バイオインフォマティクスを使い尽くす
秘訣教えます！【第1回】

ウェブ上に散在する情報を生命科学研究にどう役立てるか

坊農　秀雅

新連載「バイオインフォマティクスを使い尽くす秘訣教えます!」をはじめるにあたって

今日の生物工学研究において，研究対象となる生物のウェットなバイオリソースだけでなく，ウェブ上に存在する多種多様なデータベース（DB）から得られる情報（ドライなバイオリソース）も同様に重要であり，そのコンテンツの取捨選択が研究の進展に大きな影響を及ぼしうることは周知の事実である．多くの研究者は，DBやウェブブラウザからアクセスできるツール（ウェブツール）の恩恵にあずかり，日々の研究を進めている．しかしその一方で，数多くのウェブサイトから必要な情報を正確に抽出してくる方法や，日々進化するウェブツールの扱い方に頭を悩ませている研究者が多いのも現実である．これはまた，現状のDBやウェブツールが，すべてのユーザーにとって満足できるものになっていないためであるとも考えられる．

そこで本連載では，ウェブ上に散在する情報を自らの研究にいかに役立てるか，複雑化する研究の詳細をできる限り日本語環境で情報収集・解析する方法について，その手助けとなる方策をこれから7回にわたって解説する．また，ばらばらに構築されたDBを統合的に扱うために整備されつつある新しいDB統合化技術について紹介し，日々生み出される多様・膨大なウェブデータから必要な情報をより効率的に得る秘訣を伝授する．

予定している連載の内容は以下のとおりである．

1. イントロダクション・情報収集の秘訣
2. 英語論文からの情報収集（PubMed）
3. 日本語で作成されたコンテンツ（統合TV，新着論文レビューおよび領域融合論文レビュー）の紹介
4. 核酸配列DBの現状とその有効活用方法
5. パスウェイ解析とネットワーク可視化ツールCytoscapeの利用
6. RDFによるDB統合化技術（NBDC RDF portal）の紹介
7. 新規モデル生物でゲノム解析する際に必要なアッセンブルやゲノムアノテーションの実際

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

はじめに：研究でも「ネットで検索」

みなさんは研究に必要な情報をどうやって入手しているだろうか？自らの分野の研究に有用な情報がさまざまなDBとして公開され，誰でも使える状況になっている．そればかりか，学会の年会などの情報や，旅行に必要な新幹線，飛行機やホテルの予約のみならず，試薬のカタログなどもインターネット上で検索（いわゆるネット検索）できるようになってきている．もっと言えば，自ら探しに行かなくても，受動的にダイレクトメールとして電子メールや紙媒体でさまざまな情報が送りつけられてきている状況ですらあろう．これは生命科学分野に限ったことでなく，まずネット検索せよという意味の「ググレカス」というスラングがあるぐらいである．ウェブ上にある情報を利用する第一段階としてまずは自ら**ネット検索する**ことを実践してもらいたい．そして，マニュアルを読んで一から使い方を覚えようとするのではなく，**まず使ってみる**ことが大事である．

しかし残念ながら，そのネット検索だけでは必要な情報を見つけ出してくるのが容易でない状況にすぐに直面するだろう．では，どうしたらいいか？本稿では，ウェブ上に分散する情報を自分の研究にいかに役立てるかの方法・秘訣について指南する．

一歩進めて：生命科学DB横断検索

広く一般に使われているネット検索は汎用であり，研究目的で欲しい情報にすぐに辿りつけないことがままある．たとえば，「高血圧」で検索してみるとどういう

著者紹介　大学共同利用機関法人 情報・システム研究機構データサイエンス共同利用基盤施設ライフサイエンス統合データベースセンター（DBCLS）（特任准教授）　E-mail: bono@dbcls.rois.ac.jp

ことが起こるか，実際やってみて欲しい．サプリメントや代替医療の広告がならび，関連する遺伝子の情報などにすぐにはたどり着けない．ちなみに，論文情報を検索するときには通常のネット検索は使わず，論文情報に特化したGoogle scholar[1]や，米国National Center for Biotechnology Information（NCBI）が作成維持管理しているPubMed[2]で検索する．PubMedに収録された英語文献情報の活用方法は，本連載2回目の次回に詳しく紹介する．

