修 士 論 文

画像処理と次数制約付き最小全域木に基づく細胞系統擬似時間の予測手法の改良

指導教員：阿久津　達也 教授

京都大学大学院情報学研究科

岩城　拓磨

2025 年 2 月 6 日

画像処理と次数制約付き最小全域木に基づく細胞系統擬似時間の予測手法の改良

岩城　拓磨

内容梗概

シングルセル RNA シーケンシングは、細胞のスナップショットを取得する技術であり、データ解析を通じて生物学的プロセスにおける変化の連続性を再構築することが可能である。細胞系統擬似時間解析のためのツールとして Slingshot や Monocle が広く利用されており、これらは最小全域木を用いて生物学的プロセスの軌道を予測している。しかし、これらのツールは「細胞の状態が一度に 3 つ以上には遷移しない」という生物学的前提を考慮していない。本研究では、Slingshot を改良し、この課題を解決した。また Slingshot では入力データをクラスタリングする必要があるが、適切なクラスタ方法やパラメータの決定方法などは不明である。画像処理によってクラスタ時に使用するパラメータを自動決定することでこの課題を解決した。具体的には次の方法で解析を行う。2 次元に次元削減後の シングルセル RNA シーケンシングデータをヒストグラムを用いて画像として認識する。その画像に対して細線化処理をすることで大まかな木構造を計算する。その木構造の分岐と葉の数を調べることでシングルセル RNA シーケンシングデータのクラスタ数を決定する。シングルセル RNA シーケンシングデータをクラスタリングしその重心と決定したクラスタ数を用いて次数制限付き最小全域木を構築することで、「細胞の状態が一度に 3 つ以上には遷移しない」という生物学的前提を考慮した細胞状態の遷移構造とする。得られた最小全域木を principal curve で曲線にフィッティングすることで細胞系統擬似時間の軌道とする。この改良の有効性は、Bone marrow mononuclear cell、Hematopoiesis、Human fetal immune cell、および C. elegans の公開データセットを用いて実証した。

Improved method for predicting cell lineage pseudo-time based on image processing and constrained minimum spanning tree

Iwaki Takuma

Abstract

Single Cell RNA sequencing (scRNA-seq) provides snapshots of individual cells, enabling the reconstruction of sequential changes in biological processes through data analysis. Various tools exist for pseudo-time analysis of cell lineage in biological processes, including Slingshot and Monocle, which utilize minimum spanning trees to predict trajectories. However, these tools do not account for the biological assumption that cellular states transition no more than three ways simultaneously. In this study, we present an improved version of Slingshot that addresses this limitation. Slingshot also requires clustering of input data, but the appropriate clustering method and parameter determination methods are not yet known. We solved this problem by using image processing to automatically determine the parameters to be used for clustering. Specifically, the following methods are used to analyze the data. Single Cell RNA-seq data reduced to two dimensions are recognized as images using histograms. The rough tree structure is computed by performing a skeletonization process on the image. The number of clusters in the Single Cell RNA-seq data is determined by examining the number of branches and leaves in the tree structure. By clustering the Single Cell RNA-seq data and constructing a degree-limited minimum spanning tree using the center of gravity and the determined number of clusters , the transition structure of the cellular state is constructed considering the biological assumption that cellular states transition no more than three ways simultaneously. The

resulting minimum spanning tree is fitted to a curve with a principal curve to obtain a cellular phylogenetic pseudo-time trajectory. The effectiveness of this improvement was validated using publicly available datasets, including those from Bone marrow mononuclear cells, Hematopoiesis, Human fetal immune cell, and C. elegans.

画像処理と次数制約付き最小全域木に基づく細胞系統擬似時間の予測手法の改良

目次

1　はじめに

本研究の前提知識として、シングルセル RNA シーケンシング、擬似時間解析、そしてそれに使用されるソフトウェアについて説明する。

1.1　シングルセル RNA シーケンシング

生物の細胞は多種多様な種類から構成されており、1 つの細胞が分裂を繰り返すことで多様性を持った細胞群へと分化していく。このような細胞の多様性を生み出す運命決定の仕組みについては、いまだ多くの謎が残されている。細胞の分化や発生のメカニズムを解明するには、各過程で個々の細胞がどのような振る舞いをしているかを正確に把握することが不可欠である。これまで、マイクロアレイ解析や RNA シーケンシングを用いた網羅的な遺伝子発現解析が主に利用されてきた。しかし、これらの手法では数千以上の細胞から抽出した RNA を解析するため、複数の細胞群全体の平均的な転写状態を示す結果しか得られなかった。この課題に対して、近年大きく進展を遂げた シングルセル RNA シーケンシング は、個々の細胞レベルで遺伝子発現プロファイルを取得できる画期的な技術である。シングルセル RNA シーケンシングを用いることで、個々の細胞における遺伝子発現データを大量に取得・解析することが可能となり、細胞の分化プロセスを従来よりも詳細に追跡することが可能になった。この技術により、細胞運命の決定過程や分化の仕組みに関する新しい知見を得ることができるようになっている。例えば、胎生初期における組織の細分化から臓器が形成されるプロセスをシングルセル RNA シーケンシングを用いて解析することで、iPS 細胞から尿管芽を作製する方法が確立された [1]。シングルセル RNA シーケンシングでは細胞内の RNA を断片化した後、逆転写を利用して cDNA を作製し、それをシーケンサーで読み取ることによって、1 つの細胞内で発現している遺伝子の量を計測する。シングルセル RNA シーケンシングでは一般的に図 1 のように使用する。細胞をどんどん細胞分裂させていく過程で細胞を

サンプリングする。それらを１つ１つシーケンサと呼ばれる機械にかけ細胞内の RNA などを読み取ることで１つの細胞の遺伝子発現データを得ることができる。シーケンサは細胞の中の RNA を全て取り出して断片化等の処理を行うため読み取りに使用された細胞はそれ以降の実験において細胞分裂させることはできない。得ることができるのは細胞のその時点でのスナップショットだけであり、１つの細胞だけを用いて分化の全ての過程における遺伝子発現量をシーケンサで計測することは現時点ではできない。よって多くの細胞のスナップショットデータを作成し解析することで細胞の分化を分析する必要がある。

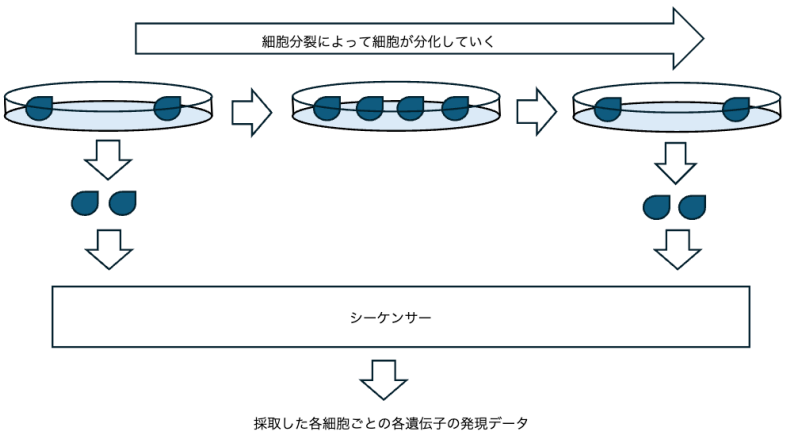

図 1: シーケンサーを用いた実験の流れの概要

1.2 擬似時間解析

各細胞の内部状態の変化は時間的に同期されているわけではなく、各細胞ごとの時間軸で細胞は変化していく。しかし研究者が欲しい情報は初期の細胞がどのような分化過程を辿るのかであるので、細胞全体として細胞の変化を追う必要がある。それらを解析するために擬似時間解析は使用される。擬似時間解析とは細胞の分化過程や疾患の進行などにおける細胞状態の遷移を理解するため、遺伝子発現データに基づき、細胞間の類似性を指標として擬似的な時間軸に沿って並び替える手法である。高次元データを擬似時間という１次元の特徴として捉えることであり、分化や疾患などの時系列の挙動を要約することがで

きる [2]。シングルセル RNA シーケンシングデータを使用して擬似時間解析を行う場合、各細胞の遺伝子発現は連続的に変化すると仮定する。つまり遺伝子発現プロファイルが似ている細胞は同じ分化の状態にあり時間軸としても近い位置にあると仮定するということである。擬似時間解析では高次元データであるシングルセル RNA シーケンシングデータの各細胞の位置の情報を元に各細胞を並び替える。この並びは分化の過程という時間軸のなかの順番ということになる。一般的に次の図 2 のように解析は行われる。与えられたデータ (図 2.a) に対して何らかの方法で細胞の分化の過程である軌跡 (図 2.b) を描く。よく最小全域木やグラフ埋め込みなどが使用される。その線に向かって各細胞を投射し (図 2.b)、スタートの細胞から各分化経路ごとに細胞の並び (図 2.c) を見ることで擬似時間を得ることができる。

