最尤推定法を利用した異数倍数体の遺伝子発現量定量化手法の開発

理学専攻 情報科学コース　山田　恵

1　はじめに

網羅的に遺伝子発現量を得る方法として，RNA-seq がある（図 1）．RNA-seq は遺伝子の転写物である mRNA を大量に読み，既知の遺伝子配列やゲノム配列と対応（アラインメント）させる．対応する配列の本数を数える事で遺伝子の発現量を計測する．近年のシーケンサでは，連続して読める配列（リード）の長さは 100 塩基程度と短いが，1 千万を超える本数が容易に計測可能となったことで，RNA-seq が実現可能となった．一方 RNA-seq には欠点もあり，対象種が持つ複数の遺伝子がほぼ同じ配列を有する場合，リードがどの遺伝子由来か特定できず，発現量の測定を誤る可能性がある．

図 1: RNA-seq の概要

本研究は，この RNA-seq の利点と欠点を上手に利用することで，異数倍数体における遺伝子発現の定量を行う．異数倍数体とは，異なる二種が交配し，子が両親のゲノムを持つ種で，生物の新規機能獲得プロセスの解明に有用であると考えられているが，両親のゲノム両方とも働くのか，働くとすれば同程度に働くのか，など未知の点が多い．本研究では，これらのゲノムに含まれる遺伝子の発現を同定する手法を確立する．異数倍数体の例として，本研究では *Cardamine* 属の *C.flexuosa* を用いる．この種は *C.amara* と *C.hirsuta* を親種に持つ．親種は，同属で進化的距離が近いため，互いに数塩基のみが異なる，ほぼ同一遺伝子集合を有している（図 2）．この種に対し RNA-seq を行うと，上記に述べた欠点によって，遺伝子の発現を定量化することが困難であるので，この問題点を解決する．

2　手法

RNA-seq では，リードをアラインメントするゲノムが必要だが，*C.amara* と *C.hirsuta* のゲノムは未知のため，*Cardamine* 属の近縁で，ゲノムや遺伝子が既知の *A.thaliana* を利用する．図 2 のように，*A.thaliana* は遺伝子 X，*C.amara* は遺伝子 X’，*C.hirsuta* は遺伝子 X”，*C.flexuosa* は X’，X ”を持つとすると，X，X’，X ”は互いに類似した配列を持つ．子である *C.flexuosa* の X’，X” 由来のリードを考えると，類似の配列を持つ *C.amara* の X’ と *C.hirsuta* の X” の配列両方に対応する．RNA-seq では，このような異なる遺伝子由来の配列が同一の遺伝子に対応することが問題となるが，今回は親種間の遺伝子配列に数塩基の違いにがある事を利用していずれの親由来であるか分離する．

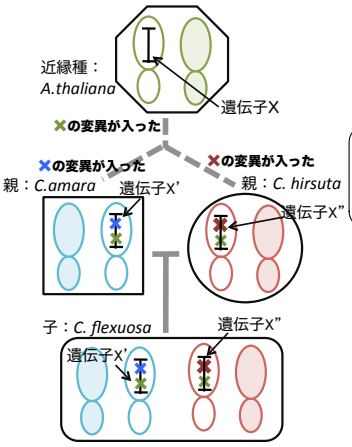

図 2: 実験対象の種

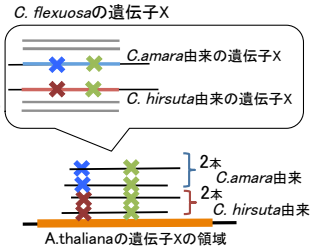

図 3: 手法の概要

2.1　変異の種類と求める方法

A.thaliana のゲノムと *C.flexuosa* のリードの間の変異には，親種特有の変異，祖先種から親種へ進化する過程で起きた変異（以下，それぞれ親由来の変異，祖先由来の変異）の 2 種類がある．図 2 の，*C.flexuosa* の X’，X” 上の x 印の，青と赤の x 印が前者で，緑の x 印が後者を示す．図 2 の X’，X” のリードを，*A.thaliana* の X にアラインメントすると図 3 のように表せる．x 印は図 2 の同じ色の x 印に相当し，x 印の赤は *C.hirsuta* 由来の X” の変異と同じで，青は *C.amara* 由来，緑は祖先由来の変異とみなせる．*A.thaliana* の X と異なる部分に注目することで，*C.flexuosa* の X’，X” の変異が得られる．
図示で色が付けられているが，実データでは変異の場所と変異の塩基の種類が分かるのみなので，変異が，親由来か祖先由来かを見分ける方法が必要になる．

