DP レポート

生命情報・DDBJ センターの データベースと生命科学ビッグデータ

秦千比呂（国立遺伝学研究所）　児玉悠一（国立遺伝学研究所）

概要　生命情報・DDBJ センターは国際塩基配列データベースの一員として，多種多様な生命科学データのための公共データベース群を運用している．生命科学データは多様かつ複雑であり，研究手法の発展により新しいデータの種類が生み出され続けるという特徴がある．さらに次世代シークエンサの登場によりデータサイズはペタバイトスケールになった．本稿では DDBJ センターのデータベースにおける生命科学ビッグデータに対する取り組みを紹介する．

1. はじめに

静岡県三島市にある情報・システム研究機構国立遺伝学研究所（遺伝研）に設置されている生命情報・DDBJ センター（DDBJ センター）は 1987 年に正式に活動を開始し，米国 National Center for Biotechnology Information（NCBI）と欧州 European Bioinformatics Institute（EBI）と共同であらゆる生物種の塩基配列情報を 30 年以上にわたって収集している[1]．この国際協力体制は International Nucleotide Sequence Database Collaboration（INSDC）と呼ばれており，オープンサイエンスを支えるデータベース基盤のあり方として規範とされることが多い．活動を開始した頃はデータベースの担当者が論文に記載された塩基配列を手作業でデータベースに入力していたが，1990 年代に科学雑誌が論文掲載の条件として塩基配列を INSDC に登録することを義務付けるようになり，データが INSDC に集積する体制が整った．論文著者は INSDC を構成する 3 拠点のいずれかのデータベースに塩基配列を登録し，配列に対して発行されるアクセッション番号を引用して論文を投稿する．アクセッション番号は拠点ごとに割り当てられたプレフィックスで区別されているため，拠点間でユニークになるようになっている．論文公開とともに塩基配列データは公開され，INSDC 間で共有される．そのため，論文中の番号で INSDC のいずれかのデータベースを検索することで対象データにアクセスし，誰でも制限なく利用することができる［2]．この国際的なデータ共有体制は相互バックアップと負荷分散にもなっており，塩基配列という生物の基本情報を安定して供給するインフラとして生命科学の発展を支えてきた．

2005 年頃からスループットが飛躍的に向上した次世代シークエンサが登場し，かつてない規模で塩基配列データが生み出されるようになった．次世代シークエンサは塩基配列決定の基本原理は第一世代のキャピラリ・シークエンサと同様であるが並列数が第一世代では数百であったものが次世代では数百万～数十億と桁違いに向上している．2003 年に完了した国際ヒトゲノムプロジェクトは第一世代シークエンサを大量投入してヒトの全ゲノムを 3,000 億円の費用と 13 年間の時間をかけて解読した[3]．しかし，最新の次世代シークエンサを使えばわずか 10 万円で数日も

あれば一人の全ゲノムを解読することができる．このシークエンス能力の劇的な向上は生命科学の分野で革命を引き起こし，1つの研究で得られるデータ量が爆発的に増大するビッグデータの時代に突入した．最近では，従来用いられていた蛍光色素を使用せず，塩基の電位を直接測定してシークエンスする新しい原理の機種も登場している．また，読み取れる塩基配列長が数百であった従来のシークエンサに対して，連続して読み取れる塩基配列長が数万にまで達するロングリード・シークエンサと呼ばれる機種も出てきており，生命科学分野に新たな手法と知見をもたらしている．

INSDC はこの爆発的なデータ量増大に対応するため，2008 年に次世代シークエンサから出力された生データを対象とする Sequence Read Archive（SRA）の運用を開始した[4]．2020 年 7 月末時点で SRA から公開されている塩基数は 1.6 京のオーダーに達している．また，一連のデータが複数のデータベースに分けて登録されるようになったため，データベースを横断して情報を整理する必要が生じた．これに対し INSDC はデータを「研究」という単位で取りまとめる BioProject，及び，「サンプル」という切り口で整理する BioSample データベースを立て続けに構築し，運用を開始した[5]．DDBJ センターも INSDC の一員としてこれらのデータベースを立ち上げるとともに，塩基配列を中心として生物学情報を網羅的に収集する体制を整えるべく，個人レベルのゲノムデータを扱うアクセス制限データベース Japanese Genotype Phenotype Archive（JGA，2013 年）[6]，遺伝子発現等の機能ゲノミクスデータのための Genomic Expression Archive（GEA，2018 年）[7]，メタボロミクスデータのための MetaboBank（2020 年 10 月）[8]，及び，ヒトゲノムのバリエーションデータのための Japanese Variation Archive（JVar，2020 年度開始予定） と急ピッチでデータベースの整備を進めている（図 1, 2）．また，DDBJ センターは大学共同利用機関法人である情報・システム研究機構傘下のセンターとして遺伝研スーパーコンピュータ（スパコン）を所有している．遺伝研スパコンは，国内研究者のための計算機資源としての用途だけでなく，上記の公共データベース群の運用基盤としても使われている．

