「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」
平成 27 年度採択研究代表者

H28 年度
実績報告書

津田　宏治

東京大学大学院新領域創成科学研究科
教授

離散構造統計学の創出と癌科学への展開

§1. 研究実施体制

1)津田グループ
① 研究代表者:津田　宏治 （東京大学・大学院新領域創成科学研究科・教授）
② 研究項目
・離散構造統計学の創出・普及

(2)門松グループ
① 研究代表者:門松　健治 （名古屋大学・大学院医学系研究科・教授）
② 研究項目
・癌検体の収集、実験データの取得および介入実験

(3)瀬々グループ
① 研究代表者:瀬々　潤 （産業技術総合研究所・人工知能研究センター・研究チーム長）
② 研究項目
・統計的検定手法構築、高速化、大規模化及び手法の普及

(4)竹内グループ
① 研究代表者:竹内　一郎 （名古屋工業大学・大学院工学研究科・教授）
② 研究項目
・網羅的遺伝情報の複合要因探索アルゴリズム構築・ソフトウェア実装・癌科学における実証

(5)山田グループ
① 研究代表者:山田　亮 （京都大学・大学院医学研究科・教授）

② 研究項目

・離散構造統計学の遺伝疫学・コホートスタディへの展開

§2. 研究実施の概要

選択的推論による組合せ要因発見手法の理論深化と高速化: 多数の仮説から統計的に有意なものを探す際、予め簡単な手法を用いて n 個選択しておき、そのあと、選択された仮説に対して統計検定を行うことは一般的に行われる。この際、選択したという事実を無視して検定を行うと、不当に優れた P 値が得られることが知られている。選択イベントに対して P 値を補正する方法を選択的推論(Selective inference)と呼ぶ。Lee and Taylor (2016)による Polyhedral Lemma の発見以降、選択的推論は NIPS, ICML などの機械学習分野でも存在感を高めている。本年度は、組み合わせ要因の選択的推論アルゴリズムの設計・提案を行った。まず、Safe screening という原理を用いて、従来の、Kudo-Morishita bound による方法に比して、数百倍以上の高速化率を得られるパターンマイニング手法を開発し、KDD 2016 に発表した。次に、選択的推論をパターンマイニングの枠組みで実現する世界初の組み合わせ要因探索法を設計し、論文が ICML2017 に採録された。

一細胞 RNA シーケンスを用いた神経芽腫発生の解明: 平成 28 年度は、主に 3 週齢の TH-MYCN マウスの SMG を対象に 1 細胞 RNA シーケンスを行った。2016 年に発売された最新型である 10x Genomics 社の「Chromium Single cell 3' Assay」を用いて実験を行い、トータルで約 5800 細胞の遺伝子発現データを取得した。各細胞につき 1000 以上の遺伝子発現データを含むため、概算で 5800 cells × 1000 genes のデータを取得した。t-SNE による結果から、約 5800 細胞は神経芽腫細胞を含む 7 つの細胞タイプに分かれることが明らかになった。

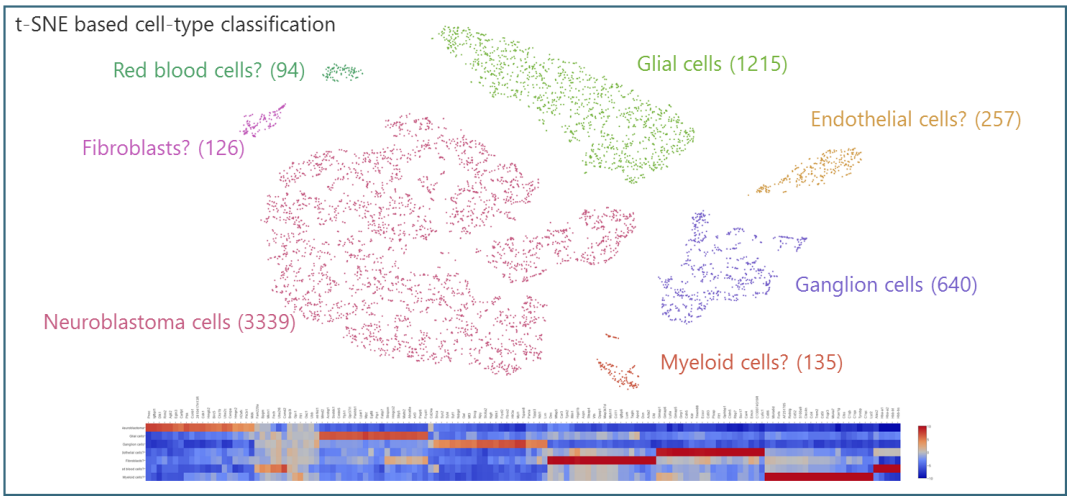

図: t-SNE を用いた一細胞 RNA-seq データの可視化。ヒートマップは、7つの細胞種で特異的に発現している遺伝子を示す。

代表的な原著論文

1) K. Nakagawa et al., Safe Pattern Pruning: An Efficient Approach for Predictive Pattern Mining, KDD, pages 1785-1794, 2016

2) A. Terada et al., LAMPLINK: detection of statistically significant SNP combinations from GWAS data, Bioinformatics, 32 (22), 3513-3515, 2016.

3) D.A. duVerle et al., CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data, BMC Bioinformatics, 17, 363, 2016.