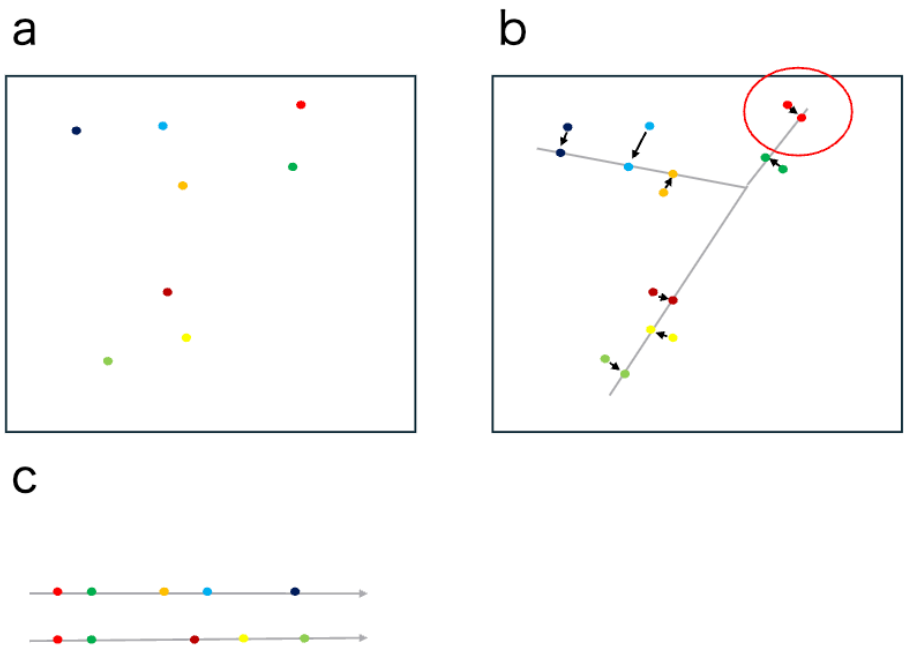

図 2: 擬似時間解析の流れ **a** 前処理等を行ったデータ (次元数は問わない) **b** データ空間上に状態遷移の軌跡を描いた後に各点を奇跡にプロットする **c** スタート点から点を並べ替えた結果

1.3 擬似時間解析ツール

擬似時間解析で使用されるソフトウェアについて説明し、それらの課題と本研究でどのように課題を解決したかについて説明する。擬似時間解析を行うためのソフトウェアには、Monocle や Slingshot などが広く利用されている。Monocle は最も普及しているツールの 1 つであり、特に Monocle2 では、高次元データから次元削減と木構造の同時構築を可能とする DDRTree アルゴリズムを採用し、細胞分岐イベントを効率的に検出している [9]。一方、Slingshot は他の解析ツールと併用しやすい設計が特徴であり、細胞が生物学的プロセスを経る中での変化を明らかにするために活用されている。Slingshot では、次元削減後のデータをクラスタリングし、その重心を基に最小全域木を構築する。この木構造に対して Principal Curve アルゴリズムを適用し、各系統ごとに細胞を滑らかな曲線にフィッティングすることで、連続的な擬似時間を生成する [3]。しかし、Monocle や Slingshot では、グラフ構造の構築時に「細胞状態が一度に 3 つ以上に分岐しない」という生物学的前提が考慮されておらず、計算結果として 3 つ以上の分岐が生じる可能性がある。また、Slingshot ではクラスタ数をユーザーが指定する必要があり、この数が最小全域木や擬似時間の結果に大きな影響を及ぼすにもかかわらず、その最適な決定方法は明確に示されていない。本研究は、Slingshot を改良し、細胞の内部状態データを画像として捉え、細線化アルゴリズムを活用してクラスタ数を自動決定する手法を提案する。また、最小全域木の構築時に分岐制約を導入し、細胞状態が 3 つ以上に分岐しないように改善することで、生物学的に解釈しやすい擬似時間の計算を実現した。

2 方法

2.1 解析方法の概要

前処理済みの シングルセル RNA シーケンシングデータを入力とし、擬似時間推定に至るまでの解析手順を以下に示す。まず、高次元データの次元削減を行い、高次元空間における細胞の分布を低次元にマッピングする。次に、得られた低次元データを画像化し、細線化処理を施すことで細胞間の関係を表す木構造を抽出する。続いて、適切なクラスタ数を決定し、細胞間の関係性を考慮した制約付き最小全域木を構築する。最後に、Slingshot を用いて曲線フィッティングを行い、細胞分化の軌跡を推定する。各解析ステップにおける結果を図 3 に示した。

1. **データの次元削減**

   高次元データを umap 等の次元削減アルゴリズムで 2 次元データに次元削減する (図 3.a)。

2. **データの画像化**

   二次元データをヒストグラムを使用して画像化する (図 3.b)。画像化することによりデータの輪郭を把握することができる。

3. **細線化による木構造の抽出**

   データの輪郭 (図 3.b) を細線化アルゴリズムにより処理し、おおまかな木構造 (図 3.c) を取得する。この木構造は、データの大まかな分岐と形状を反映する。

4. **クラスタ数の決定**

   木構造から葉と分岐の数を抽出し、それらの合計を n としてクラスタ数 (図 3.c) を自動的に決定する。その後、入力データをこのクラスタ数に基づいて Kmeans でクラスタリングする (図 3.d)。

5. **制約付き最小全域木の構築**

クラスタリング済みデータを用いて、細胞状態が一度に 3 つ以上に分岐しないという生物学的制約を考慮した最小全域木を構築する (図 3.d)。クラスタの重心の次数を 3 以下に制約をつけることで、これを実現する。

6. **Slingshot による曲線フィッティング**

   制約付き最小全域木と入力データを基に、`Slingshot` の principal curve を使用して滑らかな軌跡を描画する (図 3.e)。これにより、生物学的に解釈しやすい細胞の擬似時間推定が可能となる。