本稿ではオープンサイエンスを支えるビッグデータ基盤である SRA，及び，1987 年から続けている機能注釈された塩基配列データベースである DDBJ/ENA/GenBank（以降，塩基配列データベース）を中心に紹介する．

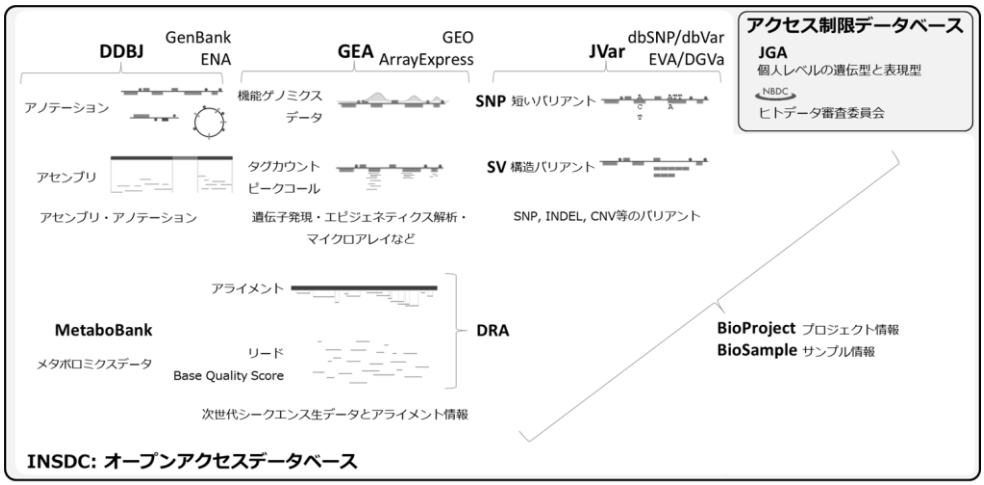

図 1　DDBJ センターが提供する公共データベース群

	Annotated sequences	NGS reads	Study	Sample	Assembly	Functional genomics	Variation	Genotype and phenotype	Metabolomics
NCBI	GenBank	Sequence Read Archive	BioProject	BioSample	Assembly	GEO	dbSNP/dbVar	dbGaP	
EBI	European Nucleotide Archive (ENA)					ArrayExpress	EVA/DGVa	EGA	MetaboLights
DDBJ	DDBJ	Sequence Read Archive	BioProject	BioSample	Assembly	GEA	JVar-SNP/SV	JGA	MetaboBank

INSDC (国際塩基配列データベース共同事業)

図 2　INSDC で対応するデータベース

2. 2.　DDBJ Sequence Read Archive (DRA)

世界初の次世代シークエンサであるロシュ社の「454」が 2005 年に米国で発売され，大規模な塩基配列データが NCBI に到着するようになった．当初 NCBI は第一世代のキャピラリ・シークエンサから出力される生データのためのデータベースである Trace Archive に 454 のデータを格納していた．しかし，Trace Archive は 1 配列を 1 レコードとして扱う構造になっており，一度のランで数百万配列が出力される 454 データを格納する上での限界がただちに認識され，NCBI で新しい発想に基づいたデータベースの設計と開発が進められた．こうして開発されたのが次世代シークエンサから出力される生データのための Sequence Read Archive（SRA）であり，NCBI によって 2007 年に運用が開始された．翌年には正式な INSDC 事業となり，日米欧で分担して巨大な次世代シークエンスデータを収集することとなった[9]．以下では次世代シークエンスデータというビッグデータを保存・提供・解析するシステムについて米国の NCBI を中心として紹介したい．

SRA は，データがどのようにして得られたのかを説明する「メタデータ」と「配列情報」の二つの部分から構成されている．「メタデータ」は BioProject，BioSample，Experiment，Run という相互に関連して階層構造を形成したオブジェクトに分かれており，図 3 に SRA に対応する DDBJ Sequence Read Archive（DRA）のデータモデルを示した．

