Vol.2011-BIO-24 No.10
2011/3/11

比較トランスクリプトーム向けマイクロアレイの設計

福崎睦美[†1] 吉田真明[†2]
小倉淳[†2] 瀬々潤[†1]

比較ゲノム解析が盛んに行われ，次なるターゲットとして比較トランスクリプトーム解析が着目されている．しかし複数種向けマイクロアレイの設計は，特異的かつ種間で結合可能なプローブの定義が明らかでないため困難だった．本研究ではマイクロアレイでの結合性質を実験により検証し，その結果から複数種の遺伝子発現を同時に観測可能なアレイの設計手法を構築した．また，実際に種共通アレイを設計し，提案手法の有用性を示した．

Microarray probe design for comparative transcriptome microarray

Mutsumi Fukuzaki,[†1] Masaaki Yoshida ,[†2]
Atsushi Ogura[†2] and Jun Sese[†1]

Comparative genomics become one of the major areas in biology, and next step is to compare gene expressions between species. Although microarray is one of the most popular gene expression experiment, it is difficult to design microarray which can measure plural species. In this study, we first check the relation of observed expression levels with nucleotide mutations. Based on the result, we then introduce probe design strategy. Finally, we develop the inter-species array and performed real microarray experiment using the inter-species probes.

†1 お茶の水女子大学 大学院人間文化創成科学研究科
Dept. of Computer Science, Ochanomizu University
†2 お茶の水女子大学 お茶大アカデミック・プロダクション
Ochadai Academic Production, Ochanomizu University

1. はじめに

近年，シークエンサの高速化によりゲノムが容易に読めるようになり，比較ゲノム研究が盛んに行われている．その一方で，ヒトとチンパンジーの遺伝子領域には 1.23%の違いしかない事がわかるなど，遺伝子配列のバリエーション以上に，種間の差異は大きい．この原因を知るため遺伝子発現量を比較する比較トランスクリプトームに注目が集まっている[1)]．比較トランスクリプトーム解析は，種間で類似している遺伝子（ホモログ）の発現頻度を網羅的に比較する研究であり，創薬研究のプロセスの短縮や発生過程の違いを生む原因の発見に役立つことからも重要であると考えられる．しかし，比較トランスクリプトームの既存研究のうち既存のマイクロアレイを用いた研究は，実験コストや観測された発現量の正確さ等に問題があった[2)3)]．マイクロアレイを独自設計する場合にも，正確な発現量を保証しつつ網羅的にプローブを設計することは困難であった[4)]．

本研究ではこの問題点を乗り越え，定量かつ網羅的な比較トランスクリプトーム解析を行うため，プローブの結合・非結合条件を実験により検証し，複数種向けマイクロアレイのプローブ設計手法を構築した．また，実データによる実験で，本手法により設計可能なプローブ数が大幅に増加し，種間で相関の高い発現観測が可能となったことを確認した．

2. マイクロアレイプローブの特異性検証実験

2.1 検証実験内容

本研究では，プローブ－遺伝子間の結合条件を明らかにするため，プローブと遺伝子間に変異があった場合，観測される発現量がどのように変化するかを実験的に検証した．検証では，遺伝子との間に変異がないプローブを使用したアレイと，故意に変異を加えたプローブを使用したアレイを用いて，観測された発現量の比較を行った．プローブは Agilent 社[⋆1] のヒト用プローブを利用したが，(1) 1 遺伝子からは 1 プローブのみ使用する，(2) プローブと他の遺伝子との編集距離の最小値が 11 以上である，という 2 つの条件を満たす 7,937 個を選択し，使用した．加える変異には，変異の数や変異位置を系統的に変化させた 51 個のパターンを作成している．サンプルは Clontech 社[⋆2]の qPCR Human Reference Total RNA とし，変異有り/無しの 2 枚のマイクロアレイ間で同じサンプルを用いた実験を行っ

