はじめに ―データサイエンスの時代を迎えて

「バイオインフォマティクス」という言葉が生まれたのはもう20年ほど前であろうか．そのはじまりがいつかは定かではない．しかし少なくとも言えるのは2020年の今，「大量に産生されるバイオデータを処理する必要性から計算機を駆使して必要な情報を抽出する技術は必須のものとなっている」というのは厳然たる事実である．ますますの先端解析技術の革新，裾野の拡大によって，従来，1つの概念であった「バイオインフォマティクス」も多様化の局面を迎えている．1つはピーク技術のますますの先鋭化へ，もう1つは既存の技術を駆使して大量のデータから生物学的意義の抽出へと向かう．後者を定義する明確な言葉はまだ定まらないが，筆者らはそれを「生命データサイエンス」と称して本書を企画した．特に本書ではその題材を仮にがんを中心としたヒトの医学的応用を想定して構成した．ただし，そのノウハウは他のモデル生物，非モデル生物にも援用可能であると考えている．

実際，近年のバイオデータ産生技術は20年前には想像しなかった速度で進歩している．はじめてのヒトゲノムが解読されて20年，公的/私的なデータベースには何十万人分，何百万人分というヒトゲノムデータにあふれる．遺伝子発現等の多層オミクスデータについてもシングルセルレベルでのデータ産生が普及，さらには空間情報を保持した形，例えば病理画像データの各スポットでの解析も実施されるようになった．今後もますます大規模データの産生と解析の流れはその速度を増して展開していくであろう．急速なデータ蓄積を背景に，ヒトに関する分子生物学的理解も飛躍的に拡大した．多くの遺伝子機能とそのネットワークが明らかになっている．20年後のヒトの生物学は，これまで主であった「疾患」をこえて「健康長寿社会の実現」をめざす状態に達しているのかもしれない．そうなればそれは，ヒトが生物学的に規定されるヒトの限界に挑戦しようとするはじめての試みとなろう．計算機科学においては，技術的特異点，いわゆるシンギュラリティーが2045年をめどに訪れるという．以降，開発された人工知能が飛躍的に自己複製的に技術を発展，データをさらに蓄積して，ついにはヒトの知能を凌駕する進化をとげると夢想される．人工知能の本質が大量に蓄積されたデータの活用にあるのであれば，生物学においても飽和量のデータが大きな転換をもたらすときが来るのかもしれない．少なくともリアルワールドにおいて真にヒトを完全に理解し，さらにはヒトを超える領域まで健康/治療を推進しようとするには，これまでにない規模での生物学的データの産生，その解析の深化は必須の要素である．もちろん現在のところ，バイオ関連データはこれらのいわゆる人工知能解析に供するには，そのデータ蓄積量，解析深度は遠く及ばない．またその計測形式，データ形式もいまだ体系化されていない．しかし，一般にあいまい性をのこす生体関連データに比して，本質的にオミクス関連データはより高次の計算機解析にむく．さらにヒトの外に目を転じても，地球上に存在するあらゆる生物はDNA/RNAを遺伝情報として格納し，それをプロテオームとして読み出す同様の分子機構で機能する．生物相を構成するこれらの計測点は膨大な数ではあってもあくまで有限個である．そのすべてには至らなくてもそのシステムを理解するに足る規模での大規模，精密オミクス解析が実践されるような未来こそが，本質的な意味での「生命データサイエンスの時代」なのかもしれない．現在，その黎明期にあって，本書が未来を担う若手研究者への第一歩を後押しするものになれば，と思う．

2021年10月

編集を代表して
鈴木　穣