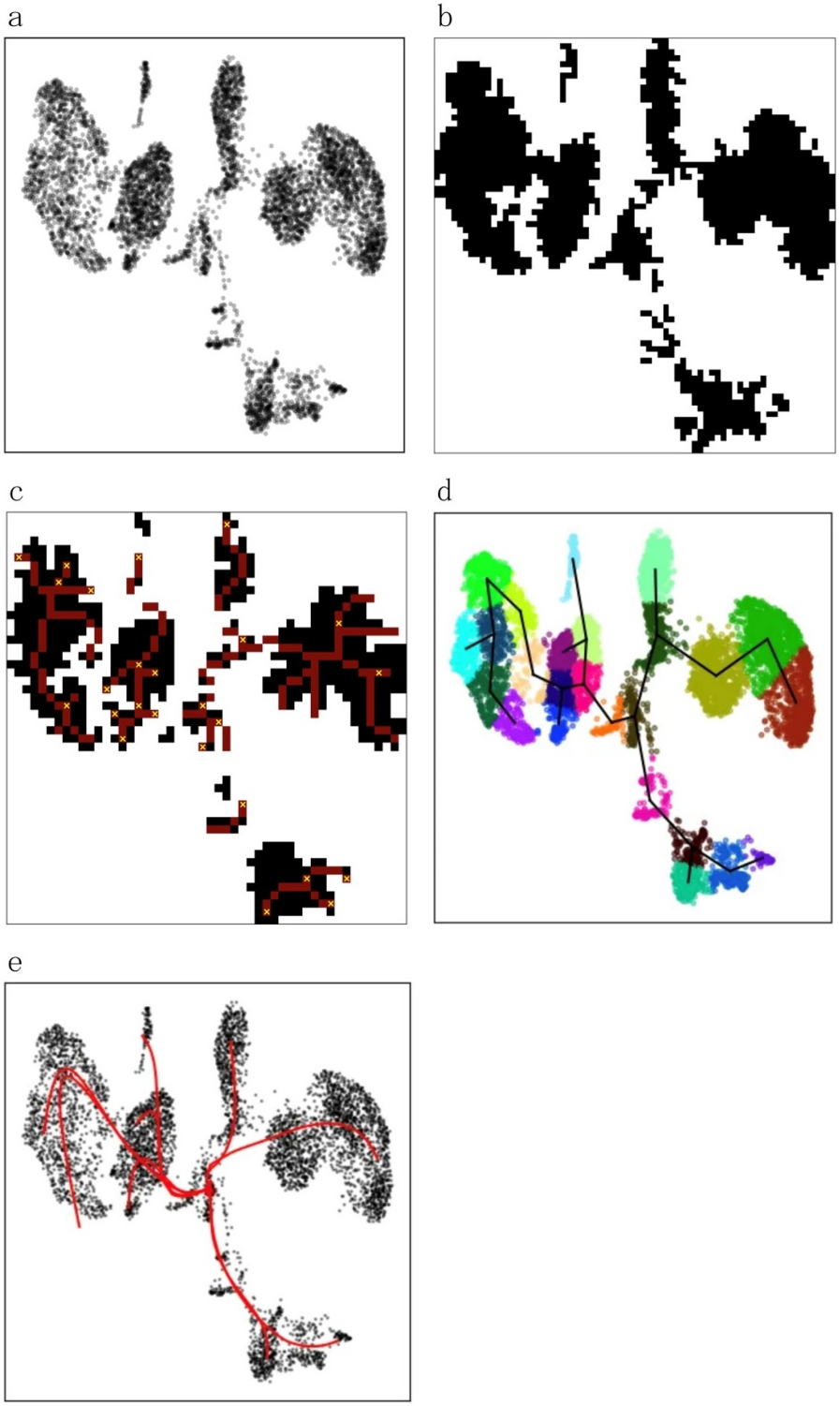

図 3: Bone marrow mononuclear cell を用いた改良後の slingshot による解析結果 **a** 入力データの散布図 **b** 解像度を指定して画像化した入力データ **c** 赤線は細線化した結果。黄点の数をクラスタ数 n とする。**d** クラスタ数 n でクラスタリング済みデータに対して最小全域木を構築した結果 **e** 最終的な解析結果

2.2 画像処理

本研究では、データを画像化した後、その画像から木構造を計算し、得られた木構造に基づいてクラスタ数を決定する。画像化からクラスタ数決定までの方法を説明する。

- 凸包

  点の集合に対してその外郭を形成する手法で、計算幾何学における形状抽出の技法の一つである (図 4)[13]。データが示す「空間的なアウトライン」を表現するために使用される。SLICER[4] において、近傍数の選択に凸包の面積を利用している。本研究ではその手法にヒントを得てデータの画像化の際に凸包の面積値を基準値としている。

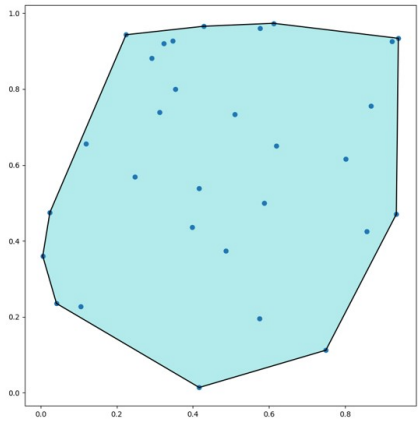

図 4: ランダムな散布図に対する凸包

- データの画像化

  データの各点は大きさを持たないため、そのままでは画像化できない。ヒストグラムを使用して点を膨らませ、ピクセルに詰め込んで画像化する。ヒストグラムでは、データを一定のビンに分割して表示する。ビンの境界値はデータ範囲を区切る位置を示し、ヒストグラムのビン境界を変更する

ことで、データを画像化する際の解像度を調整できる。ビン境界によってデータがどのような画像に変化されるかを比較した結果を図5に示した。本研究は、データの分散値に基づいて初期解像度を決定し、初期解像度の10倍までを9回の解像度で指定して画像化する。画像化後、連結オブジェクトの数、オブジェクトの面積を計算し、細線化処理を実施してクラスタ数nを取得する。

画像が荒すぎず細かすぎないビン境界を指定したい。ビン境界が高くなるとヒストグラムの区切る範囲が細かくなるのでデータが画像化された時の解像度が高くなり面積は小さくなる。ビン境界が低くなるとヒストグラムの区切る範囲が荒くなるのでデータが画像化された時の解像度が低くなり面積が大きくなる。画像化した後の面積ができるだけ大きいものを選択する。凸包の面積よりも画像のオブジェクトの面積が大きい場合、画像が荒すぎると判断し、その結果を選択から除外する。

画像化した後の連結オブジェクトの数が少なすぎるまたは多すぎるとクラスタ数に大きな影響を与えるため、オブジェクト数は10以上50以下の範囲に設定した（議論5.2）。

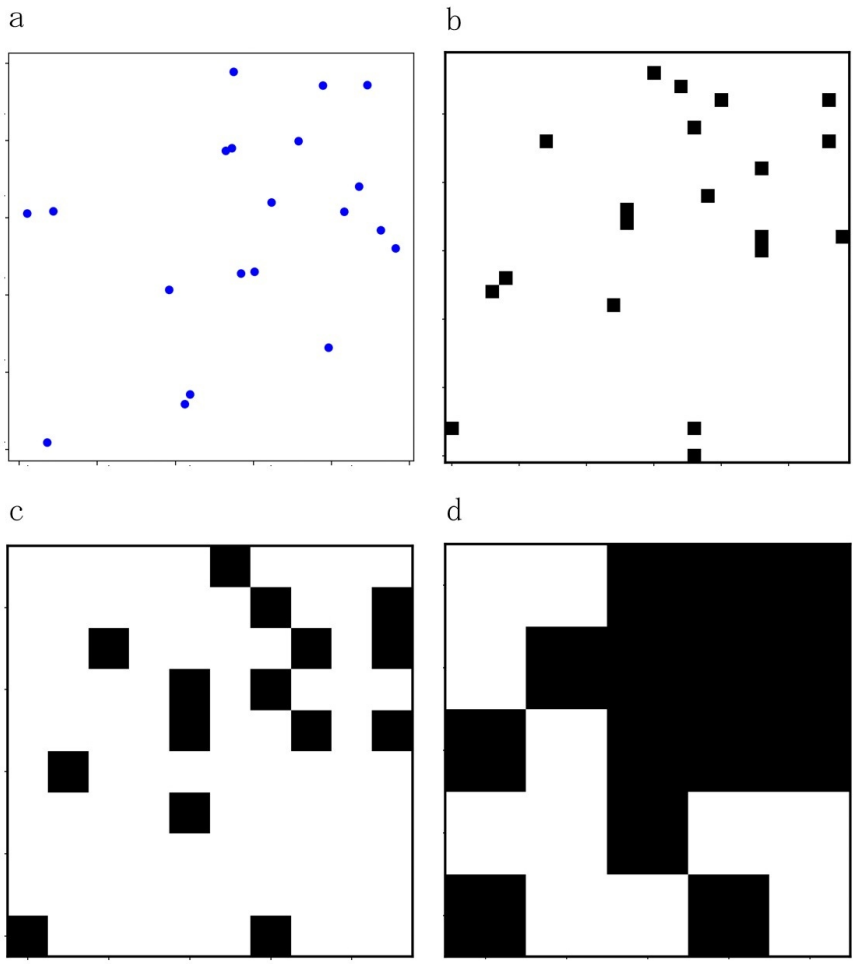

図 5: 画像化したデータの比較図 **a** ランダムなデータの散布図 **b** 境界ビンに 30 を指定した時の画像 **c** 境界ビンに 10 を指定した時の画像 **d** 境界ビンに 5 を指定した時の画像

- 細線化処理

  データの大まかな輪郭から木構造を作り、その木構造の葉と分岐の数をクラスタ数 n とする。画像の輪郭から木構造を構築するために、細線化を使用する。細線化は、2 値画像の画素領域を外側から削減し、幅 1 画素の線画像に変換する処理である [11]（図 6）。これにより、物体を中心線で表現でき、物体固有の特徴を把握することができる。得られた木構造の葉と分岐点をクラスタの初期重心として設定し、点の数をクラスタ数 n として使用する。

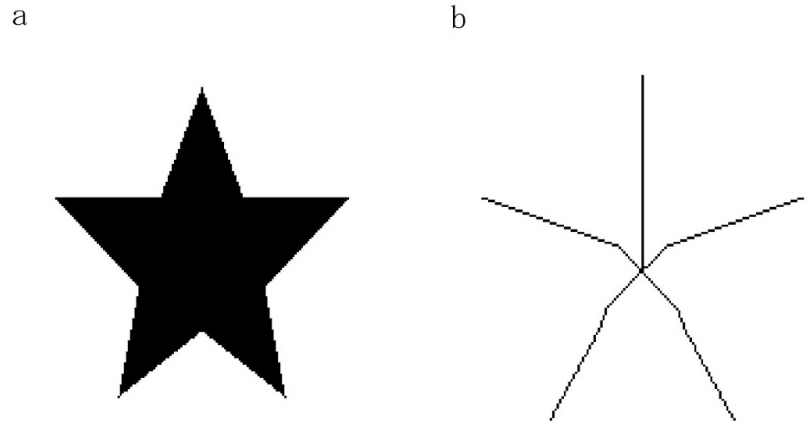

図 6: 細線化処理 **a** 元となる画像 **b** 細線化した結果

2.3 次数制約付き最小全域木

クラスタリング済みデータの重心を使用して最小全域木を構築し、細胞状態の親子・兄弟関係を予測する。細胞の状態が一度に 3 つ以上に分岐しないようにするため、一般的な最小全域木問題を整数計画問題として定式化 [12][15] し、追加の制約を導入した。具体的には、各頂点の次数を 3 以下に制限する制約を加えている。