⋆1 http://www.chem.agilent.com/
⋆2 http://www.clontech.com/

Vol.2011-BIO-24 No.10
2011/3/11

た．観測された発現量を比較することで，変異による観測精度の変化を検証している．観測された発現量の値は底が 2 の対数を取り，中央値が 10 になるように正規化した．以降，発現量として用いている値は正規化後の値とする．また今回の実験では，再現性の高い結果を得るために各アレイで同じサンプルによる 6 回の重複実験を行っている．変異有り/無しのプローブで比較する発現量は，6 回観測された発現量の平均とした．

2.2 変異数による発現量への影響検証

観測結果を用い，変異数が同じならば観測される発現量への影響も同程度となるか検証を行った．図 1 はそれぞれ，(A) プローブの末尾に連続した 5 個の変異，(B) プローブの末尾に連続した 7 個の変異，(C) 等間隔に配置された 5 個の変異，という変異パターンでの比較結果である．上部にある表にはプローブに加えた変異の位置を黒い四角で表している．グラフは，横軸に変異無しのプローブでの発現量をとり，縦軸は変異による発現量への影響を表している．また，各点で一つのプローブの結果を示した．縦軸の値が低ければ変異により発現が低下し，0 のライン上にあるものは変異による発現量の変化が見られないものである．変異数のみによって発現量の低下の程度が決定されるのであれば，図 1(A)(C) の比較では同程度の影響が見られ，図 1(A)(B) の比較・図 1(C)(B) の比較では変異数の多い図 1(B) の発現量低下が最も大きいはずである．しかし，図の示す通り結果は図 1(A)(B) はほとんど発現の低下がなく，2 つの結果に有意な差は見られない．一方で，図 1(C) は明らかに図 1(A)(B) よりも大きく発現量が低下している．この結果から，発現量の変化に最も大きく影響する要因は変異数ではなく，変異の入る位置であると予測できる．

2.3 変異位置による発現量への影響検証

次に，変異位置による発現量への影響を検証するため連続した長さ 3 の変異の位置を変化させ，影響の程度を比較した． 結果は図 2 の通りであり，プローブの前半に変異のある図 2(D) で大きく発現量が低下している．そして，図 2(E)(F) のようにプローブの中央から後ろの変異位置では徐々に発現量の低下が見られなくなる．このように，プローブと遺伝子が結合しなくなる要因は変異位置にあり，主にプローブの前半に入った変異の影響が大きく，後半には変異が入っていたとしても影響が小さいことがわかった．

2.4 検証実験結果のまとめ

これまでの認識とは異なり，プローブと遺伝子間の変異が発現観測に影響を与えない場合があった．また，発現量への影響を決定づける要因としては変異数よりも変異の位置がより重要であることがわかった．このため，比較トランスクリプトームに向けて種間に変異がある領域でも共通プローブを設計する場合には，変異の位置がプローブの後半部分となるようにプローブ領域を取ることで，種間で同程度の発現量を観測できると思われる．

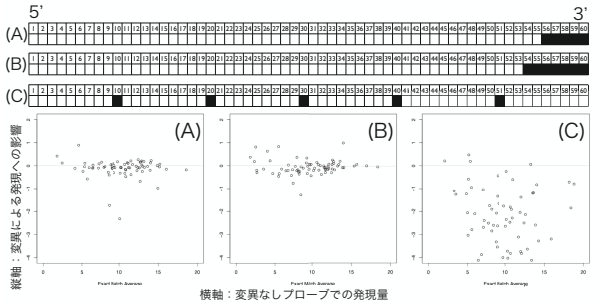


図 1 変異数と変異パターンによる影響：(A) と (B) はほとんど発現の低下がなく，(C) は明らかに大きく発現量が低下しているため，発現量への影響は変異数により決定されるわけではないことが推察できる．
Fig. 1 Effect on expression levels according to the number of mutations and mutation patterns.

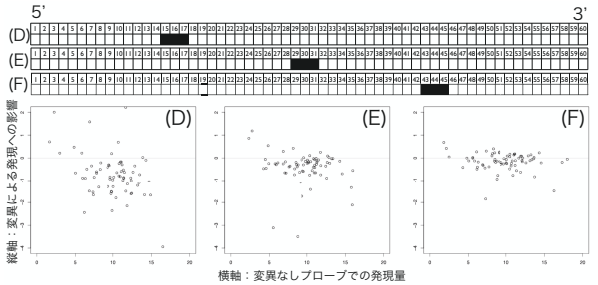


図 2 変異位置による影響：(D) のようにプローブ前半にある変異が発現量に大きな影響を与え，後半に変異を持つ (E)(F) のパターンは影響を与えない．
Fig. 2 Effect on expression levels according to mutation positions.

