BECKMAN COULTER
Life Sciences

サイトメトリーデータ解析における次元削減アルゴリズムの選び方

このアプリケーションノートでは、以下について述べます。

- 次元削減ツールの選択で考慮すべき点
- viSNE, UMAP, opt-SNE および tSNE-CUDA についての基礎知識と Cytobank の実行
- Cytobank プラットフォームにおける各アルゴリズムのイベント（細胞数）上限
- ご自身のデータセットに適切なアルゴリズムの選び方

フローサイトメトリーによる高次元データセット解析の最大の課題の1つは、各細胞から複数のマーカー発現を同時に解析することです。解析の対象が2、3のプロットのみである場合は、関係性の発見や、類似のフェノタイプを持つ細胞サブセットの同定は極めて容易と言えるかもしれません。しかしながら、近年のハイパラメータ化の潮流では、比較的少ないとされる色数（15カラー）でさえも、特定のサンプルでの細胞がどのような挙動をするかの全容を把握するためには、研究者は100以上のプロットを全て観察する必要があります。

次元削減アルゴリズムはデータを高次元空間から低次元空間へと変換することで、低次元上でも、元データの重要な特徴の多くを保持できます。フローサイトメトリーデータでは、次元削減によって各細胞が点で表示される二次元マップが作成され、このマップ上で近接表示される細胞同士はフェノタイプが類似している、つまり複数のマーカーの発現レベルが類似しているということになります。こうしたマップを作成することで、マルチパラメータデータセットが簡単に視覚化され、探索的なデータの迅速な解析も可能となります。研究者はこのマップから、あるフェノタイプを持つ細胞の「アイランド」（サブポピュレーションなど）を見出し、異なるサンプルからのマップと比較することで、この細胞サブセットの変動をサンプル横断的に見ることができます。

2008年のviSNEアルゴリズムの最初の論文[1]以降、フローサイトメトリーの分野では、少しずつ特徴が異なるものの、結果をよりわかりやすく、簡単に可視化するという同じ目的を持った次元削減アルゴリズムがいくつも現れています。

次元削減アルゴリズムが追加されたCytobank 9.0では、ユーザーは4つの次元削減アルゴリズムを選択することができます。それぞれのアルゴリズムについて、詳しく見てみましょう。

viSNE

Cytobankプラットフォーム上においてviSNEは、Barnes-Hut を実装したt-SNEアルゴリズムを使用しています[2]。viSNEはデータ内にtSNE1、tSNE2と呼ばれる、高次元空間での各細胞の位置情報を反映する2つの新たなチャネルを作成します。この2つのチャネルでプロットすると、viSNEマップが表示されます。このマップでは、それぞれの点が1つの細胞を表し、近接する細胞同士はviSNE解析に使用されるn次元において類似していることを示しています。Cytobankでは viSNEを実行する場合、Cytobank Premiumでは最大130万イベント、Enterpriseでは最大200万イベントを解析することができます。

Opt-SNE

Opt-SNE[3]はtSNEよりもはるかに速度の速いアルゴリズムです。Belkinaラボによる論文では、「カルバックライブラーのリアルタイム評価を使用して、計算初期の移動度とグラジエント演算の回数をデータセット特異的に適合させるt-SNEパラメータを自動設定するツールキット」と定義されています。つまり、opt-SNEは、各データセットの設定選択を自動化することにより、結果を得るまでの時間の短縮や、t-SNE空間での最適[3]なデータ解像を可能にするのです。このアプローチでは、試行錯誤を繰り返して経験則を得ることをせずともt-SNEに最適なパラメータを選択できるので、1つの研究プロジェクトにつき何時間もの計算時間を短縮できる可能性があります。Cytobankでopt-SNEを実行する場合、最大300万イベントを解析できます。

tSNE-CUDA

tSNE-CUDA[4]は データセットの視覚化を目的としたGPU搭載のt-SNEです。tSNE-CUDAは、処理速度が現行のほかのtSNEアルゴリズムを大きく上回っており、結果を得るまでの時間を大幅に短縮できます。CytobankでtsNE-CUDAを実行する場合、最大1,000万イベントの解析が可能です。

UMAP

UMAP[5]はUniform Manifold Approximation and Projectionアルゴリズムの略称です。CytobankではUMAPをGPU化することにより、実行速度をさらに速めることができました。UMAPは、各点が細胞を示す二次元のマップが得られる点はtSNEsと同じですが、ほかのどのアルゴリズムよりもグローバル構造の維持に優れていると広く考えられています。事実、高密度のアイランドの方が、グローバル構造は容易に維持できます。UMAPでは、アイランド間の距離に意味があり、互いに離れているアイランド同士よりも、互いの距離が近いアイランド同士の方が、類似度が高いのです。CytobankでUMAPを実行する場合、最大1000万イベントの解析が可能です。

