

遺伝子のみを考慮した細胞データに対し、あらかじめ階層的クラスタリングで推定した個数の細胞クラスごとに異なる多項分布を仮定する。次に、そのクラスの下で各細胞を出力する確率分布のパラメータを発現プロファイルから推定する。その後、各クラスが各細胞を出力する確率の対数（対数尤度）をもとに、事後確率が最大となるクラスタリングを求めることで細胞分類を行っている。

2.2.2 筋芽細胞における細胞の分類

Trapnellらは単一細胞 RNA-seq 解析を応用してヒトの筋芽細胞の時系列遺伝子発現解析を行っている⁹⁾。ここでは、複数時点での RNA-seq データを用いることで、トランスクリプトームのダイナミクスを考慮していることに注意されたい。具体的に、分化による細胞の発展時系列を以下のような教師なしアルゴリズムで計算している。まず、発現プロファイルから独立成分分析を用いて細胞の遺伝子発現データの次元圧縮を行う。次に、圧縮された発現空間上において、各細胞をグラフの頂点、各頂点間の距離を辺の重みとみなしたときの重みつき完全グラフを考え、その最小全域木を計算する。そして、最小全域木の中の最長パスを探索し、パスを構成する頂点系列（細胞群）を転写に関連する類似の細胞とみなし、それらは分化を通じて発展したものであると判定する。

2.3 Non-coding RNA と細胞情報解析

これまで RNA の全体を表すトランスクリプトームの解析に主眼を置いたが、これらは原理的にタンパク質に翻訳されない non-coding RNA も含んでいる。ヒトのような高等生物では多種多様な non-coding RNA の制御機構が複雑な生命システムを維持していると考えられており、これまで述べてきた遺伝子発現量は non-coding RNA の遺伝子発現制御機構による影響を強く受けていると思われる。そのため、細胞情報解析を行うにあたり、RNA-seq などにより得られた遺伝子発現量という情報のみを考慮するのではなく、遺伝子発現現象の背後にある non-coding RNA の分子機能や他の生体分子との相互作用を併せて多角的に考察することが、今後精度の良い解析結果を得るために必要になってくると思われる。具体的に、現在までに大量配列解析に適した高速 RNA 配列情報解析ツールセット¹⁰⁾などが提供されているので、それらとの連携を模索することが新たな知見を得る手がかりとなるかもしれない。

2.4 細胞の周期表

単一細胞解析と細胞分類から得られる結果により、そもそも細胞は何種類存在するのか、という問題を再考することができる。一般に、細胞は 200 ~ 400 種類あると言われているが、遺伝子発現量といった細胞の内部状態を考慮すると、より精緻な細胞のカテゴリーを示すことができる可能性がある。化学における構成単位が原子であり、その周期表が存在するように、生物の構成単位である細胞にも周期表に相当するものが存在するかもしれない。周期表の存在の有無に対する初等的なアプローチとして、当研究室では、マイクロアレイで取得したヒトの約 3,000 細胞の遺伝子発現量から得られるヒートマップを用いて細胞を自動的に分類し、分類後の細胞を系統的に並べ表にした結果、組織ごとに細胞のクラスターが形成されていることを確認している (<http://shogoindb.stemcellinformatics.org/>)。これは細胞間の関係において周期表のような細胞間の規則を予想させる結果である。

細胞の辞書ともみなせる周期表を完成させるためには、細胞の法則などを記述する数理モデルを推定する必要がある。細胞を支配する法則の解明のために、ビッグデータを利用した網羅的な細胞情報解析が鍵を握るのは想像に難くない。細胞周期表が完成した暁には、工学的に効率良く細胞を設計することが可能になるなど、その応用範囲は広いと予想される。

2.5 現状の問題点と今後の展開

上で紹介した内容は発展途上の研究分野であり、現状では様々な問題点がある。まず、単一細胞解析から得られる RNA の取量が非常に小さく、配列解読に至るまでに技術的なノイズが混入する可能性が高い。また、細胞の中で起きる遺伝子発現は環境によっても変化するので、生物学的な揺らぎのある確率的な事象と考えるべきである。したがって、これらの因子を柔軟に扱える数理モデルを開発することが必要であろう。また、本節では主にトランスクリプトームに

よる解析について言及したが、DNAのメチル化やヒストン修飾を代表とするエピジェネティクスも細胞を理解する上で重要な要素となるため、これらを扱えるように解析手法を発展させていくことが求められる。さらに、これまでの解析法は組織ごとに着目して行われていたが、将来的には生物そのものをまるごと細胞解析することで、細胞レベルで見た生物の設計図を一度に作るができるようになるかもしれない。

おわりに

遺伝子発現量などのビッグデータを用いて細胞解析を行うと、これまで見えてこなかった細胞間の関係性が見えるようになる。すると、我々はまだ細胞のごく一部を知っているに過ぎないのではないかと認識させられる。今後、多種多様のビッグデータが量産されていくと考えられるので、既成の枠にとらわれないデータベースの柔軟な構築とともに、それに連携する高速解析ツールの開発が活発に行われるであろう。これらの技術の創出により、細胞の詳細、ひいては周期表といった細胞の規則が分かるだけでなく、コンピューターを用いて細胞を設計する時代の到来を予感させる。

文 献

- 1) Hatano A et al., Database, 2011, bar046 (2011)
- 2) Taylor FC et al., Nat. Biotech., 8, 889-896 (2008)
- 3) Brazma A et al., Nature Genetics, 29, 365-371 (2001)
- 4) Opinion: Microarray standards at last. Nature, 419, 323 (2002)
- 5) Gene Ontology consortium. Nat. Genetics, 25, 25-29 (2000)
- 6) Tang F et al., Nat. Methods, 6, 377-382 (2009)
- 7) Islam S et al., Nat. Protoc., 7, 813-828 (2012)
- 8) Jaitin DA et al., Science, 343, 776-779 (2014)
- 9) Trapnell C et al., Nat. Biotechnol., 32, 381-386 (2014)
- 10) Kato Y et al., Nucleic Acids Res., 40, W29-W34 (2012)