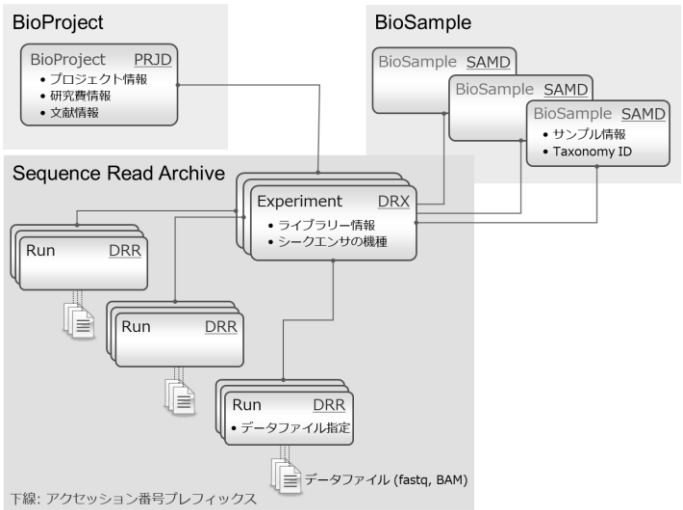

図 3　DDBJ Sequence Read Archive のデータ構造

次に「配列情報」は塩基配列（リード），塩基配列の決定精度を表す Base Quality Score（BQS），及び，各リードに付けられた名前であるリード名から構成され，これらはバイナリーの SRA ファイルとして Run に紐づけて格納される．アクセッション番号は各メタデータオブジェクトに対して発行され，個々のリードには Run 番号に１から始まる整数を付した連番が機械的に割り振られる．塩基配列データベースでは一塩基配列を１レコードとして扱っているが，SRA では次世代シークエンサの１ランで出力される数百万～数億の塩基配列を Run 単位でまとめて扱うようになっている．また，塩基配列データベースでは塩基配列と付随するメタデータが一緒になったフラットファイルを各配列毎に作成してユーザに提供している．そのため，例えばある研究プロジェクトに由来する１万の塩基配列に論文情報を追加する場合，１万ファイルの書き換え処理が発生する．しかし，SRA では配列情報とメタデータは分離されており，かつ，メタデータは階層構造を形成しているため，この例の場合であれば BioProject を更新するだけで済む．また，塩基配列を修正する場合，塩基配列データベースでは１万ファイルの内１ファイルだけ書き換えるといった更新処理が必要になり管理コストが高い．一方，SRA では塩基配列の修正は Run の差し替えで対応するため，SRA ファイル自体は更新不可のファイルとして扱うことができ，管理コストは低くなっている．

大量データの格納形式としてテキストファイルを採用すると扱いやすくて容量が小さいがインデックスが利用できないという欠点がある．一方，データベースを使うとインデックスは利用できるが容量が大きくなるという欠点がある．SRA ファイルは塩基配列，BQS や各種インデックスを別々に保持したバイナリーファイルとなっており，インデックスによる高速アクセスを提供しながらもコンパクトなサイズになっている．この構造により SRA ファイルから特定の連番レンジ内のリードを抽出したり，塩基配列だけを取り出したりすることが高速にできる．ファイル構造の定義情報は SRA ファイル自体に埋め込まれているため，前方互換性を保ちながら次世代シークエンサの新しい出力形式に対応することができる．SRA では次世代シークエンサが出力する様々な形式のオリジナルファイルを SRA ファイルに変換して半分強のサイズでアーカイブしている．

次世代シークエンサの性能向上とコスト低下により SRA に登録されるデータ量は爆発的に増えており，ストレージコストを減らすべく，SRA ファイルサイズの圧縮が試みられている．圧縮方法は（１）塩基配列をリファレンス配列との差分のみ保存する，（２）BQS を圧縮あるいは除去する，の二つである．まず（１）ではリファレンス配列にアライメントされたリードとリファレンス配列との差分のみを保存する．この方法はリファレンス配列を使って元の配列情報を復元することができる可逆圧縮であり，サイズを半分程度にすることができる．次にファイルサイズの６～７割を占める BQS を圧縮するツール（２）であり，40 段階に渡るフルスケールの BQS を 8 段階，2 段階に圧縮，あるいは完全に除去する程圧縮率が高くなり，最大でサイズを 3/10 程度にまで圧縮することができる．