2.2　遺伝子発現量の算出

変異の種類に注目すると，リードの組み合わせは 2 種類ある．親種特有の変異（赤と青の x 印）を見ると，リードの本数は赤と青で 2 本と 2 本に，祖先由来の変異（緑の x 印）を見ると，緑だけで 4 本と 0 本となる．各変異が，どちらに由来するものかは分からないので，実際には各変異位置での変異の有無の比率がわかるのみである．祖先由来の変異に関わるリードの本数の比は 0 に近く，親由来の変異に関わるリードの本数の比は，他の親特有の変異の比に近いと考えられる．この特徴を用いて，変異が親由来か祖先由来かを見分ける．ここでは特定の変異に着目して親種の比率を求めたが 1 点のみの推定ではシークエンスエラーによる影響も多く確度が低い．そこで本研究では，リード毎にいずれの親由来であるかを決定し，その本数を遺伝子全体にわたって数える事で計算の精度を向上させる．

2.3　アルゴリズム

次の 4 つのステップからなる．
ステップ 1 では，アラインメントを行い，祖先種の遺伝子配列からの変異を全て抽出する．ステップ 2 では，抽出した変異から，親由来の変異を求める．ステップ

3では，各リードを親種別に分ける．ステップ4では，片親由来と判定されたリード数を数え，両親のリード数の比と遺伝子の発現量から，親ごとの遺伝子発現量を求める．以下詳細を述べる．

ステップ1：子のリードを祖先種のゲノムに，いくつかの変異を許容して全リードをアラインメントする．以降，遺伝子ごとに計算する．変異が遺伝子上の k 箇所にあるとき，変異を有する位置の集合を $\mathbf{M} = \{M_1, M_2, ..., M_k\}$ で表す．$R1_i, R2_i$ を，M_i の位置を含むリードを M_i の塩基の種類ごとに分けたリードの集合とする．ここでは変異の有無のみを考えるので，2種類が考えられる．$R1_i$ の要素数を $|R1_i|$ と表す．

ステップ2：変異 M_i に着目した時，この変異は親由来か祖先由来の2種類の可能性がある．RNA-seq はランダムに配列を選択する手法なので，変異 M_i の状態が起こる確率 $P(M_i)$ は，ポアソン分布を用いてモデル化できる．ポアソン分布は平均 λ, 確率変数 k のとき，$\lambda^k e^{-\lambda}/k!$ と書ける．λ_p を親由来の平均，λ_a を祖先由来の平均とすると，M_i が親由来である確率は，$P(M_i) = \lambda_p^{|R1_i|} e^{-\lambda_p}/|R1_i|!$ ，祖先由来である確率は，$P(M_i) = \lambda_a^{|R1_i|} e^{-\lambda_a}/|R1_i|!$ と表せる．λ_a は任意の値 t ($t <= 1$, 実験では0.1) のとき，$\lambda_a = t \times (|R1_i| + |R2_i|)$ で計算する．それぞれの変異は独立な事象であるとすると変異 $\mathbf{M}$ の尤度 $P(\mathbf{M})$ は $P(M_1) \times P(M_2) \times \cdots \times P(M_k)$ で表せるので，この尤度が最大になるように，すべての変異を親由来か祖先由来かに分ければ良い．この時，親由来と考えられる変異セットを $\mathbf{M_P}$ ($\subseteq \mathbf{M}$) とする．

予め親種の平均 λ_p がわかっている場合には，簡単に $\mathbf{M_P}$ を決定することができるが，λ_p は求めたい変数であり不明である．つまり，$P(\mathbf{M})$ を最大化するには，最悪全ての考えうる $\mathbf{M_P}$ 及び λ_p を試さねばならない．これには膨大な時間を要する上に，現在考えているのは単一の遺伝子なので，実際にはこの計算を2万以上の遺伝子に対して行わなければならない．そこで，現実的な時間で計算できるよう，以下の工夫を行った．

