

データベースからの再発見

荒木 通啓

マイケル・ポランニーの著書「暗黙知の次元」には、「発見は、現行の知識が示唆する探求可能性によってもたらされる」という一節がある。ここで言う知識を、先人達が築き上げてきた膨大な知識と置き換えると、それらを一定のルールで整理し直し、知識データの抽出を容易にしたもの、すなわち探求可能性を拡げたものが、今で言うところのデータベースということになるだろうか。

ご存知のように、現在の科学研究におけるデータベースの代表例としては、米国のNational Center for Biotechnology Information (NCBI) のPubMed (文献データベース)¹⁾があげられる。PubMedでは、膨大な文献情報を集積し、ユーザがキーワード検索などによりデータ抽出できることから、知識確認に必要な基盤データベースとしてその地位を築いている。また、個々の研究に特化したデータベースについても年々多様化とともにその数は増大しており、Nucleic Acids ResearchのDatabase Issue²⁾には種々のデータベース群が紹介されている。ユーザとしては、多岐にわたるデータベースの中から個々の研究に見合ったデータベースをどのように選択・利用し、新たな知識の発見につなげていけばよいだろうか。

データベースの第一義的な利用法としては、キーワード検索、(配列・構造)類似検索によるデータ抽出といった辞典的な利用があげられるが、各データベースの特徴を知ることで、より有効な活用が期待できる。実際に、データベースの開発目的は実にさまざまであるが、そのデータ構成に関しては共通点も多くみられ、要素情報(遺伝子・タンパク質・化合物など)、ペア情報(相互作用・制御関係・反応情報など)、それらの関係性を統合したネットワーク情報といったデータが集積されている。ユーザとしては、利用に際してこうしたデータ属性を常に意識しておく、意味ある情報抽出により有効となる。

一方で、個々の研究に個人の思い入れがあるように、データベースにも作り手の個性が色濃く反映されており、その質・量には大きな違いが見られる。たとえば、代謝・酵素反応分野における汎用的なデータベースであるKEGG³⁾・BRENDA⁴⁾を比較してみよう。KEGGデータベースの最大の特徴は、パスウェイマップに化合物・反応・遺伝子データを統合している点にあり、データ集積そのものもパスウェイマップを中心に行われている。このため、化合物エントリとしては天然物が中心である。

一方で、BRENDAでは酵素反応の網羅的なデータ集積を目的として、非天然物を含む反応データも多く含まれており、こうしたデータの偏向性を認識しつつ利用していく必要がある。

また、作り手依存の文献・特許情報キュレーションをもとにしたデータベースの他、近年ではWikipediaの充実ぶりに見られるように、ソーシャルネットワークの知識を集積する枠組みを提供し、集合知を活用したデータベース構築といった取り組みも行われてきている。こうしたデータベースには、配列データベースはもとより、遺伝子発現情報データベースのNCBI Gene Expression Omnibus (GEO)⁵⁾といった実験データそのもののデータベース化も含まれており、今後ますます参加者の増加、データの充実とともに重要性が増していくものと推測される。実際に、配列データベースを利用した類似度比較により遺伝子の機能推定が行われているように、GEOデータベースの遺伝子発現パターン解析をもとに、細胞分類・評価あるいは遺伝子制御ネットワーク同定を行うなど、その応用可能性は多岐にわたってきている。

最後に、データベースからの再発見という観点から、代表的な研究事例を簡単に紹介しよう。Barabasiら⁶⁾は、代謝パスウェイ・タンパク質相互作用に関わるデータを抽出し、独自のネットワーク解析を行うことで、細胞内分子ネットワークの構造的特徴を見だし、その頑強性との相関について明らかにしている。またMiloら⁷⁾は、遺伝子制御関係のデータを利用して、新たにネットワークモチーフの探索・分類方法を開発することで、遺伝子発現制御に重要とされる遺伝子制御モチーフを同定している。こうした研究に共通しているのは、データベースに整理された知識が示唆するモノ・コトを独自の視点・方法で捉え、新たな知識を得ている点にあり、さらなるデータベースからの再発見の鍵は、多様な知識背景を持つ個人の積極的な関わりとその(暗黙)知をいかに発揮できるか、にあるだろう。

- 1) <http://www.ncbi.nlm.nih.gov/pubmed>
- 2) <http://www.oxfordjournals.org/nar/database/c>
- 3) <http://www.genome.jp/kegg/>
- 4) <http://www.brenda-enzymes.info/>
- 5) <http://www.ncbi.nlm.nih.gov/geo>
- 6) Barabasi, A. L. *et al.*: *Nat. Rev. Genet.*, **5**, 101 (2004).
- 7) Milo, R. *et al.*: *Science*, **298**, 824 (2002).