Cytobankでは、次元削減解析から有意義な見識を得るために使用できる機能を複数備えています。Color-by-Channelはマップをチャネルごとに色付け表示するもので、測定パラメータ（図1A）のシングルセルにおける発現の視認や、対象の細胞をゲートしてポピュレーション化（図1B）が可能です。Color-by-Sampleはサンプル別に色付け表示し、別のサンプル（例：異なる処理条件）で作成されたマップとの比較（図1C）ができます。Color-by-Densityはカラー等高線表示で、密度分布をわかりやすく表示（図1D）します。こうした機能は、その背景にある原理が同じであるため、全ての次元削減アルゴリズムで使用できます。

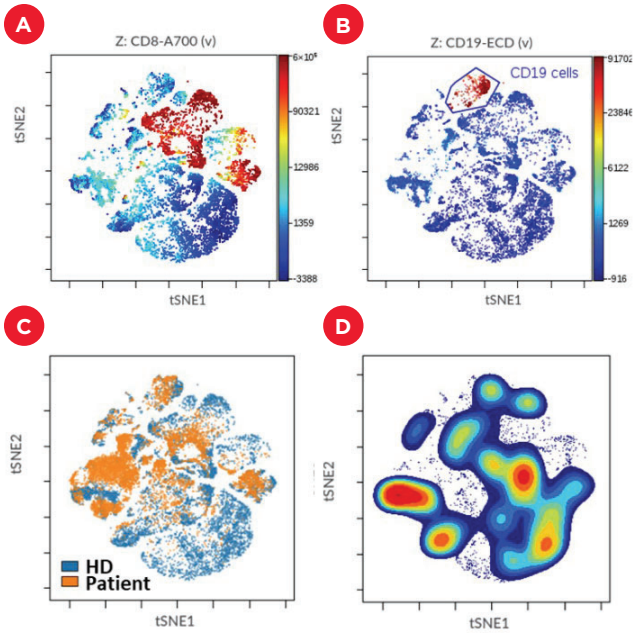

図1： 健康なドナー（HD）と患者（patient）から得た2種類のサンプルをCD45+細胞までプレゲートし、CytobankのviSNEをデフォルト設定で実行した。それぞれ、CD8チャネル強度での色付け例 **（A）**、チャネル強度色付け表示によるCD19+ゲーティング **（B）**、サンプル別色付け **（C）**、カラー等高線 **（D）** で表示。

また、CytobankでのtSNE-CUDAおよびUMAPは、計算性能を増強する大規模並列処理能を備えたGPU（graphics processing unit）を用いているため、従来のアルゴリズムよりも実行速度が格段に速いということを、最後に強調しておきます。tSNE-CUDAやUMAPのローカルへのインストールは非常に難しいかもしれません。また、効率的な解析の実行には、特殊なハードウエア要件を満たす必要がある場合もあります。Cytobankの最大の利点は、ユーザーはクラウドにアクセスすることで、そのパワフルな計算性能を利用し、ハードウエアやリソース、プログラミングの知識がなくても、次元削減アルゴリズムによる迅速な解析ができることです。

どのアルゴリズムを使えばよいでしょうか？

各アルゴリズムの基本的な特長について一通りわかったところで、ご自分のデータ解析にはどのアルゴリズムを選ぶべきかとお考えのことと思います。選択に役立ちそうな重要なポイントをいくつかご紹介します。

- tSNE-CUDAは、デフォルト設定でほとんどの検証済データセットにおいて非常にうまく機能することから、大規模データセットをより速く解析するための選択肢となり得ます。ただし、opt-SNE とviSNEの2つ以外のアルゴリズムは乱数シードの設定ができませんので、論文にシードを入れたい場合にはご注意ください。
- マップ表示の面では、tSNE-CUDAで表示されるアイランドはより大きく、密集度も低いです。また、点はより広がりをもって表示されUMAPほど互いに重なり合いません。そのため、UMAPで表示される密集度の高いアイランドよりもtSNE-CUDAのアイランドの方が、サブポピュレーションや、細胞間の微細な違いの識別が容易です（図2）。

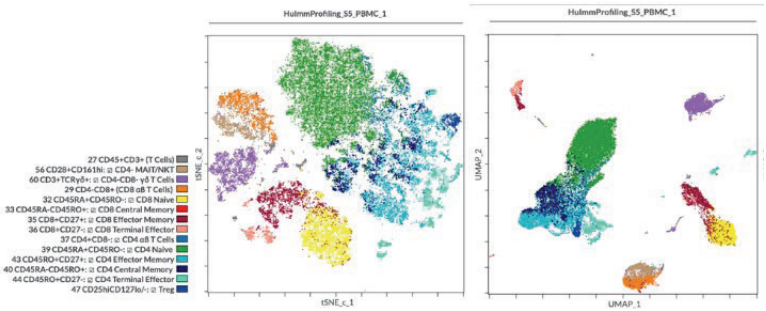

図2：どちらのプロットも、マニュアルゲートしたT細胞群を次元削減マップ上にオーバーレイしたものです。tSNE-CUDAでは、CD4細胞群（青と緑）がUMAPよりも広がって表示されています。