また，ネット検索には，一般によく使われそうなウェブサイトが上位に表示されるようになっており，実は検索漏れが多々ある．すなわち，**すべてのインターネット上のページがネット検索の対象になっているわけでない**．たとえば，文献情報はPubMedから検索しないと結果が得られない．また，DDBJ/ENA/GenBankに収録されているはずの塩基配列やアミノ酸配列断片をネット検索しても有用な情報が得られないように，配列類似性検索は不可能で専用のウェブサイトで行う必要がある．つまり，**ネット検索は万能ではない**ということである．先ほど書いたことと矛盾するようであるが「ググるなキケン」なわけである．

この必要な情報がインターネット上のどこにあるのかわからない，といった状況は今に始まった状況ではなく，ヒトゲノム配列解読が完了した2003年以降，指摘されてきたことであった．その状況を改善すべく，2006年から生命科学系DBの利便性を向上させる「統合データベースプロジェクト」（文部科学省）が行われてきた．その主力サービスの一つとして日本語を母国語とする生命科学研究者用に筆者の所属しているライフサイエンス統合データベースセンター（DBCLS）で開発されたネット検索が，JSTのバイオサイエンスデータベースセンター（NBDC）で引き続き維持管理されており，生命科学DB横断検索と呼ばれている（図1）[3]．生命科学DB横断検索の特徴は以下のとおりである．

1. 広告情報等なし　最近ではスマートフォンアプリで利用料を払うと広告が表示されないという機能があることからわかるように，ネット検索はウェブサイトに表示される広告収入によって発達してきた．その結果，ユーザーは利用料を支払わなくても高機能なネット検索の恩恵にあずかることができる．その代わり，検索結果も広告主にとって都合がいいように順番が操作されている．生命科学DB横断検索ではこういったことがないよう，生命科学研究に適した「ネット検索」が実現されている．「高血圧」で検索した結果は，図1のように生命科学系DBに対する検索結果とともに関連する遺伝子に関する情報が表示される（図1右カラム上部）．

図1．生命科学DB横断検索

2. グーグル八分されているところにも届く　グーグルの基準に該当するウェブサイトが，グーグルの検索インデックスから除かれ，検索しても内容が表示されない状態になることを「村八分」になぞらえてグーグル八分という．しかしながら，研究目的ではグーグル八分されているページに閲覧したい情報がある場合もある．たとえば，NCBIのウェブサイトにある情報はグーグル検索では引っかかってこないこともあるのは経験しているのではなかろうか．ネット検索は広く浅く検索用のインデックスが作成されているのに対し，生命科学DB横断検索では選ばれたDBに対してのみ，深くインデックスを独自に作成しており，そのDB数は約500となっている[4]．

3. 日本語でも翻訳して検索　とくに生命科学系では日本語の情報は乏しく，英語でなら情報がある状況が多々ある．その状況を鑑みて，生命科学DB横断検索ではライフサイエンス辞書（LSD）を利用して，生命科学系の専門用語の検索語を英語に翻訳し，それも検索語として加え，それと併せた検索結果が得られるようになっている[5]．たとえば，図1の「高血圧」で検索した結果ではその英語の「hypertension」も同時に検索した結果が得られている．

何を検索したか他人に知られるかもしれない，不安だという考え方もあろう．この生命科学DB横断検索はそういった個人が何を検索したか特定されるような情報漏えいがないよう，細心の注意を払っている．Google Analyticsなどの外部のログ解析ツールを使わない，検索履歴情報から個人が特定できるような情報は公開しな

いなど，管理を徹底している．より多くの方に実際に使っていただいて，情報検索するメリットを享受してもらいたいからだ．

欲しいDBを探す：生命科学DBカタログ

かつては生命科学系のDBといえば，塩基配列はGenBank，タンパク質配列はSwissProt，タンパク質立体構造はPDBであった．現在，それらの後継が存在するものの，それらを取り巻くさまざまなDBが開発され，種類は多様になり，その総数がわからないほどである．そして，そのデータの質もさまざまである．そういった状況の生命科学系のDBを，単純なキーワード検索だけでなく，どんなDBがあるか，研究対象や対象生物種で絞り込んだり，それらの分類で眺めたり，できないものか？