3. 提案手法

3.1 プローブ条件

種間共通プローブの設計を行うため，プローブ条件をまとめると，

- 一つのホモロググループ内の遺伝子が，一つのプローブに結合できること
- プローブが，比較対象全種の全遺伝子配列について，標的とするホモロググループ以外には結合しない（特異的である）こと

という二つの条件が上げられる．本研究では，これらの条件をプローブと遺伝子間の距離を用いて定義する．なお，プローブの長さは今回の実験で利用している Agilent 社のプローブ長である 60 で固定するが，任意の長さで同様の議論が可能である．

3.2 重み $w(i)$ と重み付き編集距離 $D(p,g)$ の定義

本研究ではプローブ一遺伝子間の結合・非結合条件の定義に，変異位置を基にした重み付き編集距離を用いる．配列間の類似性を測る際に用いられる編集距離の重みは一般的に挿入・削除・置換で重みを変えたり，アフィンギャップを用いる．しかし本研究では 3 章の実験結果を利用し，変異位置に着目した距離定義の方が適している．よって，変異位置に応じて重み付けした編集距離を定義した．プローブ配列 $p = p_1p_2...p_{60}$ とし，プローブ上の変異位置 i をもとにして以下のように重み $w(i)$ を定義する．

定義 1. 変異位置による重み付け

$$w(i) = \begin{cases} 10 & (i \leq 30) \\ 5 & (30 < i \leq 40) \\ 1 & (40 < i) \end{cases} \tag{1}$$

これにより，プローブ前半（プローブの 1-30 塩基目）に一カ所変異が入ると距離は 10 加算される．以降，プローブの後半にかけて重みを軽くした．この重みを用い，遺伝子を長さ n の配列 $g = g_1g_2...g_n$ としたとき，配列 p, q 間の重み付き編集距離 $D(p,g)$ は各塩基の重みを考慮して以下のように定義する．

定義 2. プローブと遺伝子間の距離

$$D(p,g) = \min\{d(p_{60}, g_{60}), d(p_{60}, g_{61}), \ldots, d(p_{60}, g_{n-1}))\} \tag{2}$$

$$d(p_i, g_j) = \min \begin{cases} 0 & (i \leq 0) \\ \sum_{k=1}^{i} w(k) & (j \leq 0) \\ d(p_{i-1}, g_j) + w(i) & \\ d(p_i, g_{j-1}) + w(i) & \\ d(p_{i-1}, g_{j-1}) + w(i) & (p_i \neq g_j) \\ d(p_{i-1}, g_{j-1}) & (p_i = g_j) \end{cases} \tag{3}$$

$D(p,q)$ は q の中で最も p に近い位置においての p と q の距離を表している．この値が小さい場合，q 内に p に極めて近い配列が存在する．一方大きい場合には q 内には p に近い配列が存在しないことが分かる．

3.3 重み付き編集距離を利用した結合条件・非結合条件の定義

本研究では重み付き編集距離を利用し，プローブ p が遺伝子 g に結合できる条件を以下のように定義した．

定義 3. 結合条件

$$D(p,g) < 10 \tag{4}$$

これにより，定義 1 で導入した重みを用いた場合，プローブ一遺伝子間で結合すると判断するとき，プローブの 1-30 文字目までが必ず遺伝子 g の部分配列となる．検証実験で，プローブの 1-30 文字目までに変異のないパターンはほとんど発現量に影響を受けなかった．よって，この条件を満たすことにより，プローブと遺伝子が結合する可能性が高い．

一方，プローブ p と遺伝子 g が結合できない条件を以下のように定義した．

定義 4. 非結合条件

$$D(p,g) > 20 \tag{5}$$

この式を満たす例としては，プローブの前半に 2 カ所+その他の領域に 1 カ所の変異を持つような場合がある．検証実験で，プローブ前半部に 2 カ所の変異があるとき，ほとんどの場合は有意に発現が低下する結果となっている．よって，この閾値により結合できないと判定をすることとした．