NCBI/EBI/DDBJ センターの各 SRA 拠点は公開された SRA ファイルを相互にミラーリングしており，2020 年 7 月時点で DDBJ センターは 8 ペタバイトの SRA ファイルを提供している．SRA は活発に利用されており，再現性の確認，バイオインフォマティクスツールのテストや網羅的な解析により使いやすいかたちで情報を整理したサイトの構築など多面的に利用されている．しかし，現在のデータ量の増加ペースはデータベース運用の観点からすると持続可能なレベルを超えて

しまっている．データ量の増加による一番の問題はストレージコストの増大であり，各 SRA 拠点は安価なテープ装置をディスクと組み合わせて使っている．しかし，テープ装置は応答が遅いため，大型の研究プロジェクトに由来するデータは数十テラバイトに及ぶことも珍しくなく，データのダウンロードだけで数日間を要するようになっている．NCBI の親組織である National Institutes of Health（NIH）はこれらの問題を解決してデータサイエンスによる生物医学研究の発展を促すため，パブリッククラウド上にビッグデータを保管・提供・解析する持続可能なエコシステムを構築する STRIDES イニシアティブを推進している[10]．このイニシアティブではクラウドプロバイダーは割引料金を提供することになっており，現在のところアマゾンウェブサービス（AWS）とグーグル（GCP）の 2 社が参加している．このイニシアティブの一環として 2020 年初頭には 10 ペタバイト以上の全 NCBI SRA データが米国にある AWS と GCP のデータセンターへコピーされた [11]．なお NCBI SRA には個人の権利保護のためにアクセス制限が課された個人由来のデータと制限の無いオープンアクセスのデータがおよそ半分ずつ含まれており，INSDC の対象はオープンアクセス分のみである．ユーザはデータが保持されている AWS/GCP のデータセンターにログインして直接データにアクセスすればダウンロードは不要になる．しかし，自身のアカウントで計算処理すれば計算料金がかかってしまい，データをコピーすれば保存料金，さらにデータセンター外にデータをダウンロードすればダウンロード料金が発生する仕組みとなっている．現在のところ NCBI はオンプレミスサーバからも SRA データを提供しているため，ユーザはそちらからダウンロードすれば今まで通りにデータを利用することができる．しかし，NIH はストレージコストを持続可能なレベルに抑えるために次のようなさらに踏み込んだ方針を打ち出している[12].

- NCBI のオンプレミスサーバでは BQS を含む SRA ファイル（SRA+BQS）は一定期間しか提供せず，一定期間経過後は BQS が削除された SRA ファイル（SRA-BQS）しか提供しない．
- AWS/GCP では SRA+BQS を提供するがアクセス頻度が少ないものはディスク（Hot storage）からテープ等のアクセスは遅いが安価なストレージ（Cold storage）に移される．ユーザはデータの Cold から Hot への移行をリクエストできるが移行料金は NCBI 持ちのため月当たりの移行量には上限が設けられる．
- オリジナルファイルは Cold storage のみで提供される．

この新しいモデルだとユーザは BQS を必要とする解析を行いたい場合，NCBI オンプレミスサーバに SRA+BQS が無い場合は米国の AWS/GCP にアクセスする必要があり，無料ではなくなる．今のところ DDBJ センターと EBI は SRA+BQS をそれぞれのオンプレミスサーバで提供し続ける方針であるため，こちらでは無料でアクセスすることができる．NIH は 2020 年 8 月頃からこの新しいモデルに移行することを表明している．

NCBI はデータだけではなくシステムもクラウドへの移行を進めており，GCP の BigQuery を使った SRA メタデータの検索システムなどクラウドネイティブなサービスを展開している．

米国の NCBI がクラウドに舵を切っているのに対し，EBI はオンプレミス志向である．EBI は合計 307 ペタバイトのオブジェクトストレージを運用しており，10 年以上の時間をかけてオープンソースのソフトウェアを活用した情報基盤を構築している[13]．しかし，EBI 単独でソフトウェアを開発・運用することは厳しい

ため，Elixir や European Open Science Cloud といったイニシアティブに参加して，EU 各国と協力してバイオインフォマティクスのソフトウェアやインフラを共同で開発するようになってきている[9].