- 最小全域木の定式化

  最小全域木とは重み付き無向グラフにおいて、すべての頂点を連結しつつ、全ての辺の距離の合計が最小 (1) となるように選ばれた部分グラフである (図 7.b)。最小全域木を整数計画法で定式化する必要があるため頂点間に辺があれば 1, 無ければ 0 となる決定変数 x を導入する (2)。全域木であるということは、全ての頂点が連結であり、かつ閉路を含まないことが必要である (3)。また全域木が閉路を含まないという条件を満たすためには部分的な閉路 (図 7.d) がないことを保証する部分順回路除去制約 (4)(5) が必要である。y_{ij}^{k} は頂点 k から見たとき、辺 (i, j) にフローが流れているかを管理する補助変数である。決定変数 x と補助変数 y の関係式 (4) は、辺 (i, j) が選ばれた場合のみ、フロー変数 y が活性化されることを保証する。フロー制

約 (5) は各辺 (i,j) において、どこかの頂点 k を経由する流れが必ず 1 つ存在することを保証する。フロー制約によりグラフ上に 1 つの有向木が形成される。有向木の根から頂点 k に向かう道の一部として辺 (i,j) が使われているかどうかを表すのが補助変数 y である。これらを踏まえると定式化した最小全域木は以下になる。

$G=(V,E)$: 無向グラフ

$i,j \in V$: グラフ G の頂点集合

$(i,j) \in E$, $e \in E$: グラフ G の辺集合

d_{ij} : 頂点 i,j 間の長さ

minimize

$$\sum_{(i,j)\in E} d_{ij}x_{ij} \tag{1}$$

s.t.

$$x_{ij}, y_{ij}^k, y_{ji}^k \in \{0,1\}, \quad \forall (i,j) \in E, k \in V \tag{2}$$

$$\sum_{(i,j)\in E} x_{ij} = |V| - 1 \tag{3}$$

$$y_{ij}^k + y_{ji}^k = x_{ij}, \quad \forall (i,j) \in E, k \in V \tag{4}$$

$$\sum_{k\in V\setminus\{i,j\}} y_{ik}^j + x_{ij} = 1, \quad \forall (i,j) \in E \tag{5}$$

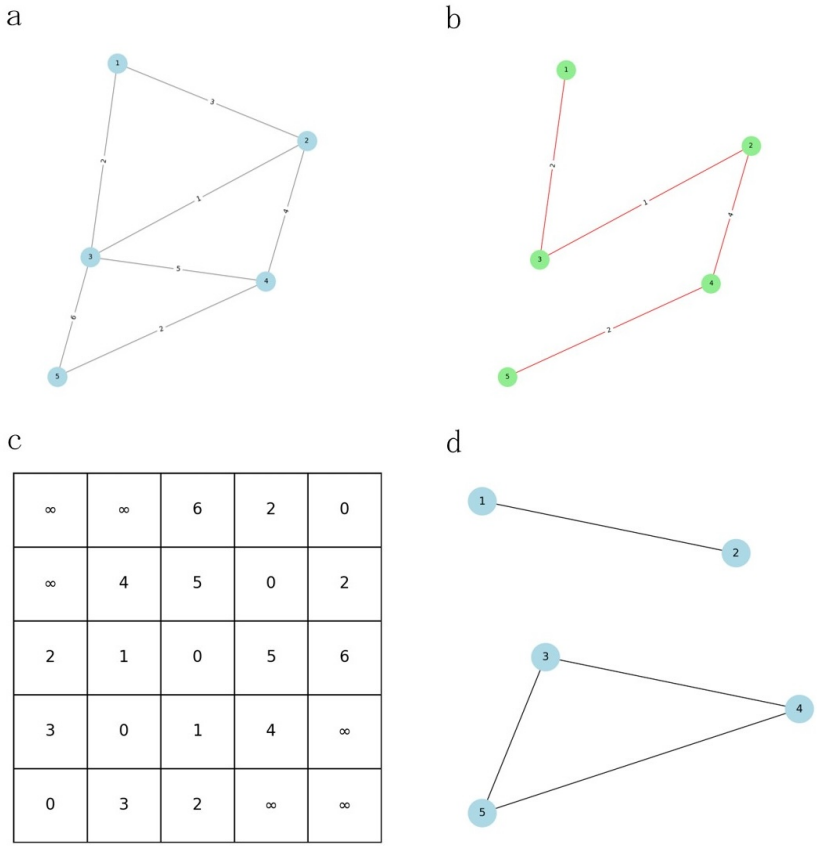

図 7: 細線化処理 **a** 元となる画像 **b** 細線化した結果 **c** 隣接距離行列 **d** 部分巡回路

- 次数制約

  細胞状態が一度に 3 つ以上に分岐しないという生物学的制約を追加するために次数制約を使用する。次数とは注目している頂点に接続されている辺の数である (図 8)。$\sum_{e \in \delta(\{i\})} x_e \leq 3 \quad \forall i \in V$ として次数を 3 以下にすることで実現する。最小全域木を解く方法は整数計画法以外にも存在するが、次数制約を追加する必要があるため整数計画法を使用している。

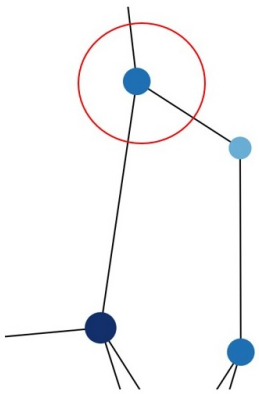

図 8: 頂点の次数

最終的な次数制約付き最小全域木の定式化は以下になる。

minimize

$$\sum_{(i,j)\in E} d_{ij}x_{ij}$$

s.t.

$$x_{ij}, y_{ij}^k, y_{ji}^k \in \{0,1\}, \quad \forall (i,j) \in E, k \in V$$

$$\sum_{(i,j)\in E} x_{ij} = |V| - 1$$

$$y_{ij}^k + y_{ji}^k = x_{ij}, \quad \forall (i,j) \in E, k \in V$$

$$\sum_{k\in V\setminus\{i,j\}} y_{ik}^j + x_{ij} = 1, \quad \forall (i,j) \in E$$

$$\sum_{e\in\delta(\{i\})} x_e \leq 3 \quad \forall i \in V$$

制約なしの最小全域木はクラスカルのアルゴリズムなどを使用して多項式時間で計算できることが知られているが、制約付きの場合は NP 困難となる。NP 困難となるのは、例えば、次数がすべて 2 とすると、NP 困難である巡回セールスマン問題を帰着できるからである。

3　結果

3.1　制約の効果

次数制約の効果を検証するために最小全域木に制約ありとなしの条件で、細胞の軌跡にどのような影響があるかを比較した。制約なしの場合、細胞の軌跡において 3 つ以上の分岐が確認されたが、制約を加えることでこの現象が解消された。本研究の方法でクラスタリング数 n を決定し、制約ありなしで同じクラスタリング結果を使用している。

- Bone marrow mononuclear cell

  次数制約無しの解析結果が図 9.c および 図 9.d である。図 9.d および 図 9.e には、次数が 4 の頂点が存在し、細胞が 3 つの状態に遷移していることが示されている。次数制約有りの解析結果が図 9.a および 図 9.b である。図 9.a および 図 9.b では、当該頂点の次数が 3 に減少しており、この現象が解消されている。

- Human fetal immune cell

  次数制約無しの解析結果が図 10.c および 図 10.d である。図 10.d および 図 10.e には、次数が 4 の頂点が存在し、細胞が 3 つの状態に遷移していることが示されている。次数制約有りの解析結果が図 10.a および 図 10.b である。図 10.a および 図 10.b では、当該頂点の次数が 3 に減少しており、この現象が解消されている。

