイエンスとインダストリー*, **65**(11), 28，2007.

〔その他〕
研究代表者ホームページ
http://www.iu.a.u-tokyo.ac.jp/~kadota/

６．研究組織

(1)研究代表者
門田　幸二（KADOTA KOJI）
東京大学・大学院農学生命科学研究科・特任助教
研究者番号：60392221

(2)研究分担者

(3)連携研究者