これらの条件を用いて比較トランスクリプトーム解析向けマイクロアレイの設計を行う．

3.4 比較トランスクリプトーム向けマイクロアレイの設計手法

本研究は，2 つのステップにより構成されている．図 3 で，提案手法の手順をヒト-マウス間で実行する場合の例を示した．なお，以降本論文では種間でホモログ関係にある遺伝子の集合をホモロググループとして記述する．H をホモロググループとし，設計するプローブは 1 つのホモロググループ H に対し 1 プローブとする．

ステップ 1:プローブ候補配列の探索

ステップ 1 ではホモログ間で共通して結合できるプローブ候補配列の探索を行っている．プローブ候補配列は H に含まれる全遺伝子と結合できる配列とする必要があるため，H に含まれる全遺伝子に対して定義 3 を満たさなければならない．よってステップ 1 で探索する，長さ 60 塩基のプローブ候補配列 p を以下のように定義する．

定義 5. プローブ候補配列 p

$$D(p,h) < 10 \ \text{ for } \ \forall h \in H \tag{6}$$

ただし，p は H の要素のうち 1 遺伝子の部分配列として探索するため，$h = h_1h_2h_3...h_n$ $(h \in H)$ として，

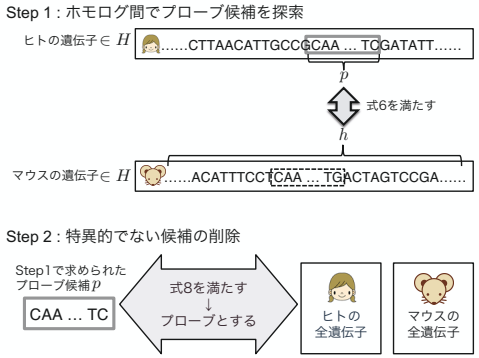


図 3 提案手法の概要
Fig. 3 Overview of proposed method.

$$p = h_i h_{i+1} h_{i+2} ... h_{i+59} \quad \text{where} \quad 1 \leq i \leq n - 59 \tag{7}$$

を満たす配列とする．p を種間で結合可能なプローブの候補配列とし，次にステップ 2 の処理でその特異性を検証する．

ステップ 2:非特異的プローブ候補の除去

ステップ 2 ではステップ 1 で求めたプローブ候補配列 p について，H に特異的であるかの判定を行い，特異的であった場合にはプローブとして採用する．特異的であるとは，標的ホモロググループ以外と結合できないことを指す．このため，G を発現量の比較に用いる種全ての遺伝子集合としたとき，$G - H$ に含まれる全遺伝子に対し式 4 を満たす必要がある．よって，H に特異的なプローブ p' の条件を以下のように定義する．

定義 6. 特異的プローブ p' の条件

$$D(p', g) > 20 \quad \text{for} \quad \forall g \in G - H \tag{8}$$

ステップ 1 の探索によって求められたプローブ候補配列が式 8 を満たすとき，p' は H 内の全ての遺伝子に結合可能であり，かつ，H 内の遺伝子に特異的な領域であると言えるので，H 内の全遺伝子の発現観測が可能なプローブとして採用する．p' が式 8 を満たさない場合，p' は H 内の遺伝子以外と結合し，発現量を誤検出する可能性がある．よって，p' はプローブ候補から除外し，ステップ 1 に戻って別のプローブ候補を探索する．

3.5 重み付き編集距離計算の高速化

3.4 節に示した順で計算を行う際，大量の $D(p, q)$ の計算が発生する．式 (2), (3) より，$D(p, q)$ の計算は動的計画法 (DP) を用いて高速化が可能である．よって，ステップ 1,2 の処理を動的計画法を用いた編集距離の実装を行うことによって高速化した．ただし，ステップ 2 の処理では $G - H$ に含まれる全遺伝子と 1 プローブ候補対 1 遺伝子の $D(p, q)$ の計算を行う必要があるため，DP を用いた場合でも計算コストが高いと考えられる．このため，本研究では長大な参照配列に対する高速なアラインメントを実現するツールに用いられている Burrows-Wheeler 変換 (BWT)[5] を応用した実装を行い，疑似データを用いて実行時間を計測し，DP による実装との比較を行った．