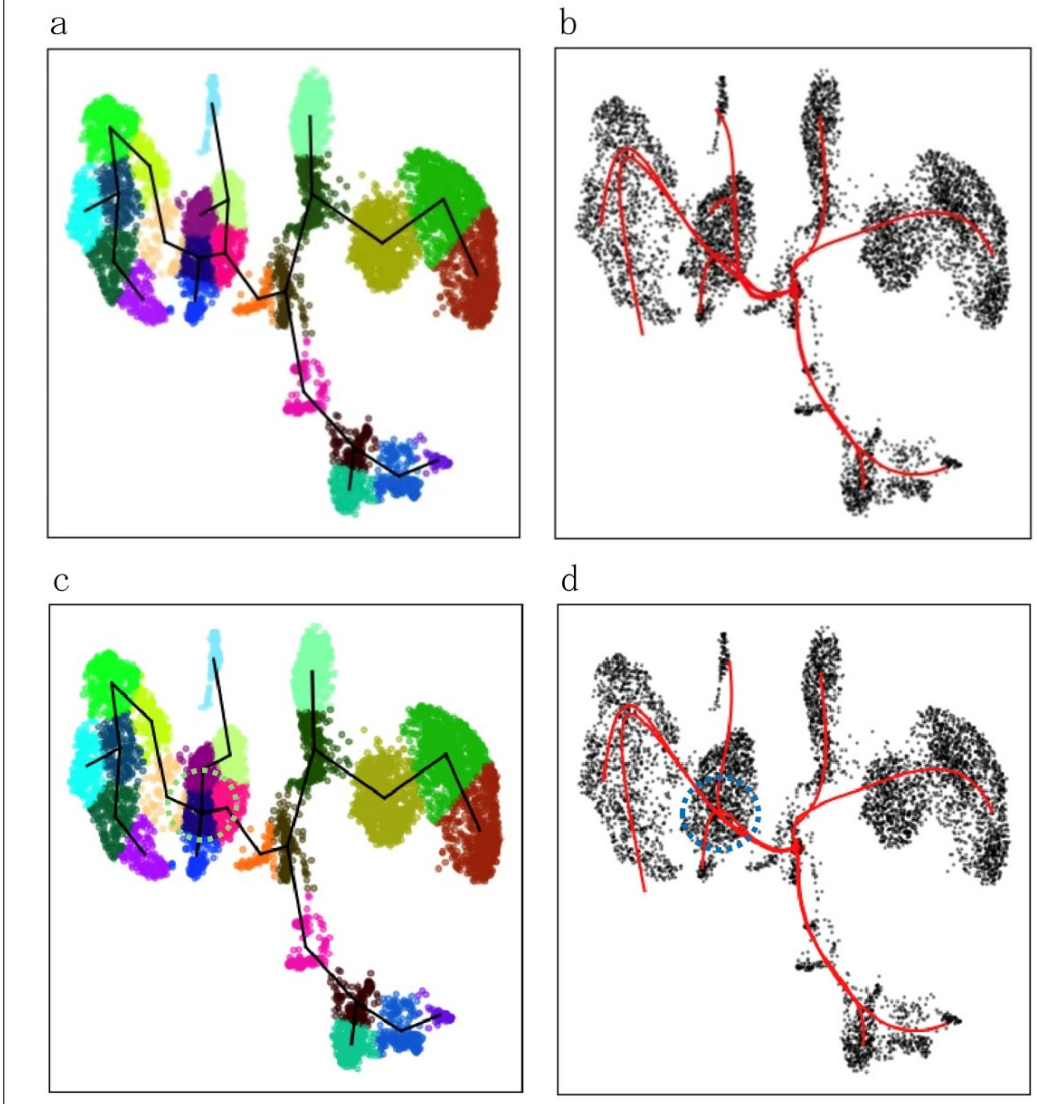

図 9: Bone marrow mononuclear cell のデータを使用した解析結果の比較図 **a** 制約有り最小全域木 **b** 制約有りの解析結果 **c** 制約無しの最小全域木 **d** 制約無しの解析結果

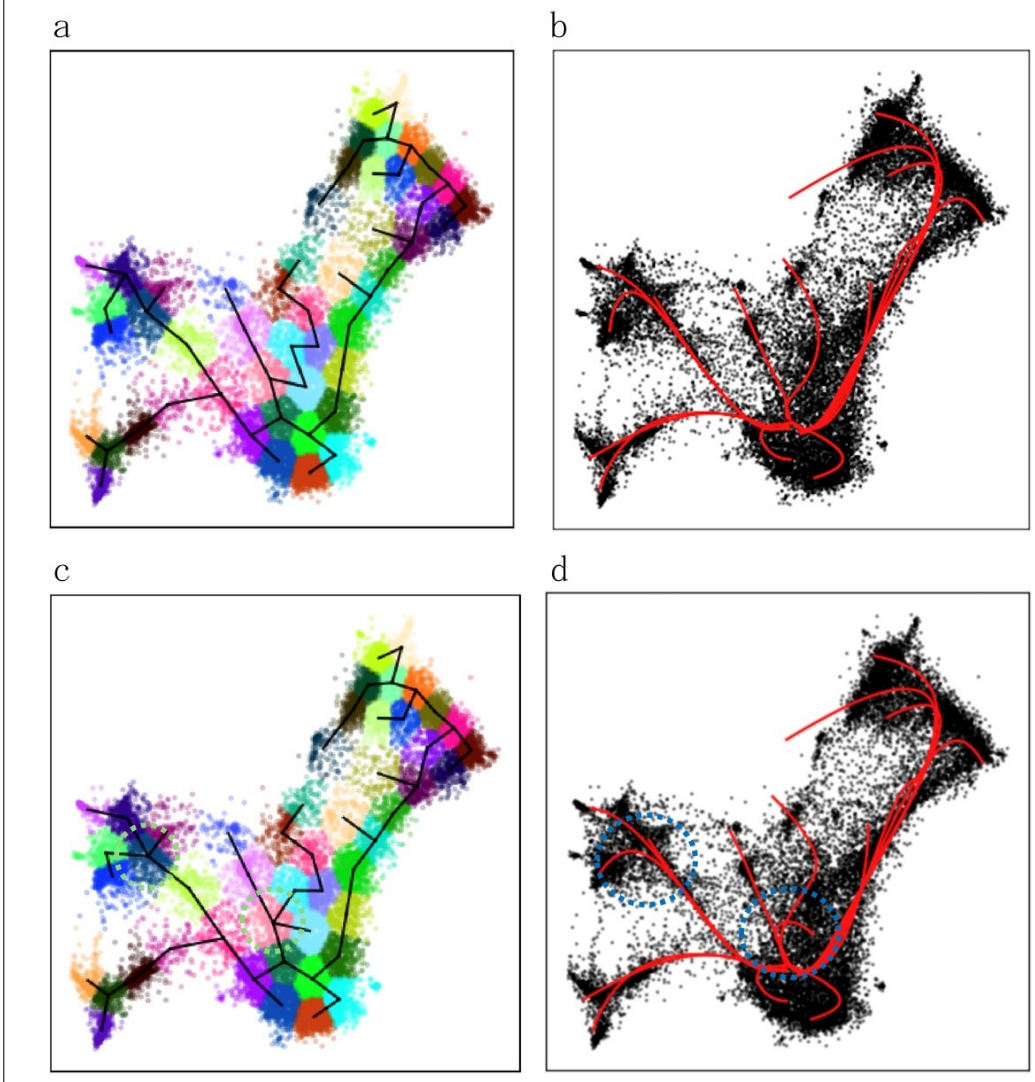

図 10: Human fetal immune cell のデータを使用した解析結果の比較図 **a** 制約有り最小全域木 **b** 制約有りの解析結果 **c** 制約無しの最小全域木 **d** 制約無しの解析結果

3.2 堅牢性

入力データに対する最小全域木の構築の堅牢性を検証するため、入力データとサンプリング済データに対して最小全域木を構築し、それらを比較した。サンプリングには、データをランダムに60%削減したものを使用した。検証には、Bone marrow mononuclear cell、Hematopoiesis、および Human fetal immune cell のデータを用いた。各データの検証結果は図 11 に示した。また各データのクラスタ数の情報を表 1 に示した。検証の結果、最小全域木の大まかな形状には大きな変化は見られなかった。しかし、細かい分岐の認識には違いが生じた。具体的には、図 11.c および図 11.f に示すように、左側の分岐の認識順序が変化した。また、クラスタ数にも変化が確認された。データを画像化する際に各点を膨らませて画素として認識する過程で、各点の距離が急激に変化することで、画像の形状が大きく変わり、クラスタ数に影響を与えたと考えられる。特に、Bone marrow mononuclear cell と Human fetal immune cell のデータは分散が低く、データの密度が高い特徴を持っている。このため、サンプリングによって40%のデータが削減されると、各点間の距離が急激に変化し、その影響でビン境界やクラスタ数に影響を与えたと推測される。

