1B-06

モンテカルロ法を用いた生物学的要因の最適化法

石井　一夫†　　古崎　利紀‡

東京農工大学農学府農学部　「農学系ゲノム科学人材育成プログラム」†

東京農工大学農学府農学部　「農学系ゲノム科学人材育成プログラム」‡

1、はじめに

筆者らは、次世代シーケンサーやマイクロアレイから得られたデータを用いて、生物学的現象を説明する数理モデルを作成する検討を実施している。そのモデル作成の際の説明変数の選択の組合わせの選択の問題は、組合わせ最適化問題であり、NP 困難であるために、実際のモデルを構築するために、大量の計算が必要になる。このため、モンテカルロ法により説明変数の選択の最適化を行うという工夫を実施している。今回これらの検討状況を報告する。

2、ビッグデータと数理モデル

生物学分野では次世代シーケンサーの実用化により大量の塩基配列データが産生されるようになり、大量の説明変数を持つデータ（要因；説明変数）を扱い生物学的現象を数理モデルなどにより説明することが可能になってきた。

これらの説明変数の組み合わせはたとえ数百程度の数であっても、天文学的な数になる。

したがって、説明変数の最適化問題を解決すること自体が、大量の計算プロセスを必要とするビッグデータ解析であると言える。

これらの大量データ解析に対応するため、メニーコア CPU によるハイパフォーマンスコンピューティング（HPC）、Hadoop などのクラウドコンピューティングによる分散処理、GPGPU(General-purpose computing on graphics processing units; GPU による汎目的計算)による並列計算などを用いて実施する必要がある[1,2]。しかし、実際には、そのようなパワフルな方法を用いたとしても、結局は解決できない問題である。

Optimization of biological factors using Monte Carlo methods

†Kazuo Ishii・Department of Applied Biological Science, Faculty of Agriculture, Tokyo University of Agriculture and Technology

‡Toshinori Kozaki・Department of Applied Biological Science, Faculty of Agriculture, Tokyo University of Agriculture and Technology

3、従来からの数理モデルの最適化法

3.1、経験的方法による順位付けによる説明変数の選択法

従来から数理モデルの最適化には、t 検定による p 値などの経験的指標により、順位付けをし、変数減少法、変数増加法、変数増減法などの方法で説明変数を絞り込む方法が一般に取られる。

現在、我々は徳島大学大学病院精神神経科との共同研究により、変数減少法により多変量解析による DNA メチル化部位のうつ病診断マーカの選択を行ってきた（Numata, Ishii *et al* 投稿中）。実際の、生物学分野における多変量解析において説明変数の選択は、ほとんどがこの経験的方法に基づいており、SAS や、SPSS などの市販の統計ソフトにおいても、この方法を用いて説明変数の選択を行う。

3.2、総当たり法による説明変数の選択法と経験的方法の限界

一方で、この方法により変数の最適化を行ったとしても、必ずしも最適の組み合わせにはならず、総当たりで検討しなければ、最適化な説明変数の組合わせを選択したことにならないことも、多変量解析による遺伝子発現を用いたうつ病診断マーカの選択により示してきた（Watanabe, Iga, Ishii *et al* 投稿中）。

これらは、要因間の交互作用や交絡因子などの影響など不測の要因が絡んでいると考えられるが、これらを考慮した最適化要因の選択は、最終的には、総当たり法によるしかないと思われる[3]。しかし、その組合わせの数は膨大となり、実際に実施することは困難であることが多い。

3.3、モンテカルロ法による説明変数の選択法

そこで、これらの問題を解決するために、モンテカルロ法による無作為抽出により、説明変数を選択し、近似的な最適化を行う方法を検討した。

この方法は、事前の経験的な情報がなくても、いろいろな組合わせ最適化に用いることができる。

筆者らは、進化系統樹を最適化する場合の遺

伝子の選択にこの方法を試み、非常に有望な方法であることを示唆してきた[4]。また、次世代シーケンサーデータの網羅的発現定量データによる殺虫剤抵抗性遺伝子の選択にも有用であることを示してきた（Kozaki, Ishii *et al* 投稿準備中）。

モンテカルロ法による方法であっても大量計算が必要であり、上記検討の中で、並列化計算は計算時間の短縮のために必要であった。これをメニーコア CPU による大量計算やコンピュータクラスタの利用が有用であることも示してきた。

