バイオテクノロジーにおける “Mathematics is the language of Science.” －ブラックボックスに陥らないために－

石井　一夫

バイオテクノロジー分野において，次世代シーケンサーをはじめとした大規模データ解析が日常化し，ゲノムビッグデータによるデータマイニングやデータモデリングの手法を用いて回帰モデルを作成し，有用性産物の最適化条件を決定したり，モニタリングしたりすることが行われるようになってきている．数値データのみならず，画像データや電気信号データを用いて機械学習や深層学習，強化学習といった人工知能技術のバイオテクノロジー分野への応用も進んでいる．

これらの解析には，統計学やプログラミングのスキルは必須になっているが，同時に初心者にも使えるということを謳った分析ツールも普及してきている．高度なデータ分析を初心者でも使えるということは，一見便利なようであるが，中身のアルゴリズムなどが不明で，“ブラックボックス”のまま使われることが多いと思われる．

実際のところ，データマイニングやデータモデリングはもちろんのこと，機械学習にしろ，深層学習にしろ，ブラックボックスであることは決してない．これらの理解には，線形代数や，解析学（微分積分をはじめとする）などの数学の素養は必須となっている．“Mathematics is the language of Science.”は，イタリアの科学者ガリレオが残した言葉らしいが，少なくともデータ分析に関しては，まさにその通りと実感することが多い．というよりは，数学の素養はデータ分析をブラックボックスに陥ることを防ぐことができる．つまり，物質に基づく化学を基礎としたバイオテクノロジーは，情報に基づく数学を基礎としたサイエンスに変わりつつあることを実感することが多い．

ここで，このようなデータ分析の数学的理解の重要さを示すために，回帰分析でよく使われる最小二乗法について考えてみよう．最小二乗法は，誤差を伴う測定値の処理において，その誤差の二乗の和を最小にすることで，もっとも確からしい関係式を求める方法である．回帰分析は，生化学や分析化学において検量線を用いた回帰直線による濃度測定に普通に使われるので，馴染みが深い人も多いと思う．ここで，統計学においてパラメータ推定を行う手法に最尤推定法というのがある．「統計的推定の際，実際に得られた標本があるとき，それが得られる確率が最大になるような母数の値をその推定値とする手法」などと辞書的には説明されている．生物学的には，進化系統樹の作成や，遺伝統計解析に使われる．しかし，測定したデータが等分散ガウス分布（等分散の正規分布）に従っていると仮定できるときには，最小二乗法は最尤推定法と同じ結果が導かれ両者は等価になる．つまり，

「最小二乗法＝等分散ガウス分布＋最尤推定法」

の関係が成立している．要するに，最小二乗法は，最尤推定法の特殊例であり，データが等分散の正規分布に従っている場合には，最尤推定法と最小二乗法から同じ式が導ける．生物学ではまったく違った使われ方をする最尤推定法と最小二乗法が等価であるということはなかなか興味深い．

測定データが等分散ガウス分布である場合に，分布最小二乗法と最尤推定法が等価になるという話題は以下の高校生・大学生向けの数学のウェブサイト（具体例で学ぶ数学：https://mathwords.net/saisyonijoho）に掲載されているので，数式の詳細な誘導に興味がある読者は確認してほしい．これらの関係をきちんと理解しておくことは，データモデリングでは重要である．

機械学習や人工知能の応用は，今後，バイオテクノロジーでは必須のものになると思われる．うまく使いこなすために，単にソフトウェアの分析ツールにデータを流し込んで，出力されてきたデータの生物学的解釈にとどまることなく，そのアルゴリズムや数学的背景にも目を向けよう．“Mathematics is the language of Science.”は，バイオテクノロジーでも真であり，ブラックボックスに陥らないために必要なことである．

著者紹介　久留米大学バイオ統計センター（准教授）　E-mail: ishii_kazuo@med.kurume-u.ac.jp