データ	データ数	ビン境界	クラスタ数
図 11.a	6224	50	25
図 11.b	3734	60	36
図 11.c	19764	50	50
図 11.d	11858	40	41
図 11.e	499116	100	71
図 11.f	29469	100	76

表 1: 図 11 の自動決定されたビン境界とクラスタ数

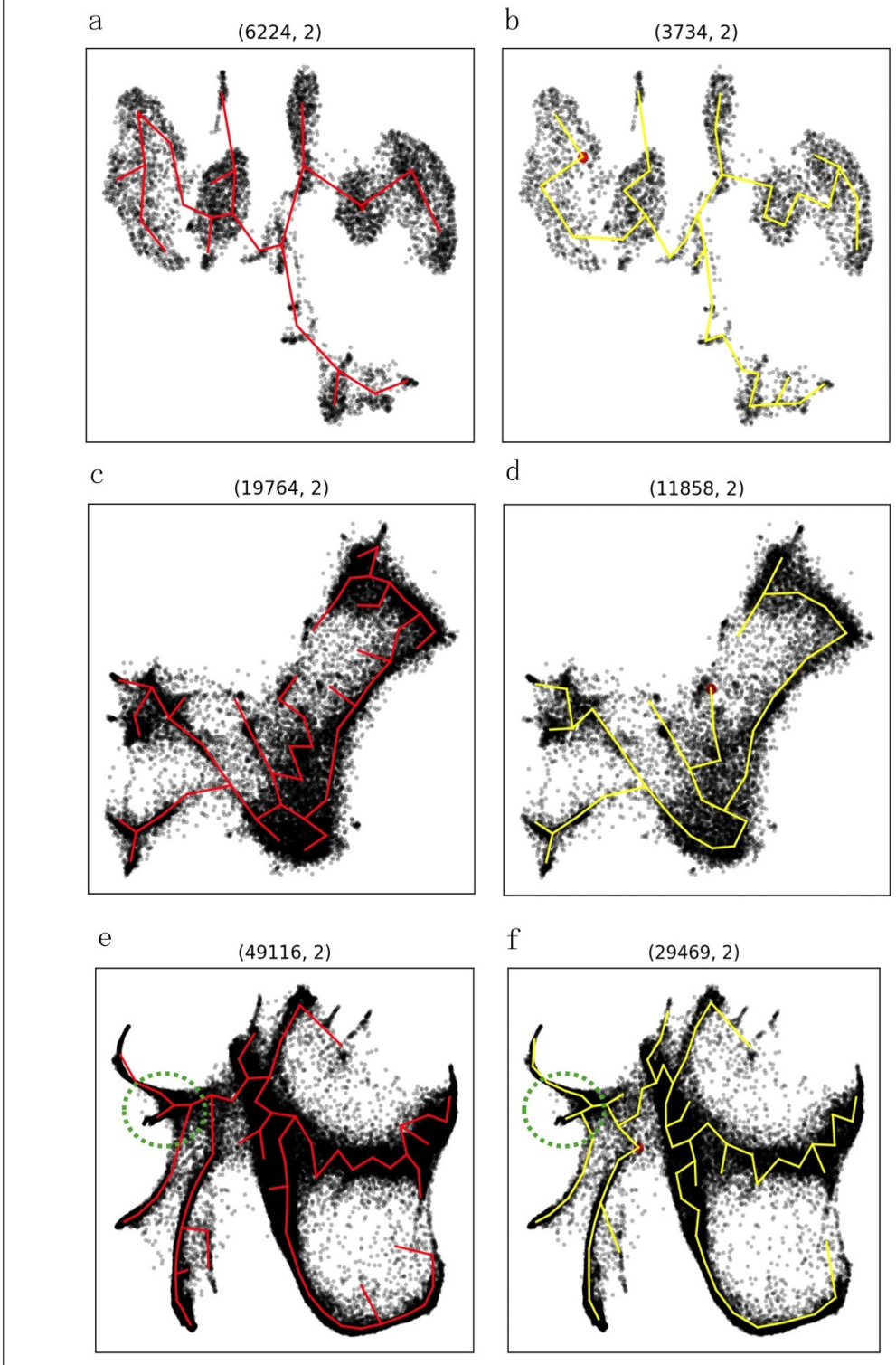

図 11: 最小全域木の比較図 **ab** Bone marrow mononuclear cell。**cd** Human fetal immune cell。**ef** Hematopoiesis。

3.3 Monocle3 との比較

本研究の実用性を検証するために C. elegans のデータを使用して Monocle3 の解析結果と比較した。結果は図 12 に記載した。Monocle3 は細胞の軌道を滑らかにフィッティングするわけではないので、きめ細やかな軌道予測になっている (図 12.c)。制約有りの結果 (図 12.a) と制約無しの結果 (図 12.b) を比較すると、制約無しの結果には Monocle3 の結果 (図 12.c) にはない 3 方向への分岐が発生している。制約をつけることにより細胞の軌道分岐の認識能力が向上した。

3.4 予測精度

本研究の予測精度を検証するために、Hematopoiesis のデータを用いて細胞の兄弟関係予測を実施した。Hematopoiesis データには DNA バーコーディング情報が含まれており、細胞の兄弟関係を記録することができる。Hematopoiesis データの DNA バーコーディング情報とシーケンスデータに基づき、CoSpar を用いて細胞状態の遷移予測を行った結果と、本研究によるクラスタリング後の最小全域木の結果を比較した。各クラスタ間の兄弟関係を正確に認識できているかを検証した。表 2 に結果を記載した。制約によってデータの約 60%の兄弟関係を予測でき、制約無しと比較して約 10%認識能力が向上した。

状態	データ数	制約付正答数	制約付正答率	制約無正答数	制約無正答率
Meg	1064	1064	1	0	0
Erythroid	365	365	1	0	0
Mast	1545	1545	1	1545	1
Baso	5998	5989	0.9985	5989	0.9985
Eos	168	0	0	0	0
Neutrophil	9231	9208	0.9975	9208	0.9975
Monocyte	9184	9069	0.9875	9069	0.3983
pDC	49	0	0	0	0
Lymphoid	64	0	0	0	0
Ccr7_DC	203	0	0	0	0
合計	27871	27240	0.5983	25811	0.4979