（1）$\mathbf{M_P}$ に対して，親由来を表すポアソン分布の平均 λ_p は，大体 $\mathbf{M_P}$ に属する変異の平均に近いと考えられるので，次の方法で計算する．$I = \{i \mid M_i \in \mathbf{M_P}\}$ とすると，$k = |R1_i|$ のとき，

$$\lambda_p = (|R1_i| + |R2_i|) \times \left(\frac{\sum_{i \in I} |R1_i|}{\sum_{i \in I} |R1_i| + \sum_{i \in I} |R2_i|} \right)$$

（2）$|R1_i|/|R2_i|$ が小さいもの程，λ_p を小さくする可能性が高いので，この比率が 1 に近いものから順に $\mathbf{M_P}$ に入れて，尤度が最大となる組み合わせを計算した．最大尤度を取る I を I ’とする．

ステップ3：ステップ2までで求めた結果を元に，各リードに親ラベル A もしくは B を割り当てる．親ラベルは，$|R1_i|/(|R1_i| + |R2_i|)$ が大きい変異 $i \in I$ ’から順に割り振る．一度親ラベルが決まったリードはラベルを変更せず，次の候補へ移る．候補がなくなるまで繰り返す．

この割当を初期値とし，リード数の多い $i \in I$ ’から順に，割り振ったラベル A, B を反転させ，尤度が大きくなるかを調べる．これは，ラベルが親種を取り違えていないかどうかを確認するための手順である．以上の手法で，リードごとに由来する親種を特定する．

ステップ4：RNA-seq による遺伝子発現量は RPKM[1] と呼ばれる．ステップ3で求まった，両親それぞれのリード数の比と RPKM の積が今回求めたい両親由来別の遺伝子発現量となる．

3 実データを用いた実験の結果

次世代シーケンサ SOLiD5 で観測した *C.flexuosa* のリード約6千万本に対し，本手法を適用した．アラインメント先に *A.thaliana*（TAIR10）[4] のゲノムを選び，ゲノム上の反復配列を省くため RepeatMasker[2] を用い，アラインメントソフトには SHRiMP[3] を用いた．

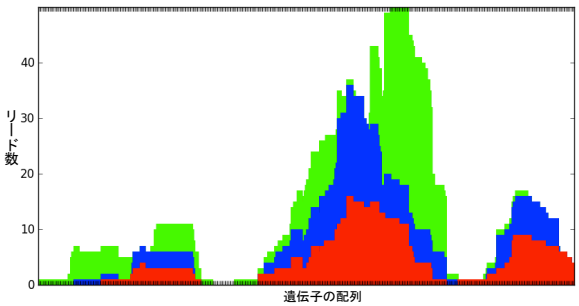

図 4: 遺伝子上の親由来の変異を持つリードの分布

図4は，一つの遺伝子に対し本手法を適用した結果を示している．横軸に遺伝子上の位置，縦軸に塩基ごとのリードの本数を表す．色分けは，リードごとに判定した親種に対応し，赤と青が片親，緑は由来が不明のリードである．この遺伝子の両親種ごとの遺伝子発現量は，それぞれ 5.93, 6.56 だった．グラフからは，親種ごとの発現量を求めるのに使われるリードが，ほぼ遺伝子の範囲上を覆っていて全体的な変異の傾向から発現量を求められていることが分かる．

4 まとめ

本研究では，異数倍数体の *C.flexuosa* の各遺伝子について，親種を区別して遺伝子発現量を計測する手法を開発した．本研究では，対象種である異数倍数体とその親種のゲノムと遺伝子が同定されていない状態で，近縁種との遺伝子セットが類似しているが完全には一致しないという特徴を使う事で，親種別の遺伝子発現量を求める手法を開発した．

謝辞

本研究にあたり，ご指導くださいました東京工業大学情報理工学研究科　瀬々潤准教授と，貴重なデータとご助言を賜りましたチューリッヒ大学　清水健太郎教授，清水理恵研究員に深く感謝いたします．

参考文献

[1] Mortazavi, *et al.* Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods, 5(7):621-628 (2008).

[2] Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2004. http://www.repeatmasker.org

[3] Rumble, S.M. *et al.* SHRiMP: accurate mapping of short color-space reads. PLOS Comput. Biol. 5, e1000386 (2009).

[4] http://www.arabidopsis.org/