ビッグデータと向き合うバイオインフォマティクス

浅井　潔
（産業技術総合研究所生命情報工学研究センター）

バイオインフォマティクスは，生命現象の計算機科学や数学の手法を用いた解明と，その方法論の研究を行う分野である．従来のバイオインフォマティクス研究では，限られた量の，（当時としては）大量のデータを，（当時としては）高速なアルゴリズムを用いて精密に解析し，多くの場合，生物学者による真実の発見の手助け，仮説の検証を行うことが課題であった．ところが，生命科学で扱うビッグデータは，従来の「精密な解析」をデータすべてに適用することを難しくしてしまった．「超高速」で「荒っぽい」解析を行ってでも，ビッグデータを処理し，結果としてより多くの知識を抽出することが求められてきているのである．

バイオインフォマティクスにおいては，2 つの局面で大きな計算複雑度（計算時間と記憶容量）が必要だと考えられてきた．ひとつは，複雑な生命現象の解明であり，もうひとつは大量データの解析である．多くの生体内分子が相互作用，化学変化に基づくネットワークを形成しているが，外界からの刺激の信号としての伝達と，その応答としての遺伝子発現，代謝等は生体内で複雑な情報処理が行われているとみなすことが出来る．これらのネットワークを解析し，シミュレートし，理解することは極めて複雑な対象を相手にする課題であり，膨大な計算量を必要とする．生命現象に関連した比較的大量なデータとしては，ゲノム塩基配列，タンパク質アミノ酸配列，生体内分子の質量分析データ，生体画像データなどがバイオインフォマティクスの研究対象とされ，私自身も「生物の大量データ」という言葉を幾度となく使ってきた．ところが，近年，我々が相手にしなければならない「大量データ」のスケールは，従来よりも数桁高いレベルとなってしまった．

ゲノム配列情報解析の分野では，いわゆる「次世代シークエンサー」の登場により，ゲノム DNA 配列や細胞内 RNA 配列を読み取る作業（シークエンシング）で，1 回の測定あたり数千万本～数億本の塩基配列データ（文字列）が得られるようになり，この膨大な数の塩基配列の処理が必要となっている．シークエンシングでの典型的な課題は，「de novo アセンブリ」，「再シークエンシング」，「メタゲノム解析」，「RNA-seq」などである．「de novo アセンブリ」とは，まだ近縁のゲノムが解読されていない生物のゲノム配列から多数の塩基配列断片をシークエンシングし，断片をつなぎ合わせることによってゲノム配列決定を行う課題である．文字列のジグソーパズルを解く，というイメージである．「再シークエンシング」では，ゲノム配列の断片をシークエンシングし，既に解読された類似ゲノム上に張り付ける（マ

ッピング）することにより，ゲノム配列決定を行う．「メタゲノム解析」では，多数の生物の DNA 配列が混合したサンプルをシークエンシングし，多数の断片の由来する生物の種類と割合，それらを繋ぎあわせてできる遺伝子の種類などを同定する．「RNA-seq」では，細胞内の RNA をシークエンシングし，ゲノム配列に張り付けてゲノムのどの部分がどのくらいの量，転写されているかを調べる．

これらの課題で主に必要な配列情報処理は，「配列断片同士の繋ぎあわせ」と「長い配列への断片の貼り付け」である．「配列断片同士の繋ぎあわせ」を，単純に行おうとすると，すべての配列ペアを比較することが必要だが，膨大な配列本数の２乗に比例する計算を行うことは不可能である．また，配列断片にはシークエンシング誤りが含まれており，問題をより難しくしている．「長い配列への断片の貼り付け」も，古典的なアルゴリズムだと「長い配列」の配列長と，配列断片群の総配列長の積に比例した膨大な計算時間がかかってしまう．現状の解析では，接尾辞配列，インディシング，ハッシング，鳩の穴原理等の技巧を用いて，配列数や配列数に対してほぼ線形な時間で動作するアルゴリズムが多数考案され，用いられている．また，アセンブリに関しては，de Bruijn グラフを用いた手法により，NP 困難なハミルトン経路問題ではなく，効率的な開放が知られたオイラー経路問題に持ち込む手法が用いられている．