4. 実験と考察

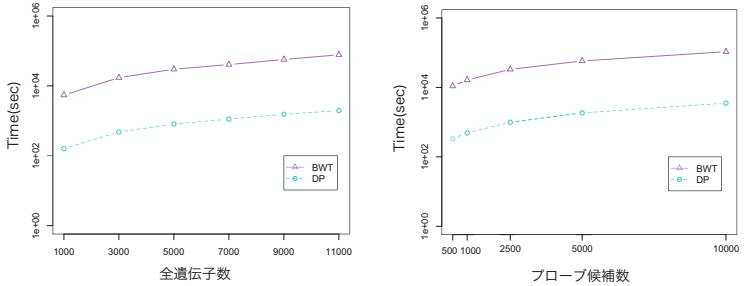
4.1 疑似データによる実行時間比較

疑似データを用いて，$D(p, q)$ の計算を DP により実装した場合と，BWT により実装した場合の実行時間比較を行う．実装は Java で行い，Java 1.6.0_20，CPU は AMD Opteron 3.2GHz，OS は Linux 2.6 を利用した．実行に際し，メモリは最大 5GB を利用した．疑似データの作成方法は，まず全遺伝子配列をランダムに作成し，特異的プローブはランダムな 60 塩基の配列とし，特異的でないプローブ候補は全遺伝子配列中に含まれる 60 塩基の部分配列にランダムな変異を与えることで作成した．疑似データのパラメータ初期値は表 1 の通りとする．$D(p, q)$ の計算は本手法のステップ 1,2 の両方で用いられているが，大量の $D(p, q)$ の計算が必要となるためより有意な差が出ると思われる，ステップ 2 の実行時間を計測し比較を行った．

まず，全遺伝子配列の数を変化させた場合の実行時間比較の結果を図 4(a) に示す．DP では $D(p, q)$ を 1 プローブ候補対 1 遺伝子で行うため，全遺伝子数が増加すると計算回数もそれに比例して増加すると考えられる．一方，BWT を用いた実装は BWT によって長大な文字列の情報を圧縮して処理するため，全遺伝子数が増加しても高速に類似配列を検索できると予測される．しかし比較の結果，BWT は実行時間が DP の 30 倍以上となった．これは，p と非常に類似した配列が存在した場合の各実装方法の振る舞いに原因があると考えられる．BWT による実装では，参照配列中に p と非常に類似した領域があると，p とその類似領域の間で許容されている全ての変異パターンによりアラインメントを試みることになるが，DP では各遺伝子に対して必ず 1 回しか計算を行わないため，より高速となると思われる．また，全遺伝子数が多くなると，p と類似した領域が増加するため，全遺伝子数に比例して実行時間も増加していると考えられる．

次に，ステップ 2 の実行回数により実行時間がどのように変化するかを検証した．ステップ 2 の実行回数が増えることは，ホモロググループ内の遺伝子が他の遺伝子と類似してお

Vol.2011-BIO-24 No.10
2011/3/11



(a) 全遺伝子数を変化させた場合.　　(b) プローブ候補数を変化させた場合.

図 4 疑似データを用いた，DP と BWT による実装の実行時間比較.
Fig. 4 Comparison of execution time between DP and BWT.

表 1 疑似データのパラメータ初期値
Table 1 Default parameters for pseudo data

パラメータ	初期値
全遺伝子数	3000
遺伝子配列長	3000
疑似プローブ候補数	1000
疑似プローブ数に対する特異的プローブ数	500

表 2 設計プローブ数
Table 2 The number of probes.