表 2: Hematopoiesis のデータに対する予測結果

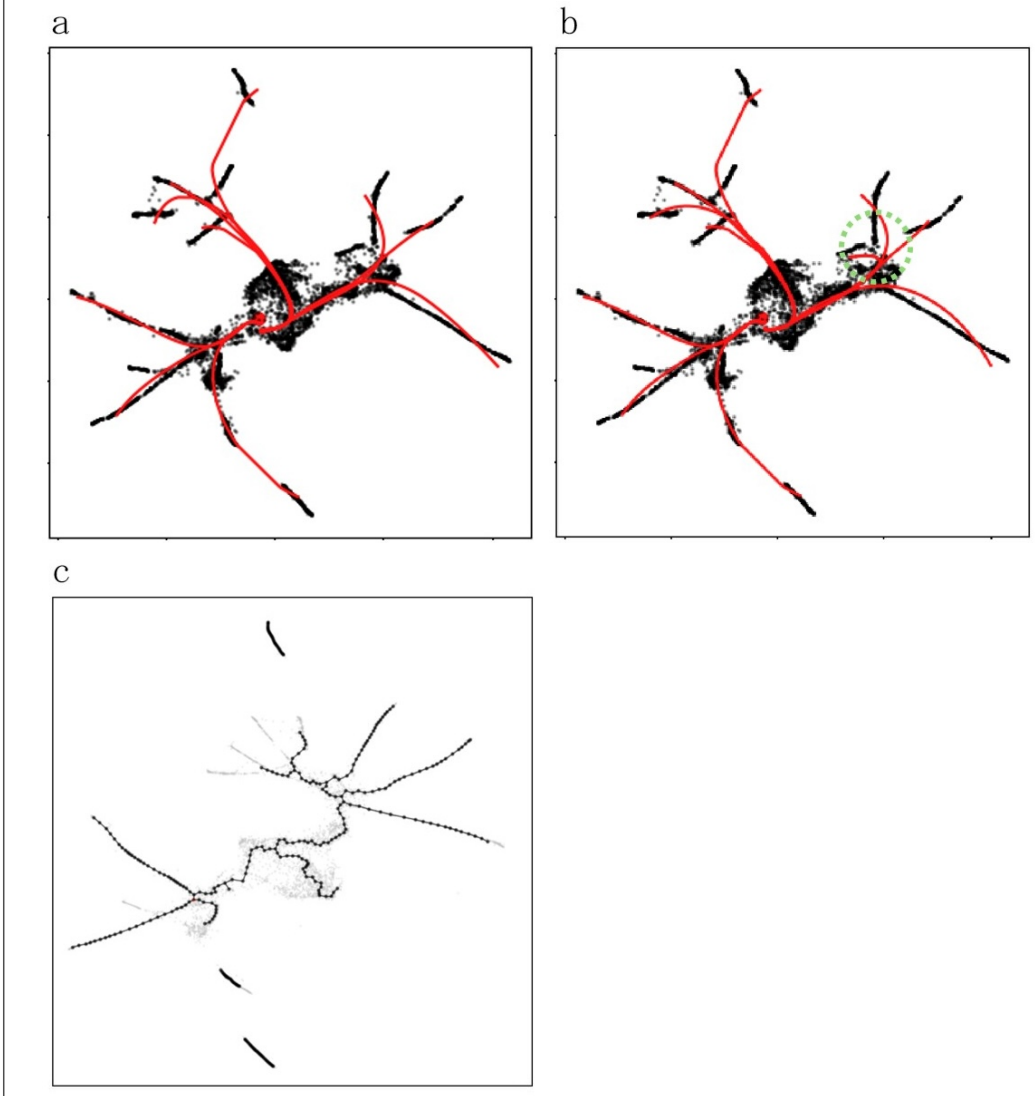

図 12: Monocle3 と Slingshot の解析結果の比較図。C. elegans のデータを使用している。**a** 制約有りの Slingshot の解析結果 **b** Slingshot の解析結果 **c** Monocle3 の解析結果

4 データ

検証に使用したデータについて説明する。

4.1 Bone marrow mononuclear cell

12 人の健康なヒトドナーから収集された骨髄単核細胞のシングルセルマルチオミクスデータである。(生データは、NCBI の GSE194122 からダウンロードできる。前処理済みデータは次のリンクから確認できる:https://openproblems.bio/events/2021-09_neurips/documentation/data/dataset.html) データのダウンロードには、単一細胞ゲノミクスにおける最適輸送アプリケーション用のフレームワークである MOSCOT[7] を使用した。具体的には、MOSCOT の「Translating multiomics Single Cell data」チュートリアルに従い、データの処理を行った。(MOSCOT のチュートリアルに関する詳細は、次のリンクから確認できる:https://moscot.readthedocs.io/en/latest/notebooks/tutorials/600_tutorial_translation.html)

4.2 Human fetal immune cell

ヒト胎児免疫データセットである。データセットのうち検証には NK 細胞関連のデータのみを使用している。(データは次のリンクからダウンロードできる:https://github.com/alexQiSong/scSTEM_sample_data/tree/master/human_fetal_immune) データの前処理には scSTEM を使用した。scSTEM はシングルセル RNA シーケンシングデータの疑似時間順序に基づいて遺伝子の動的プロファイルをクラスタリングする方法である [8]。scSTEM を用いて得られた前処理結果を使用した。

4.3 Hematopoiesis

本 scSeq を使用して DNA バーコードで細胞をクローン的にタグ付けしたマウス造血データである [6]。系統追跡と統合された単一細胞トランスクリプトミクスから細胞ダイナミクスを推測することができる CoSpar[5] を用いて前処理を行った。(前処理の方法については、次のリンクを参照して実施した:https://cospar.readthedocs.io/en/latest/20210121_all_hematopoietic_data_v3.html) DNA バーコードは、特定の DNA 領域にランダムな挿入するバーコードのことである。挿入されたバーコードを細胞が分裂するたびに引き継ぐことにより細胞の系統的な起源を識別することができる [14]。

4.4 C. elegans

線虫 C. エレガンスの初期胚発生細胞のシーケンスデータである [10]。前処理には Monocle3 を使用し、前処理済みのデータを利用している。(前処理の方法については、次のリンクを参照して実施した:https://cole-trapnell-lab.github.io/monocle3/docs/trajectories/)

5　議論

5.1　クラスタ数と分岐認識能力

クラスタ数の変化は小全域木の形に影響を与えることで Slingshot の軌道予測結果を変化させる。図 13 にクラスタ数による最小全域木の比較を記載した。自動決定されたクラスタ数 76(図 13.a) を基準として考えると、クラスタ数 35(図 13.b) の場合はクラスタ数が小さすぎるので細かな分岐が 1 つのクラスタとして認識され分岐がうまく認識されない。一方クラスタ数 100(図 13.c) の場合は、クラスタ数が大きくクラスタ数 76(図 13.a) には無い分岐が発生している。クラスタ数を適度な範囲に設定することは軌道予測において重要であることが示されている。

5.2　解像度とクラスタ数の関係

指定したビン境界ごとに、画像内の連結オブジェクトの数、クラスタ数、オブジェクトの面積の変化を表 3、表 4、表 5、表 6 に示した。オブジェクトの面積は、データの x, y 領域を切り出した際の長方形の面積を基準として相対値で表している。ビン境界は、ヒストグラムを作成する際にデータをどのようにビン分割するかを指定するパラメータであり、ビン境界が高いほど画像の解像度が高くなる。解像度が低すぎると画像が荒くなり、オブジェクト数が減少する傾向がある。それに伴ってクラスタ数も減少する傾向にある。一方、解像度が高すぎると各データ点が個別に認識され、オブジェクト数とクラスタ数が増加する傾向がある。ただし、解像度が増加することでクラスタ数が必ずしも増えるわけではないため、複数の解像度を試した結果、最適なものを選択する必要がある。

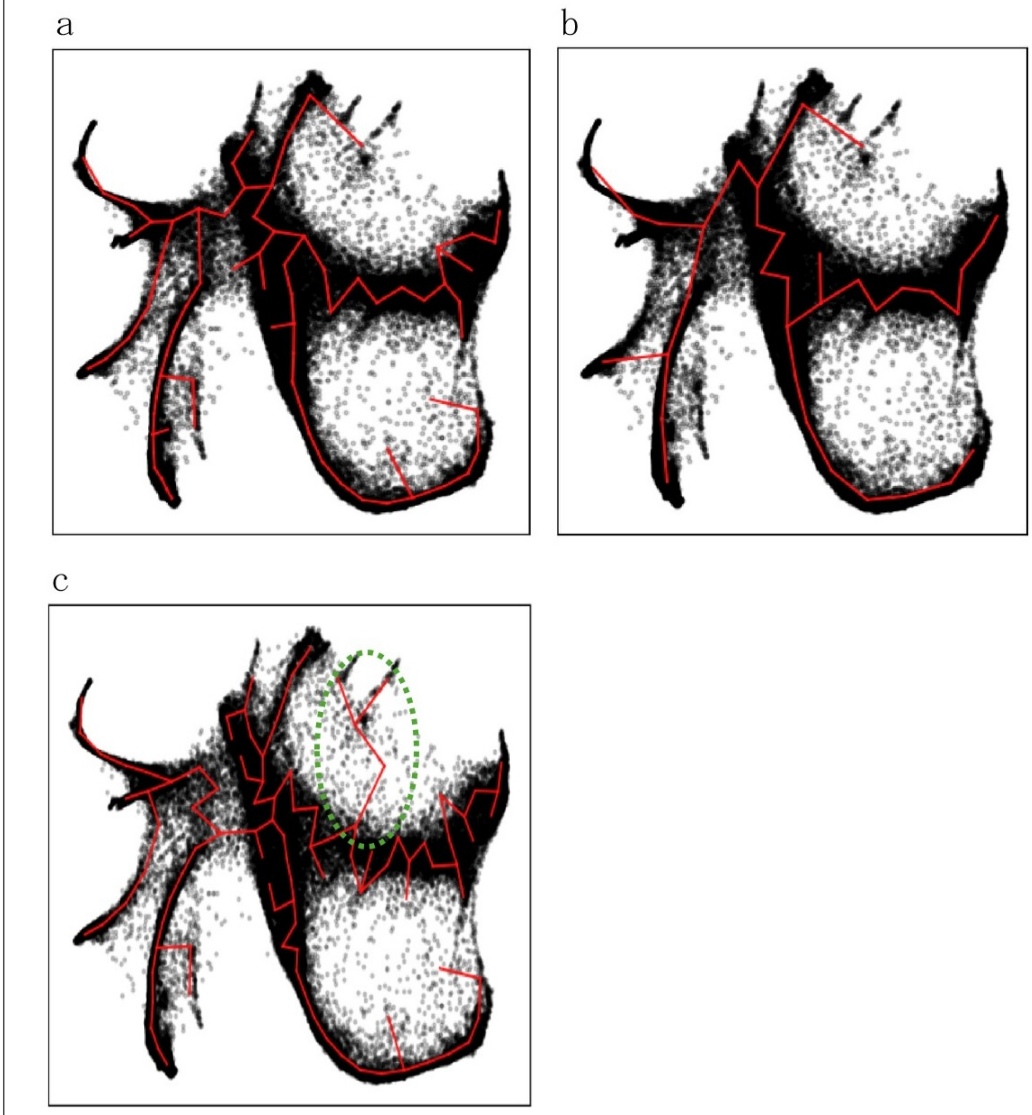

図 13: クラスタ数による最小全域木の比較図。Hematopoiesis のデータを使用している。**a** クラスタ数 76 **b** クラスタ数 35 **c** クラスタ数 100

