確率生成モデルに基づく RNA-seq データの解析

松井佑介（名古屋大学）

本講演では、自然言語処理技術から発展してきたトピックモデルを応用した RNA-seq データ解析について発表する。基本的なトピックモデルは、Latent Dirichlet Allocation（LDA）と呼ばれる確率モデルに基づくもので、複数の文書に含まれる単語の頻度から、各文書の潜在的な意味を形成すると考えられる共起単語のグループ、すなわちトピックを推定するとともに、文書毎のトピック構成比の推定を行うことで文書分類を行う解析技術である。RNA-seq データの解析でも、文書をサンプルまたは細胞、また遺伝子を単語、単語頻度を遺伝子発現量、またトピックを生物学的機能を持った遺伝子群と考えることで、トピックモデルとの同様な状況を仮定することができる。先行研究では、Genotype-Tissue Expression(GTEx)と呼ばれる組織特異的な RNA-seq による遺伝子発現量データに対してトピックモデルにより組織特異的発現遺伝子群の同定 [1]や、1 細胞 RNA-seq 発現データにおける細胞種の同定[2]、遺伝子編集である CRIPSR-CAS9 による遺伝子介入前後の発現量の摂動を細胞毎のトピック構成比の変動として捉える解析手法[3]など、応用が進みつつある。

しかし RNA-seq データ解析におけるトピックモデルを応用した解析は有用であることが示されつつあるものの、多くの場合トピック数を事前に決める必要があることや、トピック自体の生物学的な解釈づけについても十分なコンセンサスが取れていないことなど課題もある。

我々は予めトピックに対して事前情報を組み込んだトピックモデルとして半教師付トピックモデルを応用した RNA-seq データ解析の手法の開発を進めている。実際に得られる RNA-seq データでは、実験生物学的あるいは臨床的データでは、各サンプルに背景情報（正常/腫瘍やグレードなど）が紐づけられている場合や、採取した各細胞に対してマーカー等で細胞種が事前にある程度わかっている場合などが多い。あるいは、腫瘍組織ではゲノム変異によるサブタイプ等のサンプルの特徴づけができている場合や、tSNE 等の部分空間クラスタリングである程度のグループ構造がわかっている場合なども多くある。このように事前に得られている事前情報をトピックに紐づけた推定を行うことにより、生物学的/臨床的な解釈を向上させるのみならず、トピック数の選択についても、自然な形で回避することができると考えられる。本講演では、本手法の紹介とともにいくつかの実データを用いた応用例について発表する。

参考文献：

[1] Dey KK, Hsiao CJ, Stephens M (2017) PLoS Genetics, 13(5).

[2] duVerle DA, Yotsukura S, Nomura S, et al. (2016) BMC Bioinformatics, 17:363.

[3] Duan B, Zhou C, Zhu C, et al. (2019) Nat Comm. 10: 2233