生命科学分野のデータベース統合化と生物名称
ーデータ・知識の共有に向けてー

川本　祥子（ライフサイエンス統合データベースセンター）

ライフサイエンスは生物学や医学など、人間をはじめとする様々な生物について、主に分子的なレベルで明らかにする科学である。21世紀に入ってからの生命科学は、次世代シーケンサーによるゲノム解読を中心に、かつてないデータの大量生産時代に突入した（**図1**）。当初10年以上かかると言われたヒトゲノムの解読は、いまやたった数日で解析が終了してしまう。ヒト以外の生物のゲノムもとりあえずゲノムを読んでみましょうと気軽に言うような時代になった。その良し悪しはともかく、生命に関する情報が大量に生み出される現在、情報の蓄積場所であるデータベースが果たす役割は非常に大きいものである。ところが、生命科学系のデータベースは国内だけでも500以上、全世界では1万以上あると言われ、目的のデータを的確に探し出すことは専門家でも困難な状況にある。また、それぞれ独自の仕様で構築されているため、データを比較したり足しあわせて解析したりすることも難しい。これは生物多様性情報の分野においても繰り返し指摘されていることである。

このような状況の中、平成18年より文部科学省委託研究開発事業「統合データベースプロジェクト」が開始された。このプロジェクトは国内生命科学分野のデータベースの基盤整備を目的とする5年間の計画で、データベースへのガイドを果たすポータルサイトの構築、検索技術の開発、国内研究機関で作成されたデータベースの受け入れ、ウェブサービスの標準化や、辞書やオントロジーの構築、文献情報の活用、人材育成等を推進項目として掲げたプロジェクトである。その成果として、分子や文献を合わせ250以上の異種データベース全800万データを一括して検索できる横断検索や、データベースを共通のフォーマットで構築するシステムを開発し国内初のデータベースのアーカイブサイトを開設した。これらのサービスは幅広い社会への還元を考え、全て日本語で構築されており、ウェブサイトより誰でも自由に閲覧が可能である（**図2**）。また、技術開発と並行して、データベースの利用許諾にクリエイティブ・コモンズライセンスの導入や、疾患研究に関わるゲノムデータの共有方針策定など制度面の整備にも取り組んでいる。文科省でのプロジェクトは今年最終年度を迎えるが、平成23年度からはJSTに設立されるバイオサイエンスデータベースセンターを中心に、文科省以外の関係省庁とも協力して国内データベースの統合を推進することが計画されている。統合データベースプロジェクトが担当するのは分子データが主であるため、お互いの情報が蓄積し統合が進めば、今後さらに、生物多様性情報との連携の必要性が高まると考えられる。今回のワークショップではプロジェクトの成果を紹介するとともに、生物多様性に関連してプロジェクトで構築した生物名辞書とその活用事例についても紹介する。

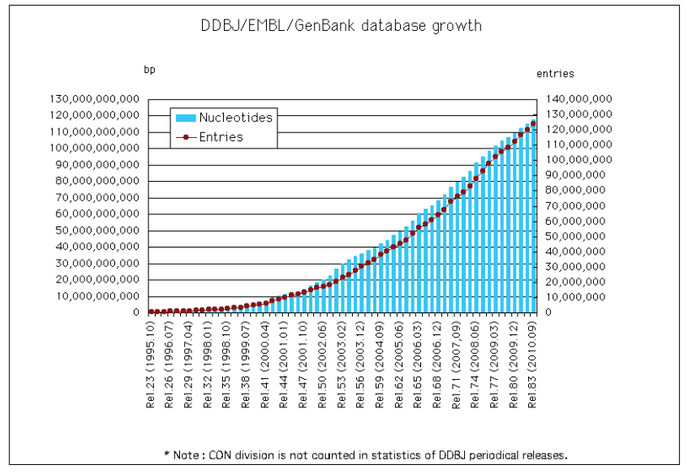

図1　国際塩基配列データベースの登録情報の増加を表すグラフ

図2　統合データベースプロジェクトのポータルサイト「統合ホームページ」
URL は http://lifesciencedb.jp/（H23年度以降変更の可能性があります）