4、モンテカルロ法による最適化法の拡張と検討

筆者らは、このモンテカルロ法による説明変数の最適化法が、有用である可能性から、さらに他の生物現象に応用することを試みた。

一つは、ゲノム編集技術による遺伝子組換えイネにおける次世代シーケンサーによる発現変動遺伝子の検出であり、もう一つは、カビの感染した植物の次世代シーケンサーによる発現変動遺伝子の検出である。

方法として、ゲノム編集技術による遺伝子組換えイネにおける次世代シーケンサーによる発現変動遺伝子においては、遺伝子発現変動の見られた遺伝子群の組合わせを無作為に選び出し、判別分析を行い正確に遺伝子組み換え群と、遺伝子非組換え群を識別できる遺伝子の組合わせで選ばれた遺伝子を選択した。

カビの感染した植物の次世代シーケンサーによる発現変動遺伝子の検出においては、遺伝子発現変動の見られた遺伝子群の組合わせを無作為に選び出し、判別分析を行い、正確にカビ感染群と、非感染群を見分けることのできる遺伝子の組合わせで選ばれた遺伝子を選択した。

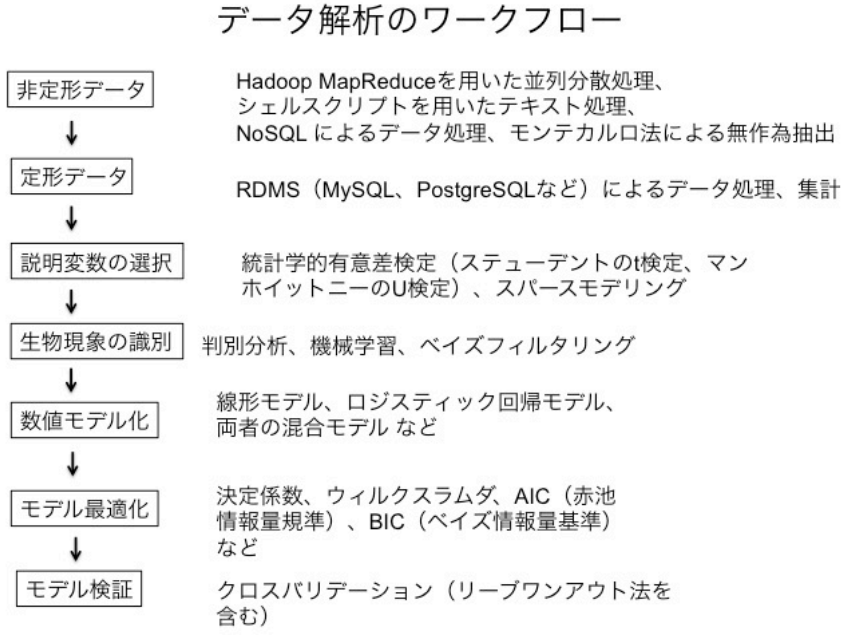
両者とも、従来の t 検定による経験的方法で見つかった遺伝子群とは若干異なる遺伝子が見つかっており、生物学的要因を選択する新しい方法として有効であることを示すことができた。

5、結論

本研究により、モンテカルロ法による数理モデルの変数最適化法や有効であることを示すことができた。詳細な、条件検討はこれからの課題であるが、今後より強力な方法を確立していきたい。

6、今後の展望

現時点では、判別分析による選択法しか検討できていないが、さらに機械学習を用いた方法との比較を実施する予定である。



謝辞

本研究は、文部科学省特別経費「農学系ゲノム科学領域における人材育成プログラム」、科学研究費基盤研究(C)（課題番号 26330325）の資金支援により実施した。また、一部、有限会社ユニバーサルシェルスクリプティング研究所との共同研究により、さらに、AWS in Education Research Grant Award の助成により実施した。本研究で用いた植物のゲノムデータは明治大学研究員中川知巳博士の提供によるものである。ここに感謝の意を表する。

引用文献、参考文献

1) 石井一夫，ビッグデータ：世界を変えていくイノベーションの原動力として：2. 医療におけるビッグデータ利活用 −精神神経系疾患の診断系の開発を中心として−,情報処理; 55, 964-969(2014).
2) 石井 一夫，古崎 利紀，ゲノム科学とビッグデータ分析・データマイニング，日本化学会情報化学部会誌: 32, 4, 80-83 (2014)
3) Michael J.Crawley 著，野間口謙太郎訳，菊池泰樹訳，統計学：R を用いた入門書,共立出版 (2008)
4) 石井一夫，松田朋子，古崎利紀，後藤哲雄：モンテカルロ法を用いた進化分岐図作成法. 研究報告バイオ情報学（BIO）,2013-BIO-36(22),1-6 (2013-12-04).