Vol.2015-MPS-103 No.59
Vol.2015-BIO-42 No.59
2015/6/25

トピックモデルによる野外トランスクリプトームデータの解析

岩山 幸治[1,a)]　本庄 三恵[2]　岩崎 貴也[3]　永野 惇[1,2,4]

概要：トピックモデルとは，従来自然言語処理の分野において，膨大な単語から成る大量の文書データを少数のトピックで表現する手法である．近年，分子生物学や生態学など他分野への応用がなされてきている．遺伝子を単語，各サンプルを文書，遺伝子の発現を文書中への単語の出現と対応させることで，トランスクリプトームデータへトピックデータを適用することができる．本研究では，野外の複雑な環境における遺伝子発現の変動を解析するため，野外で生育するイネのトランスクリプトームデータをトピックモデルにより解析した．その結果，膨大な遺伝子の発現量の変動を少数のトピックの確率の切り替わりで表現できることを示す．

1. はじめに

野外環境では実験室環境と違い，気温や日射量などの環境条件は大きく複雑に揺らぐ．このような揺らぎを持った現実の環境への植物の応答を明らかにするためには，野外で生育する植物の遺伝子発現量，すなわちトランスクリプトームデータの解析が必要となる．多数のサンプルのイネからマイクロアレイで計測された遺伝子の発現量と気象データの関係を記述するモデリングが先行研究で行われている [1], [2].

先行研究のモデルでは，環境応答は遺伝子ごとに個別に記述されている．しかし，各遺伝子の振る舞いは独立ではなく，互いに制御をする結果，似た振る舞いを示す遺伝子群が現れる．また，先行研究ではマイクロアレイによる計測データを対数正規分布で記述している．近年広く使われるようになってきている網羅的なトランスクリプトーム解析法である RNA-Seq [3] では出力が離散値となるため，記述には離散分布が適している．

トピックモデル [4] は文書の潜在的な意味を扱うための方法で，文書のトピックをそのトピックにおける各単語の出現頻度として特徴づける．トピックモデルの一つである Latent Dirichlet Allocation (LDA) [5] は，生態系の多様性 [6] や miRNA による mRNA の制御のモジュール構造 [7] の特徴づけなど，自然言語処理以外の分野にも応用されている．

本研究では，大量の離散的なトランスクリプトームデータから類似した振る舞いをする遺伝子群を抽出すると同時にそれらの遺伝子群の環境応答を特徴づけるために，イネのトランスクリプトームデータをトピックモデルにより解析した．LDA の拡張モデルである Dirichlet-multinomial regression [8] を用いて，各サンプルにおけるトピックの構成比率を決める確率分布のパラメータと気象データや田植え後日数などとの間の回帰を行った．

2. データ及び解析手法

2.1 トランスクリプトームデータ及び気象データ

茨城県つくば市の実験圃場において，田植え日の異なる複数のイネから，2 時間おきに 24 時間，計 12 回のサンプリングを 1 セットとし，計 6 セットのサンプリングを行った．収集したサンプルから RNA-Seq により得たトランスクリプトームデータの内，総リード数の多い 20,000 遺伝子を以降の解析の対象とした．

気象データとして，圃場の近くに位置する気象庁の地上気象観測地点のデータを用いた．地上気象観測地点では，気温，相対湿度，気圧，風向および風速，全天日射量，降水量などについて，1 分ごとのデータが蓄積されているが，サンプリングを行った時刻，サンプリングの 12 時間前，24 時間前，72 時間前の気温と全天日射量を各サンプルの補助情報とした．

2.2 Dirichlet-Multinomial Regression

Dirichlet-multinomial regression [8] は以下の生成モデル

[1] 龍谷大学農学部
[2] 京都大学生態学研究センター
[3] 日本学術振興会
[4] 科学技術振興機構　さきがけ
[a)] iwayama-kouji@ad.ryukoku.ac.jp

Vol.2015-MPS-103 No.59
Vol.2015-BIO-42 No.59
2015/6/25

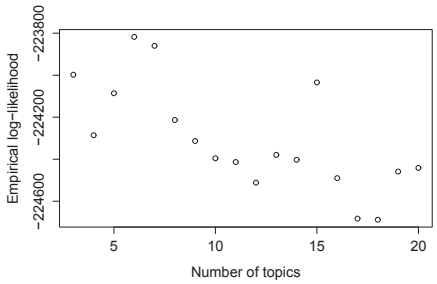

図 1　経験対数尤度
Fig. 1　Empirical log-likelihood

で記述される，

$$\lambda_{tk} \sim N(0, \sigma_k^2), \quad \alpha_{dt} = \exp(\boldsymbol{x}_d^T \boldsymbol{\lambda}_t), \tag{1}$$

$$\boldsymbol{\phi}_t \sim Dir(\boldsymbol{\beta}), \quad \boldsymbol{\theta}_d \sim Dir(\boldsymbol{\alpha}_d), \tag{2}$$

$$z_{di} \sim Multi(\boldsymbol{\theta}_d), \quad w_{di} \sim Multi(\boldsymbol{\phi}_{z_{di}}). \tag{3}$$