- 一方で、サイズの大きくないサンプルでは、UMAP結果の分布はtSNEsよりも明瞭に見える場合があります(図3)。

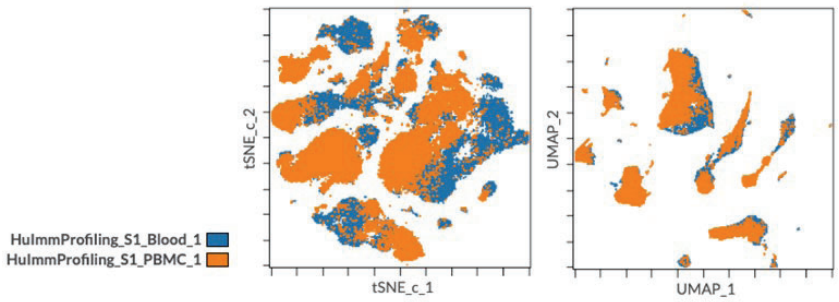

図3：tSNE-CUDA（左）とUMAP（右）での同一被験者の血液とPBMCサンプルのオーバーレイ表示。Fluidigm MaxParデータセットからのゲート01。tSNE-CUDAの方が、血液とPBMCの差異がより鮮明です。

- 図4のようなイベント数が少ないサンプルでは、tSNE では表示されるイベントが少なすぎて把握が難しいため、UMAPの方がデータパターンがわかりやすいことも重要です。

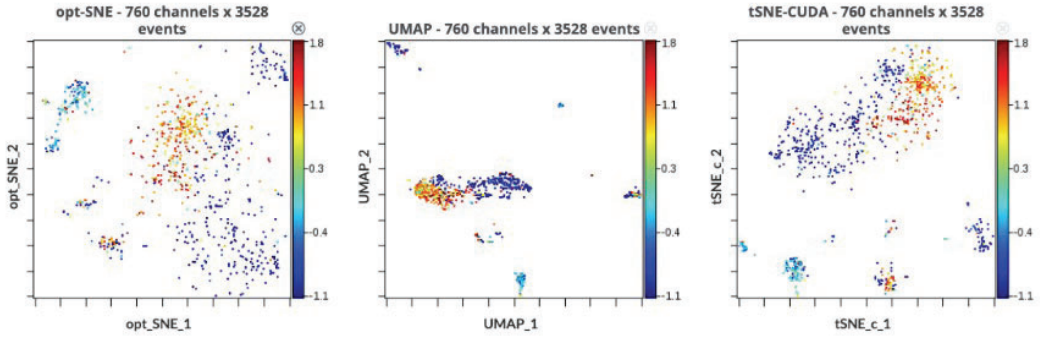

図4：760チャネルで3528細胞イベント4ファイルのCITE-seqデータでの次元削減解析結果例（Cytobank Enterprise）

- また、UMAPではデフォルトの初期化時にグローバル構造が維持されることから、tSNEよりもグローバル構造の保持に優れている点も重要です。UMAPでは細胞アイランド間の距離に意味がありますが、tSNEsアルゴリズムではそうではありません。このことから、細胞集団間の関係性についても考慮したい場合には、UMAPの方が適しています。

本アプリケーションノートでは、さまざまな次元削減法の主な相違点に焦点を当て、データセットや研究上のニーズに最適なアルゴリズムを選択する際のガイドを示しました。

ご質問やご不明点などありましたら、下記Cytobankサポートへお気軽にご連絡ください。
cytobank_support@beckman.com

References

1. Van der Maaten and Hinton, Journal of Machine Learning Research, 2008
2. Van der Maaten, Journal of Machine Learning Research, 2014
3. Belkina A. et al., Nat Communications, 2019
4. Chan DM et al., Journal of Parallel and Distributed Computing, 2019
5. Corey J. Nolet. et al., arXiv, 2020, arXiv:2008.00325 **http://arxiv.org/abs/2008.00325**

ベックマン・コールター株式会社

本　　社：〒135-0063 東京都江東区有明3-5-7 TOC有明ウエストタワー

お客様専用 0120-566-730　☎ 03-6745-4704　FAX 03-5530-2460
e-mail bckkcas@beckman.com　URL https://www.beckman.jp