比較する種	プローブ数（全体からの割合）
ヒト，ラット	6,792（56.8%）
マウス，ラット	11,388（80.0%）

り，特異的なプローブ候補を発見することが困難であることを示している．よってこの検証で実行時間が大幅に増加する場合，遺伝子に重複があるホモロググループはあらかじめプローブ設計対象から除外するなどの考慮が必要となる．ステップ 2 の実行回数はプローブ候補数に等しいので，プローブ候補数を変化させて実行時間を計測する．結果を図 4(b) に示す．どちらの実装でもプローブ候補数の増加に対して実行時間の増加は少なく，遺伝子重複をあらかじめ考慮する必要はないことがわかったが，この比較結果でも DP が BWT の 30 倍高速であった．

以上のように，疑似データによる 2 種類の実行時間比較から BWT よりも DP が高速であるという結果を得た．BWT を用いた類似配列検索は元々ゲノムに対してシーケンサのリードをマッピングするために開発されており，ゲノムとリードの間に多くの変異がある場合を想定していないことが主な原因であると思われる．よって，本研究ではより高速な DP による実装を採用した．

4.2 実データによる実験

本研究では，実際にヒト-ラット間とマウス-ラット間の比較トランスクリプトーム用マイクロアレイを設計し，実験を行った．遺伝子配列は NCBI RefSeq から，ホモロググループのデータは NCBI HomoloGene から取得した[6]．

まず，設計可能なプローブ数について検証する．提案手法で設計されたプローブ数を表 2 に示した．ヒトーラット間の共通プローブは入力したホモロググループ全体の約 57%，マウスーラット間では 80%に対してプローブが設計可能となった．これまでの認識通り一つでも変異があれば結合できないと仮定して設計した場合は 16%程度のホモロググループにしか設計できなかったため，提案手法により設計されたプローブ数が大幅に増加しており，網羅的な発現量観測が必要となる比較トランスクリプトームに有効と言える．

さらに，提案手法で設計したマイクロアレイを用いて実験を行い，種間で発現量を比較した．比較には，ヒト-ラットのアストロサイトでの発現量観測実験とマウス-ラットの神経細胞での発現量観測実験の結果を用いた．結果，提案手法で設計したプローブは発現量の相関が高く，種間で同程度の発現量を観測できていることがわかった．よって，本研究のプローブ条件が適切であったことを確認することができた．

以上により，本研究で重み付き編集距離により定義したプローブの結合・非結合条件と，マイクロアレイ設計手法は比較トランスクリプトーム解析に有効であることが示された．

5. 関 連 研 究

5.1 マイクロアレイプローブ設計

マイクロアレイの原理は，DNA が相補対を作る事を利用しており，同様の現象を利用した実験に PCR がある[7]．マイクロアレイのプローブ設計は，一般に PCR の設計手法が応用できると考えられてきた．PCR では対象領域に特異的なプライマーを設計する必要がある．PCR の各ステップでは，急激な温度変化によって DNA 同士の結合可能な時間を短時間に制御しているため，プライマーー遺伝子間は完全に相補的でなければ結合できる確率が著しく下がることが知られている．このため，PCR のプライマーの場合には結合できる領域はプライマーと完全に相補的な領域に限られていた．よって，プライマー設計では特異性の確認に完全一致領域の探索が導入されており[8]，プライマーが結合対象とする領域が，他の遺伝子の部分配列になっていなければ特異的であると判断される．この知見をもとに，これまでのマイクロアレイ実験では，正確な発現量を得るためにはプローブー遺伝子間が完

全相補的でなければならないという認識があり，少しでもプローブ－遺伝子間に変異があれば結合できる確率が著しく下がると思われていた．

5.2 既存の比較トランスクリプトーム研究

比較トランスクリプトーム解析は，以下の 3 つの場合に分類される．

手法 (1) 各々の種用の既存アレイによる発現観測を行いホモログ間で発現量比較[2)]

手法 (2) 1 枚の既存のアレイを用いて発現観測を行い，各プローブの発現量を種間比較[3)]

手法 (3) 種間共通アレイを独自設計して発現観測を行い，各プローブの発現量を種間比較[4)]