ここで，$\boldsymbol{x}_d = (x_{d1}, \ldots, x_{dk}, \ldots, x_{dK})^T$ は文書 d の著者などといった補助情報，$\boldsymbol{\lambda}_t$ はトピック t における，その回帰係数である．$\boldsymbol{\phi}_t$，$\boldsymbol{\theta}_d$ はそれぞれ，トピック t における各単語の出現確率と文書 d におけるトピックの構成比率であり，ディリクレ分布に従う．また，トピックの構成比率 $\boldsymbol{\theta}_d$ の従うディリクレ分布のパラメータは文書の補助情報の線形回帰の指数関数によって決まる．z_di は文書 d における i 番目の単語がどのトピックから生成されたかを示す潜在変数であり，$\boldsymbol{\theta}_t$ を各トピックの頻度とする他行分布に従う．最後に，文書 d の i 番目の単語は潜在変数 z_{di} の示すトピックに対応する多項分布から生成される．

回帰係数 $\boldsymbol{\lambda}_t$ を固定したもとでの潜在変数の周辺化ギブスサンプリングと，潜在変数を固定したもとでの回帰係数 $\boldsymbol{\lambda}_t$ の最適化を交互に行うことで学習する．回帰係数は，対数尤度 L の勾配，

$$\frac{\partial L}{\partial \lambda_{tk}} = \sum_d \alpha_{dt} x_{dk} \left\{ \Psi\left(\sum_t \alpha_{dt}\right) - \Psi\left(\sum_t \alpha_{dt} + n_d\right) + \Psi(\alpha_{dt} + n_{dt}) - \Psi(\alpha_{dt}) \right\} - \frac{\lambda_{tk}}{\sigma_k^2}. \tag{4}$$

に基づき，勾配法により最適化する．

3. 結果

気温と日射量，田植え後日数，また時計として 0 時と 6 時にそれぞれピークを持つ 2 種類の 1 日周期の正弦波を，各サンプルの補助情報とし，トランスクリプトームデータの解析を行った．サンプリングでは，最初の 250 回を破棄した後，1000 回のサンプリングを行った．回帰係数の最適化は 50 回のサンプリングごとに行った．3-fold の交差検定で求めた経験対数尤度 (図 **1**) がトピック数 6 の時に最大となったため，以降ではトピック数を 6 として解析を行った．

各トピックの回帰係数を図 **2** に示す．各トピックの環境応答の違いが確認できる．例えば，Heat shock protein の一つである HSP70 について $\boldsymbol{\phi}_t$ を確認するとトピック 5 の確率はほかのトピックに比べ 10^{10} 倍以上となっていた．このトピックはサンプリング時刻と 12 時間前の気温に強く応答していることがわかる．

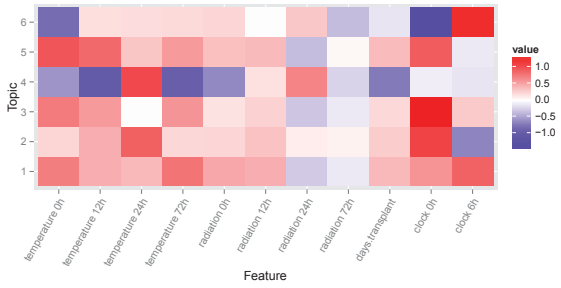

図 2　各トピックの回帰係数
Fig. 2　weights of features for each topic

4. おわりに

野外環境への植物の遺伝子発現レベルでの応答を明らかにするために，トランスクリプトームデータと気象データを Dirichlet Multinomial Regression によって解析した．その結果，環境応答の類似した遺伝子集団をトピックとして抽出することに成功した．

参考文献

[1] Nagano, A. J., Sato, Y., Mihara, M., Antonio, B. a., Motoyama, R., Itoh, H., Nagamura, Y. and Izawa, T.: Deciphering and prediction of transcriptome dynamics under fluctuating field conditions., *Cell*, Vol. 151, No. 6, pp. 1358–69 (2012).

[2] Matsuzaki, J., Kawahara, Y. and Izawa, T.: Punctual Transcriptional Regulation by the Rice Circadian Clock under Fluctuating Field Conditions, *The Plant Cell*, Vol. 27, No. 3, pp. 633–648 (2015).

[3] Wang, Z., Gerstein, M. and Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics*, Vol. 10, No. 1, pp. 57–63 (2009).

[4] 佐藤一誠：トピックモデルによる統計的潜在意味解析，コロナ社.

[5] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[6] Valle, D., Baiser, B., Woodall, C. W. and Chazdon, R.: Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method, *Ecology letters*, Vol. 17, No. 12, pp. 1591–1601 (2014).

[7] Liu, B., Liu, L., Tsykin, A., Goodall, G. J., Green, J. E., Zhu, M., Kim, C. H. and Li, J.: Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation, *Bioinformatics*, Vol. 26, No. 24, pp. 3105–3111 (2010).

[8] Mimno, D. and McCallum, A.: Topic models conditioned on arbitrary features with dirichlet-multinomial regression, *UAI* (2008).