ビン境界	オブジェクト数	クラスタ数	面積
10	1	4	0.9
20	8	8	0.245
30	3	15	0.919
40	3	30	0.826
50	11	25	0.269
60	17	51	0.237
70	9	60	0.283
80	16	82	0.247
90	44	128	0.173

表 3: Bone marrow mononuclear cell

ビン境界	オブジェクト数	クラスタ数	面積
10	1	0	0.9
20	2	5	0.345
30	4	23	0.804
40	19	26	0.633
50	19	50	0.638
60	10	42	0.293
70	10	56	0.311
80	12	60	0.27
90	25	104	0.229

表 4: Human fetal immune cell

ビン境界	オブジェクト数	クラスタ数	面積
10	1	6	0.37
20	8	15	0.315
30	11	19	0.289
40	1	16	0.968
50	5	46	0.817
60	27	83	0.627
70	33	112	0.646
80	36	168	0.66
90	22	58	0.221

表 5: Hematopoiesis

ビン境界	オブジェクト数	クラスタ数	面積
100	18	76	0.216
200	87	331	0.197
300	365	729	0.22
400	775	1242	0.161
500	1231	1544	0.129
600	2187	1959	0.062
700	2535	2465	0.062
800	3050	2603	0.048
900	3710	2774	0.036

表 6: Hematopoiesis

6　結論

本研究では、細胞系統擬似時間解析における課題を克服するために、画像処理技術と次数制約付き最小全域木を組み合わせた新たな手法を提案した。擬似時間解析ツールである Slingshot は、最小全域木を使用して細胞分化の軌跡推定を行うが、生物学的に「細胞状態は一度に 3 つ以上に分岐しない」という制約を考慮していなかった。またクラスタリングにおいて適切なクラスタ数の決定方法が明示されていないという課題も存在していた。本研究では、Slingshot に対して次元削減後の細胞データを画像として認識し、細線化処理を施すことで、細胞集団の大まかな分岐構造を抽出し、これを用いて適切なクラスタ数を決定する方法を提案した。得られたクラスタ数と重心情報を用いて、次数制約を導入した最小全域木を構築し、「細胞状態が一度に 3 つ以上に分岐しない」という生物学的制約を反映した細胞分化モデルを構築することで、より生物学的妥当性の高い擬似時間解析を実現した。本手法の有効性を検証するために、Bone marrow mononuclear cell、Hematopoiesis、Human fetal immune cell、C. elegans の 4 つの公開データセットを用いた実験を行った。その結果、提案手法は従来手法と比較してより整合性のある分岐構造を再現し分化経路の推定精度を向上させることが示された。また本手法によりクラスタ数の自動決定が可能となったことで、解析者が事前にクラスタ数を指定する必要がなくなり、解析の利便性が向上した。

しかし課題が残されている。一般的に擬似時間解析の手法は細胞の発現プロファイルの空間的な近さが分化の時間軸と関連するという前提に基づいている。その前提に基づけば、本研究のようにデータを画像化しデータの輪郭からデータ構造を分析すること自体は妥当であると考えることができる。画像から適切なクラスタ数を決定する手法に関して RNA シーケンシングデータの特徴量抽出とその理論的根拠の解明が今後の課題となる。本研究ではクラスタ数を決定する際に分散値を用いたが、より適切な特徴量の選定には至らなかった。した

がって、クラスタ数決定の精度向上には、使用する特徴量の選定とその理論的な検証が必要である。総括すると、本研究では画像処理と次数制約付き最小全域木を組み合わせることで、細胞分化の軌跡をより生物学的に適切な形で推定する手法を提案し、その有効性を示した。本手法は従来手法の課題を解決し、より解釈しやすい細胞系統擬似時間解析を実現するものであり、今後の細胞系譜研究において重要な貢献を果たすと考えられる。

謝辞

本研究を進めるにあたり御指導、御助言を賜り、暖かく見守ってくださった阿久津達也教授に深く感謝いたします。日頃から数々の有益な御助言、御協力をいただきました情報学研究科の皆様に深く感謝いたします。さらに、本研究を行うにあたり、有益な協力をいただいた香港大学博士課程 Jiaying Zhao 氏にも、深く感謝しここに誠意を表します。

参考文献

[1] 坂本智子, 前伸一, 長船健二, 岡田千尋, 樺井良太朗, 渡辺亮: 細胞運命決定機構を明らかにするシングルセル遺伝子発現解析, 日本薬理学雑誌, Vol. 153–2, pp. 61–66 (2019).

[2] 仲嶋なつ: シングルセル解析の動向と展望, *JSBi Bioinformatics Review*, Vol. 3–2, pp. 61–74 (2022).

[3] K.Street, D.Risso, B.Russell, D.Diya, N.John, Y.Nir, P.Elizabeth and D.Sandrine: Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics, *BMC Genomics*, Vol. 19–447 (2018).

[4] D.Welch, J.Hartemink and F.Prins: SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data, *BMC Genomics*, Vol. 17–106 (2016).

[5] W.Shou-When, J.Herriges, H.Kilian, N.Kotton and M.Klein: CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information, *Nature Biotechnology*, Vol. 40, pp. 1066–1074 (2022).

[6] C.Weinreb, A.Rodriguez-Fraticelli, F.D.Camargo and A.M.Klein: Lineage tracing on transcriptional landscapes links state to fate during differentiation, *Science*, Vol. 367 (2020).

[7] K.Dominik, P.Giovanni, L.Marius, K.Michal, P.Zoe, G.Manuel, M.Laetitia, S.Michael, S.Lama, J.Changying, B.Aimée, C.Perla, T.Marta, P.Shrey, G.Ilan, L.Heiko, B.Mostafa, N.Mor, C.Marco and J.Theis: Multi-

modal cell mapping with optimal transport, *Nature*, Vol. 22 (2025).

[8] S.Qi, W.Jingtao and B.Ziv: scSTEM: clustering pseudotime ordered single-cell data, *BMC Genomics*, Vol. 23–150 (2022).

[9] Q.Xiaojie, M.Qi, T.Ying, W.Li, C.Raghav, A.P.Hannah and T.Cole: Reversed graph embedding resolves complex single-cell trajectories, *Nature Methods*, Vol. 14, pp. 979–982 (2017).

[10] S.P.Jonathan, Z.Qin, H.Chau, S.Priya, P.Elicia, D.Hannah, S.Derek, T.Kai, T.Cole, K.Junhyong, H.W.Robert and I.M.John: A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution, *Science*, Vol. 20–365 (2019).

[11] N.Nabil and S.Rajjan: An investigation into the skeletonization approach of hilditch, *Pattern Recognition*, Vol. 17, pp. 279–284 (1984).

[12] 岡本吉央: 完全双対整数性：全域木 (1), 電気通信大学講義スライド, p. 9 (2014).

[13] 上原隆平: 計算幾何学特論 凸包, 北陸先端科学技術大学院大学講義スライド, p. 2 (2014).

[14] IkSooKim: DNA Barcoding Technology for Lineage Recording and Tracing to Resolve Cell Fate Determination, *Cells*, Vol. 13 (2024).

[15] G.Mehdi: Integer Programming Formulations for Minimum Spanning ForestForest Problem, *The University of Arizona, College of Science Mathematics Lecture Slides*, p. 11 (2014).

付録:本研究にて使用したソフトウェアとコンピュータ

A.0.1　ソフトウェア

Python 3.9.19

GurobiPy 10.0.3

NumPy 1.26.4

Gurobi: https://www.gurobi.com/jp/products/gurobi/

PySlingshot: https://github.com/mossjacob/pyslingshot

本研究のソースコード: https://github.com/iwakitakuma33/degree_restricted_slingshot

A.0.2　コンピュータ

PC: MacBook Pro

Chip: Apple M3

Memory: 24GB

OS: macOS 14.6.1

A.0.3　翻訳について

英文の内容梗概を作成するにあたり DeepL(https://www.deepl.com/ja/translator) を用いて文章を校正した。