その願いを実現しているのが，IntegbioDBカタログである[6]．これは，ネット検索が一般的になる前にあったインデックス型（かつてのYahoo!）のポータルサイトで，DBの情報が日本語でも維持管理されており，原稿執筆時点で約1500件のDBが登録されている（図2）．

2016年3月からの新機能として，チュートリアル動画でDBの使い方を紹介する統合TVの該当動画へのリンクがつくようになった[7]．DBのアクセス可否などきっちり維持管理されており，すでに維持されなくなったDBに関しても，左カラムの「稼働状況」からそれとわかるようになっている．

またDBが維持されなくなる前に消滅の危機から保全する，生命科学系DBアーカイブもNBDCによって維持管理されている[8]．いわゆるDBの「永代供養」サービスであり，データ説明（メタデータ）を統一して検索を容易にするとともに，利用許諾条件（ライセンス）とDB作成者のクレジットなどを明示し，多くの人が容易にデータへアクセスしダウンロードを行えるようになっている．原稿執筆時点で113件のDBがアーカイブされており，2016年3月にはアーカイブされたすべてのDBに，デジタルコンテンツに付与される国際的な識別子DOI（Digital Object Identifier）が付与されるようになり[9]，DBを引用するハードルが下がった．DBを作成する研究は論文という形での報告はされにくく，引用文献リストに入りづらい傾向があるが，こういったDBを利用した研究発表をされる際には今後はぜひ引用をお願いしたい．なお，DBの寄託は随時募集している．

図2．IntegbioDBカタログ

こういったデータ再利用に向けたサービスは，上述の日本の統合DBプロジェクトによる国産のもの以外にもfigshare[10]やwikimedia commons[11]といったものがある．figshareは前述のDOIが自動的に付与され，その後の引用が容易になるので，筆者も自分の発表のスライドやポスターのアーカイブとして利用している[12]．

DBやそのデータ解析に対する誤解

我々統合DBプロジェクトに関わっているものはDBを広く使ってもらうため，日本全国各地でDB利用講習会を開催したり，各種学会の年会でブース展示などしている[13]．そういった時にDBを利用するうえで困っている点など，気軽に相談していただければ，と思う．何を隠そう，本連載も統合データベース講習会AJACS薩摩がきっかけとなって企画されたものである[14]．とくに初めて使う種類のデータの場合，その道のプロに相談するのが一番である．ここで誤解して欲しくないのは，データ解析のやり方はお教えできるが，**実際にやるのは生物研究者自身である**ということである[15]．データ解析は，タダではない．ウェットの実験同様，コストがかかるし，良い結果が得られたいのであれば積極的にコストをかけるべきである[16]．

そういった時の相談でよくあるのは，もともとDBに掲載されないデータに関することである．たとえば，DBに載っているデータの質について，である．基本，ネット上の情報と同じ，‘As is（今あるそのまま）’で提供されているものである．すなわち，**質を担保する一般的な基準はない**．ただ，塩基配列解読の質はPhredスコアとして数値化され，それを基に配列データの客観的な質は測定することが可能で，その計算結果が次世代シークエンサーから得られたデータの公共DB検索ツールであるDBCLS SRAから辿って見られるようになっている[17]．この例はむしろ例外で，それ以外のデータに関し

ては，自明の理であるが，もともとDBに載っていない情報はどうやっても得られない．ユーザーが使いながら体得していくしかない．さらに知りたいことがあるときには，DB中に記載されている電子メールアドレスなどの連絡先を活用して個別にコンタクトするのがよかろう．

最後に：オープンアクセスの重要性

現在，論文誌を運営する出版社の合併による寡占化，そして論文誌購読料が高騰している．それにともない，古くからある論文誌の購読打ち切りが頻発しており，結果としてオープンアクセスでない論文は目に触れる機会が減ってきている．PubMed検索でその論文の存在を知っていても，所属している大学や研究所で購読していないため全文が読めない，という状況が頻発している．**自分の出した論文が，より多くの読者の目に触れるためにはオープンアクセス**であることが必須条件のようになってきているのである．