ゲノム情報の解析では，単に塩基配列の並びを決めるだけでなく，より高次の構造を求めるための研究も必要である．我々は，RNA 分子の相互作用を予測するために，分子内の水素結合による構造と，分子間の水素結合による構造の両方を考慮した手法の開発を行っている．RNA の分子内の相補塩基（A と U，C と G）の水素結合による構造は，二次構造と呼ばれている．RNA 塩基配列からの二次構造予測は，バイオインフォマティクスにおける古典的な予測問題として長らく研究されてきた．我々は，RNA 二次構造予測を２進空間における推定問題として定式化し，従来主流であった最少自由エネルギー構造の予測ではなく，塩基対の予測精度の期待値最大化に基づく「γセントロイド」という推定量を提唱した[1]．これらの RNA の二次構造予測では，配列長の３乗に比例した計算時間が必要な動的計画法が知られているが，RNA 分子の相互作用を予測するために水素結合を考慮した複合体構造を予測しようとすると，厳密には配列長の６乗に比例した計算時間が必要となる．我々は，分子内・分子間相互作用の同時確率分布を周辺分布の積で置き換える近似と，整数計画法による高速化を実現したソフトウェア RactIP[2]を「京コンピュータ」上に実装し，細胞内の膨大な RNA 分子間の相互作用を予測している．この課題は，単体でも大きな計算複雑度を持つ問題（RNA 間相互作用予測）を，比較的多くの要素（細胞内の RNA 分子の種類，10^5〜10^6）の組み合

わせ（10^{10}〜10^{12}）について解くという，従来にない挑戦であり，数理的なモデル化，アルゴリズムの工夫，スーパーコンピュータの活用のすべてが揃って初めて可能となっている．

ビッグデータとは一見関係なさそうに見えるが，バイオインフォマティクスで数学に関連した話題として，秘密検索についても触れておきたい．製薬会社における研究開発や，個人ゲノムを用いた研究では，データを秘密にしたまま様々な調査・研究・開発を行う必要がある．たとえば，薬の標的となるタンパク質のアミノ酸配列が解っているとき，その配列を用いて Web や公共のデータベースを検索すると，通信路で秘密が漏れてしまう恐れがある．また，たとえ暗号化して検索を行ったとしても，データベースの管理者には何が検索されたかバレてしまうし，データベースがハッキングされると秘密が漏れるかもしれない．もし，検索のクエリを暗号化し，それを復号化することなく検索を実行し，検索結果も暗号化されたまま得られれば，暗号そのものが破られない限り，クエリも検索結果も検索を行った本人にしか知りえない，というシステムが構築できることになる．我々は，類似化合物の検索において，Tversky index（良く使われる Tanimoto 係数の一般的な形）が類似のものを安全に検索するシステムが，準同型暗号で構築できることを示し[3]，ソフトウェアとして実装した．同様の考え方に基づく研究は多くのグループが取り組んでいて，個人ゲノムのデータ解析などにも応用が期待されている．

ところで，どうして秘密検索がビッグデータと関係があるのだろうか．たとえば，製薬会社では，データベース検索を外に対して行うことが出来ないので，データベースはまるごと社内に追ってきて運用する場合が多い．ところが，今日のビッグデータの時代に，有用なデータすべてを社内に取り込み，蓄積し続けることは不可能である．秘密検索は，その問題に一つの解決を与えてくれるかもしれない技術なのである．

[1] Hamada M, Kiryu H, Iwasaki W, Asai K (2011) Generalized Centroid Estimators in Bioinformatics. PLoS ONE 6(2): e16450. doi:10.1371/journal.pone.0016450.

[2] Yuki Kato, Kengo Sato, Michiaki Hamada, Yoshihide Watanabe, Kiyoshi Asai, Tatsuya Akutsu (2010) RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. Bioinformatics 26: 18. i460-i466.

[3] http://www.aist.go.jp/aist_j/press_release/pr2011/pr20111101/pr20111101.html