特集 ゲノム情報科学

観測技術の進展を支えるインフォマティクス
編集にあたって

森下 真一
東京大学大学院
新領域創成科学研究科
moris@gi.k.u-tokyo.ac.jp

久光　徹
(株) 日立製作所
中央研究所
hisamitu@harl.hitachi.co.jp

高木 利久
東京大学医科学研究所
takagi@ims.u-tokyo.ac.jp

2000年夏に宣言されたヒトゲノム配列解読と前後して，ゲノムやバイオインフォマティクスという言葉が，マスコミを通じて入ることが多くなってきた．インフォマティクスという言葉も入っているため，情報処理がどのような形で活用されているかについてご興味の方も多いと思う．情報科学としてどれだけ手ごわい問題があるのか，生物学や医学を進展させる上で重要な情報科学技術は何であるか，ヒトゲノム解読後の今後必要となる新しいインフォマティクスは何なのか，このような疑問も多いのではないか．

実はゲノムやバイオデータの情報処理は20年を超え

る歴史がある．ゲノム配列やアミノ酸配列の類似性を調べるためのアルゴリズムは早期から開発されている．これらの技術はFASTA，BLAST，CLUSTAL W に代表される遺伝子配列解析ソフトウェアの中で活かされている．しかし今日でも解決が待たれる基礎的な問題もある．また，ヒトゲノム配列が確定したのちに必要とされる技術の中にも新しい展開を必要とする問題が多い．本特集中の記事「ゲノム情報科学における情報科学的諸問題」では，アルゴリズムや計算量の観点からの未解決問題や，今後需要が予測される問題を解説している．

ヒトゲノム解読以前に情報科学が実りある成果を収めた例として配列解析技術がある．長さが高々数千塩基程度の遺伝子配列が数百万個集積されたデータベースに対して，類似配列を高速に検索できる技術は特に重宝され，遺伝子配列の新規性の検証や，遺伝子配列のグループ化に威力を発揮してきている．しかし約30億塩基からなるヒトゲノムの出現は新たな問題を提起している．たとえば長大なヒトゲノムを高速に処理するには，長大なゲノム配列へのインデックス生成を必要とするし，ヒトゲノム中の遺伝子構造を高速に解く際に最適化技術は有効である．「ヒトゲノム解読とヒト遺伝子地図の精緻化」では，この周辺の技術的課題を紹介している．

またヒトゲノムの解読と並行して進行している大型プロジェクトとして，マウスや類人猿のゲノム解読がある．複数の生物種のゲノム解読は，生物の進化プロセスを解明する上で貴重なデータ資源を提供している．また，マウス等のモデル生物とヒトのゲノムおよび遺伝子を比較することで，疾患関連遺伝子を同定するのに役立つことも多い．「比較ゲノム解析を中心とする進化ゲノム学の展望」では生物種間のゲノムを比較してゆくことの生物学的な意義について解説している．ゲノム進化を解析するには，どのようなソフトウェアやアルゴリズムが有用であるかを考えながら読むと面白い．

以上はゲノムに関連する話題である．ゲノムにコード化された遺伝子情報はmRNAとして転写され，mRNAはさらにタンパク質へと翻訳される．タンパク質は生体内でさまざまな機能を担っており，mRNAはタンパク質を生成するための鋳型である．しかしタンパク質に比べmRNAの観測の方が容易なため，mRNAの転写量を観測することで遺伝子の発現する量を観測したと考えることが多い．「遺伝子発現量の観測と遺伝子ネットワークの解析－遺伝子の機能解析を目指して－」ではmRNAの観測結果から，遺伝子間の相互作用を表現する遺伝子ネットワークを推定する問題について解説している．

mRNAはタンパク質へと翻訳され，遺伝子として機能を発揮する．機能未知のタンパク質の機能を予測する手がかりとして，その3次元構造を推測することが有効とされているが，難しい問題として残っている．1つのアプローチとして，1次元配列（アミノ酸配列）から3次元構造を推測するホモロジーモデリング法がある．「生物ゲノムの機能予測を目指して」ではこの技術，およびタンパク質の構造予測を競う国際コンテストCASPを紹介している．現実にCASPに参加して好成績を収めた著者らによる報告であり，3次元構造予測の最先端の状況を知るのに役立つ．

このように本特集では，ゲノムからmRNAが転写されタンパク質へと翻訳される過程に沿って，個々の段階で研究開発されているバイオインフォマティクス技術を理解できるよう配慮しながら編集した．しかしバイオインフォマティクス技術が必要となる場面はこれだけにとどまらない．

遺伝子の配列情報は記号列であり，遺伝子の発現情報は量として捉えることができる．このため問題を数学的に定義することが比較的容易であり，情報処理の対象として扱いやすいテーマといえる．しかし数学的定式化が困難な問題が数多くある．たとえば，遺伝子がどのような条件のもとで，どの組織や腫瘍で観測されたかという情報や，細胞内での局在部位は論文の中に言葉として記述されていることが多い．また，生体内でタンパク質を介してどのように情報が伝わるかという知識であるシグナル伝達は生物学的に深遠であり，解明されていない場合が多い．そのためシグナル伝達に関する知識は，研究者が予測した部分を含めて言葉や絵として論文中に書かれている．このような知識を論文中から抽出し知識ベースとして整備することには生物学からの大きな期待があり，「ゲノム情報学と言語処理」に詳しく解説されている．

情報処理に携わる人たちがバイオインフォマティクスの現状を理解していただくことに本特集が役立てば幸いである．

（平成13年11月5日）