データは論文に付随するものであり，そのデータを引用する場合にはその論文を引用すればよいのであり，論文として出版すれば十分である，という考え方がこれまでの生命科学の世界では大勢を占めてきた．しかし残念ながら，自らのデータを普及させ，その再利用を促進するためには，それだけでは不十分で，より「ネット検索」されるようにする必要がある．こういった考え方は，時代とともに変わっていくもので，価値基準もアップデートしていく必要がある．それでは，自ら出したデータを利用してもらうためにはどうしたらいいか？各種の**公的なDB（データレポジトリ）への積極的な登録によるオープンデータ化の推進**であろう．Scientific Data[18]やGigaScience[19]といったデータジャーナルなる分野の雑誌も創刊され，研究データの発表，発見，再利用，そして議論の場として今後注目されていくに違いない．

こういった「知のめぐりを良くする」オープンデータの試みは，いろいろ論議がある．そういったジャーナルでのデータ公開までできなくても，こういう研究がある，と研究室のウェブサイトやブログ，TwitterやFacebookなど，ありとあらゆるさまざまな媒体で情報発信していってほしい．それは，**情報発信するところに情報が集まる**からである．本稿をきっかけにインターネット上のさまざまなツールをうまく活用して研究に役立てていただければと思う．最後にガリレオ・ガリレイの言葉でこの稿の締めの言葉としたい．

書きとどめよ　議論したことは風の中に吹き飛ばしてはいけない

今回紹介したコンテンツについて，今後の連載の中で個別に説明をしていく．

謝　辞

紹介したサービスを維持管理しているNBDC，DBCLSをはじめとした統合DBプロジェクト関係者，データをオープンアクセスにしていただいているすべての研究者に感謝します．

文　献

1) Google scholar: https://scholar.google.co.jp/ (2016/05/25)
2) PubMed: http://pubmed.gov/ (2016/05/25)
3) 生命科学データベース横断検索：http://biosciencedbc.jp/dbsearch/ (2016/05/25)
4) 生命科学データベース横断検索Database List：http://biosciencedbc.jp/dbsearch/dblist.php?lang=ja (2016/05/25)
5) ライフサイエンス辞書:http://lsd-project.jp/ (2016/05/25)
6) Integbioデータベースカタログ：http://integbio.jp/dbcatalog/ (2016/05/25)
7) 統合TV:生命科学系DB・ツール使い倒し系チャンネル:http://togotv.dbcls.jp/ (2016/05/25)
8) 生命科学系データベースアーカイブ：http://dbarchive.biosciencedbc.jp/ (2016/05/25)
9) 生命科学系データベースアーカイブのDOI: http://dbarchive.biosciencedbc.jp/contents/doi/list.html (2016/05/25)
10) figshare: http://figshare.com/ (2016/05/25)
11) Wikimedia commons: http://commons.wikimedia.org/ (2016/05/25)
12) Bono, Hidemasa (2015)：公共データベースを活用した生命科学研究（Next generation research in life science making full use of publicly available databases）. figshare.: https://dx.doi.org/10.6084/m9.figshare.1599741.v3 (2016/05/25)
13) バイオサイエンスデータベースセンター（NBDC）が企画するライフサイエンス分野データベース関連のイベント情報:http://events.biosciencedbc.jp/ (2016/05/25)
14) 統合データベース講習会：AJACS薩摩http://events.biosciencedbc.jp/training/ajacs58 (2016/05/25)
15) 誰もが“バイオインフォマティシャン”の時代：*Nature Digest*, **12**, 22 (2015).
16) なぜバイオインフォマティクスの解析はタダではないのか：http://trattoriainutano.tumblr.com/post/132214903857 (2016/05/25)
17) DBCLS SRA: http://sra.dbcls.jp/ (2016/05/25)
18) Scientific Data: http://www.nature.com/sdata/ (2016/05/25)
19) GigaScience: http://gigascience.biomedcentral.com (2016/05/25)

（【第2回】は94巻11号に掲載予定です）