手法 (1) は，主にモデル生物間の比較トランスクリプトームに用いられる．それぞれの種向けのマイクロアレイを用いるため発現量が正確であり，種間比較が容易である．ただし，プローブ間でホモログの対応を取る必要があり，対応づけられないプローブは比較に利用することができない．また，複数のアレイを必要とするため実験コストが高く，実験環境の統一も困難となる．手法 (2)(3) は，1 枚のマイクロアレイのみ用いるため低コストであり，比較対象種がモデル生物でない場合にも実験が可能となる．また，各プローブの発現量を種間比較するため，これまで知られていなかったホモログを発見可能という利点がある．しかし，種間の変異を考慮していないため発現量の正確さを保証することができず，別途 PCR 等で発現の確認をする必要があるなどの問題がある．また，独自設計の場合では，プローブは遺伝子と完全に相補的な場合のみ結合可能という認識により，網羅的にプローブを設計することができないという問題がある．種間で変異を持つ領域を使用すればプローブ数は増えるが，特異的かつ種間で結合可能なプローブの定義が明らかでないため，発現量の正確さを保証することが困難である．

本研究では検証実験を行うことによって，遺伝子との間に変異を含む，独自設計した種間共通プローブの発現観測の正確さを保証した．

5.3 BWT を用いた類似配列探索

BWT は，ソート済みの suffix array により文字列の可逆圧縮を実現する手法である．近年は BWT により長大な参照配列に短い問い合わせ配列をアラインメントする手法が盛んに研究され，次世代シーケンサ等で読まれた大量のリード配列をゲノムにマッピングするツール[9)10)] に活用されている．これらの手法では，ソート済み suffix array により効果的に探索空間を削減している．また，BWT を使用して問い合わせ配列を圧縮することができるため、メモリ効率が良い．さらに，BWT によって複数の遺伝子配列を一つの参照配列として扱い，一度の実行で類似配列の検索を行える．本研究では BWA を改良し，本研究の重み付けに適した探索を行えるように実装した．しかしゲノムへのマッピングとは異なり，多くの変異を許容して類似配列を探索するため，多くの変異を許容することを目的としないアラインメントツールのアルゴリズムは，探索空間を効率的に絞り込めない．よって，DP による実装と比較して実行時間が大幅に増加したと思われる．

6. おわりに

本論文では，マイクロアレイにおけるプローブ－遺伝子間の結合性質を実験的に検証し，これまでの結合性質の認識とは異なる結果を得た．この結果をもとに新たなプローブの結合・非結合条件の定義を行い，比較トランスクリプトーム向けマイクロアレイの設計手法を提案した．さらに実データによる実験を行い，実際に種間で網羅的かつ正確な発現量観測を行えることを確認した．

今後の課題として，プローブ設計の計算の高速化や，さらに実験結果に則したプローブの設計を行えるよう編集距離の重みや閾値の改良を行っていきたい．

参 考 文 献

1) Fujiyama A, Watanabe H, and *et al.,* Construction and analysis of a human-chimpanzee comparative clone map. *Science*. 2002; pp 131-134.
2) Su AI, Wiltshire T, and *et al.,* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*. 2004; 101:pp. 6062-6067.
3) Ji W, Zhou W, and*et al.,* A method for cross-species gene expression analysis with high-density oligonucleotide arrays. *Nucleic Acids Research*. 2004; 32:p. e93.
4) Vallee M, Robert C, and*et al.,* Cross-species hybridizations on a multi-species cDNA microarray to identify evolutionarily conserved genes expressed in oocytes. *BMC Genomics*. 2006; 7:p. 113.
5) Burrows M and Wheeler D. A block-sorting lossless data compression algorithm. *Digital Systems Research*. 1994; Center Research Report 124.
6) Wheeler DL, Barrett T, and *et al.,* Database resources of the national center for biotechnology information. *Nucleic Acids Research*. 2006; 34:pp. 173-180.
7) Saiki RK, Gelfand DH, and *et al.,* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 1988; 239:pp. 487-491.
8) Gusfield D. *Algorithms on Strings,Trees and Sequences: Computer Science and Computational Biology*; Cambridge University Press, 1997; pp. 178-180.
9) Li H and Durbin R, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:pp. 1754-1760.
10) Li R, Yu C, and *et al.,* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 2009; 25:pp. 1966-1967.