特集

バイオデータベースの今

1. バイオデータベースの歴史と展望
2. バックボーンデータベースの課題と展望
   a）バックボーンデータベース DDBJ
   b）バックボーンデータベースの標準化：PDBj
3. 配列データベース検索の現在
4. バイオ知識の形成と表現
5. ゲノムデータの視覚化による効果的な理解
6. バイオデータサービス
7. バイオデータベースの技術的問題点

編集にあたって

平川美夏
京都大学化学研究所
バイオインフォマティクスセンター
hirakawa@kuicr.kyoto-u.ac.jp

高木利久
東京大学新領域創成科学研究科
情報生命科学専攻
tt@k.u-tokyo.ac.jp

背景とねらい

バイオサイエンス研究は，今，データベースの将来にその成功の如何がかかっていると言っても過言ではない．2003 年，国際的ヒトゲノムプロジェクトによって，ヒトゲノムの 30 億余の塩基配列が 99.99%の精度で決定された．30 億文字，コンピュータサイエンスの世界では手に余る大きな数字ではない．しかし，ここに至る 13 年間の試行錯誤の過程で 10 倍以上のデータが産出されたと言われ，その中から最も信頼の高いデータを集めた結晶である．これが意味することは，それだけのデータを管理し，国境を越えて共有し，全ヒトゲノムというスタンダードのデータを作り上げたということであり，高度なデータベース管理技術，高速なインターネットの普及，処理能力の高いコンピュータとその活用技術なくしては，決して実現できなかったということである．

ヒトゲノムプロジェクトを経て，ビッグサイエンスの仲間入りを果たしたバイオサイエンスは，ポストゲノムシークエンシング時代を迎え，さらに多様な研究法によって大規模なデータ獲得へと乗り出している．このようなハイスループットスクリーニングによって得られたデータは，測定機器からコンピュータに直接出力され，研究者の結果の分析はコンピュータの前から始まる．大量の採取データがその場でデータベース化され，公共に流通するデータベースの膨大なデータを含めて解析に投入されることも日常的になっている．近年のハイパフォーマンスコンピューティング環境の広がりは，バイオサイエンス研究の展開にも大きく影響しているのである．

一方，バイオサイエンスの目標は，データ獲得にとどまるわけではない．そこから，医学のための疾患の兆候を読み取り，有効な薬物の候補を探り当て，さらに生命の理解へと続いていくのである．つまりデータの解釈こそがその真骨頂なのである．しかし，これまでのバイオサイエンス研究のアプローチからすれば，コンピュータからの大量のアウトプットを前に，多くの研究者は途方に暮れている．今まで見たこともないこのコンピュータが教える結果を，いかに理解し，解釈すべきなのか，そしてなんと名付け，表現すればいいのか．公開されているデータベースの情報を総動員し，文献データベースを総ざらえして解釈に努めるが，容易な作業ではなく，その作業環境も十分には整ってはいないのが現状である．バイオデータベースの整備に，バイオサイエンスの行く末がかかっている理由は，ここにある．

バイオサイエンス分野のデータベース開発は，1960 年代に遡ることができるが，主に文献に発表されたデータの集中管理が目的で始まった．その後も公共のリソースとして，文献発表されたデータはすべて登録する方針をとり，今や未発表の実験データまでも研究者の登録があれば受け入れる巨大なレポジトリとなっている．つまり公共データベースは，研究成果を皆が持ち寄った共有物として構築され，バイオサイエンスの研究を支える貴重な資産を維持している．バイオサイエンスのコミュニティには，この資産の番人としてのデータベースという考えがあり，バイオデータベース独特の文化と言えるかもしれない．こうして公共データベースは成長してきたが，その実，自発的登録に依存しているためデータ記述や質などにばらつきがあることは否めない．しかし，これは公共データベースに限らず，バイオサイエンスのデータが一般的に持ち得る性質でもある．先に述べた研究の成果であるデータの解釈も，この分野ではほとんどが文献中に言葉で表現されている．昨年のバイオ自然言語処理の特集の中村桂子氏の言葉を借りれば，「データが，常に研究者自身による，もう少し広く言うならその時の研究者コミュニティにより解釈を通して意味づけされ，結果は“言語”で表現される」というものが，バイオデータベースのデータになるのである．

大量の実験データが発生し，その解釈には文献データベースを含め，さまざまな既存のデータベースの情報を参照しなくてはならない．しかし，そのデータベースもデータが質量ともに増加し，肝心の知識情報は，記述のばらつきからなかなか目的に適わない．こうしたなかで，コンピュータサイエンスへの期待はますます高まっている．

公共データベースにおいては，品質管理が課題である．由来の異なるデータを編纂していく上で，大量のデータを効率的に処理し管理するだけでなく，意味のある分類や用語の統一の工夫を行い，検索の支援に役立てる必要

がある．データ作成や更新の処理の高速化，人手に頼る作業の支援，自動化などの技術的解決が必要である．また，利用者には，中身が見えない，目的の結果が得られない，結果がなにか分からないなどの状況が恒常的になっている現状の不便さを改善し，さらに高度な情報を得るための検索支援，知識発見の技術の活用が期待されている．

さらに今，バイオデータベースにおいては「統合」が重要なキーワードになっている．バイオサイエンスの究極の目的は，生命の理解と言っていいが，言うまでもなく地球上の生命は実に多様であり，数え上げるだけでも容易ではない．したがって，あらゆる生命情報を結集して生命の本質を捉えたい，というのが切なる願いである．そしてその過程でさまざまな研究手法が生み出され，生命の構成成分のリストアップが進み，ある時間経過の中のスナップショットとして物質や状態を測定することも行われている．これらの多様化するデータをいかに結びつけ，生物の形作りや生きる営みの原理の探求につなげるのか，もはや情報処理技術の活用なくしては，実現し得ない夢なのである．それほどまでに突き詰めなくても，生命現象は非常に複雑であり，たとえば細胞分裂の仕組みを研究する場合でも，分裂のあるタイミングにかかわる特定のタンパク質というように非常に狭い範囲をターゲットにするのが現実的である．したがって，このようなデータを集積して全体像を明らかにすることが求められる．Web 技術によってデータベース間に相互リンクを持たせることは難しくないが，異種のデータベースのデータを現実の生物で見られるような関係で再構築することが，真の統合と言えるだろう．それには，世の中に存在する大量のデータを相手にするための処理速度や処理容量の問題も克服課題であるし，もっとデータの内部や意味に直結した問題もある．オントロジーやセマンティック Web など，バイオよりさらに巨大なインターネット世界の技術もヒントになるかもしれない．今後発生してくるであろう未知のデータも含めたバイオデータベースの有機的な統合を実現するためには，新たな標準化，知識化の手法やそれに合わせたデータ作成，活用のための情報技術の導入が待望されているのである．

本特集の構成

本特集は，「バイオデータベースの今」とあるように，今リアルタイムで，バイオデータベースに取り組んでいる方々に執筆をお願いしている．

1．バイオデータベースの歴史と展望

バイオデータベースの歴史をそのデータの発生源であるバイオ研究を取り巻く状況とともに紹介し，文化的背景にもなじんでもらえるよう意図した．

2．a）バックボーンデータベース：DDBJ
b）バックボーンデータベースの標準化：PDBj

日米欧の国際協力によって維持される 2 大公共データベースの日本拠点のデータベース構築について解説する．（独）科学技術振興機構は，バイオインフォマティクス推進事業（JST-BIRD）の一環として公共データベースの高度化・標準化を支援してきた．その成果も報告いただいた．

3．配列データベース検索の現在

アミノ酸配列データベースから生物学的意味を抽出するための最新の配列検索技術を紹介する．膨大なデータから意味を取り出す検索技術は，今後さらに注目されていく分野になるだろう．

4．バイオ知識の形成と表現

文献などに散在する知識データを計算機上に再現する技術について解説する．言語によって表現された情報を視覚的にも計算機処理的にも効果的に利用するためのオントロジー，ネットワーク表現によるデータベース構築の実際に迫る．

5．ゲノムデータの視覚化による効果的な理解

30 億のヒトゲノムを読み取ったものの文字データではとりつくしまもない．データベースで公開されているゲノムデータとその解釈に必要な多様なデータをバイオサイエンス研究に資するため，統合し視覚化する方法とその実際について解説する．

6．バイオデータサービス

膨大なバイオリソースの効果的な活用のためのポータルサイトの構築，バイオグリッドプロジェクトについて解説する．Jabion は日本語バイオポータルサイトであり，バイオ初心者の読者諸氏にもおおいに参考いただけると思う．

7．バイオデータベースの技術的問題点

バイオデータベースは，コンテンツ指向であるため技術面に注目した解説は少ない．本特集の総括として，バイオデータベースの具体的な問題を指摘し，コンピュータサイエンスとバイオサイエンスの橋渡しを意図している．

先にも触れたが，バイオデータベースには独特の背景があり，コンピュータサイエンスにおけるデータベースの特集とは少し趣を異にしているかもしれない．それをお断りした上で，バイオデータベースの現状を通じて，それが直面している諸問題にコンピュータサイエンスとしての，新たなシーズを見いだすきっかけとしていただけたらと願っている．それは同時にバイオサイエンスの未来を切り開く可能性にも繋がると期待している．

（平成 18 年 2 月 3 日）