DDBJ センターもオンプレミスで遺伝研スパコンを運用しており，15 ペタバイトのディスクと 15 ペタバイトのテープで構成される大容量の階層ストレージシステムで SRA データをアーカイブしている．遺伝研スパコンのユーザは直接 SRA データにアクセスすることができ，予めセットアップされているバイオインフォマティクスツールを使って解析することができる．遺伝研スパコンは SINET5 で AWS と連携しており，用途にあわせてスパコンと AWS を使い分けられるようにしておりハイブリッドを志向している．

以上，SRA を取り巻く状況について米国を中心に概観してきたが「全てのデータをオンプレミスサーバで提供する」という時代は終わりを迎えつつあり，パブリッククラウドを含めた新たなモデルの模索が始まっている．

3. 塩基配列データベース

フレデリック・サンガーが 1970 年代後半に塩基配列決定法を確立すると，世界各地で様々な生物種の塩基配列が報告されるようになった．やがて塩基配列を集中的に管理する公共データベースの設立を求める声が研究者コミュニティで高まり，1980 年代に塩基配列に機能注釈（アノテーション）が付加された情報を収集する塩基配列データベースとして欧州の EMBL-Bank（後に European Nucleotide Archive，ENA に改称）が 1980 年，米国の GenBank が 1982 年，日本の DDBJ が 1987 年に設立され，これらが後に INSDC に発展した．

塩基配列データベースである DDBJ/ENA/GenBank では塩基配列とアノテーションを一緒にしたフラットファイル（図 4）としてデータを提供しており，2020 年 7 月時点で 24 億塩基配列，9.3 兆塩基を公開している（図 5）．塩基配列データベースのデータサイズは合計 1 テラバイト程度であり，SRA と比べるとサイズは小さいが件数が億単位と多いこと，アノテーションの記載ルールが複雑であることが特徴として挙げられる．DDBJ センターではデータ登録を円滑にすべくデータ入力部分の自動化を進めており，この章ではその取り組みについて紹介したい．

前章で述べたようにデータベースを横断してサンプル単位でデータを整理する BioSample が稼働し，サンプル情報を集約管理する基盤ができた．BioSample ではサンプル情報を属性と属性値のペア（例 tissue=leaf）で柔軟に記述することができ，サンプルの種類毎に必須と任意属性のセットをパッケージとして提供することでサンプル記述の標準化を促している．DDBJ センターでは BioSample のサンプル情報をチェックするバリデータを 2018 年に登録システムに組み込み，チェック結果の登録者への提示や属性値の自動修正等によりデータ登録フローを大幅に効率化した．BioSample バリデータではチェックに使うデータや定義情報の形式としてセマンティック・ウェブテクノロジーを積極的に活用している．INSDC では塩基配列が由来する「生物」の情報は極めて重要であるため，NCBI が構築している Taxonomy データベース[14]を共通で使っている．Taxonomy データベースでは生物の分類群に taxonomy ID を割り振って管理しており，系統分類を反映したツリー構造で管理されている．例えば，ヒト（学名 Homo sapiens，id:9606）の上位分類群はホモ属（Homo，id:9605）であり，さらに上位階層に進むと哺乳類（Mammalia，id:40674）となっている．各分類群は主要な学名と ID 以外にも異名

（ニホンアマガエルの学名 Dryophytes japonicus に対する Hyla japonica など）や種を代表するバクテリアの基準株といった様々な付随情報を持っている．分類学では分類の見直しの結果，学名の変更や，別種とされていたものが同種になるといったことが絶えず起きており，NCBI Taxonomy では専門家が新しい知見を取り込んで日々データベースを更新している．DDBJ センターでは情報が日々更新され，度々データモデルが変更になる NCBI Taxonomy をセマンティック・ウェブテクノロジーの標準メタデータ記述言語である RDF（Resource Description Framework）で管理している．RDF は表形式に比べて情報を柔軟に表現することができ，データモデル変更がシステムに及ぼす影響を小さくすることができる．BioSample バリデータは Taxonomy RDF を使って，生物名と taxonomy ID の一致，異名の学名への変換や上位分類階層（バクテリアゲノム用パッケージ選択時に記載されている生物がバクテリアかどうか）のチェックを実施している．また，BioSample のパッケージ定義情報は RDF の語彙拡張である OWL（Web Ontology Language）で定義する予定である．生物学は生き物が多様である上に，同じ生物種でも塩基配列や機能は異なっていることが多く，実験手法も様々であるためデータが極めて多様性に富んでおり，生物やデータの種類といった「文脈」に依存したルールや例外が多い．OWL はこのようなルールを記述するのに適しているため，DDBJ センターでは定義情報の OWL への集約を進め，システムで共通利用するとともに外部へも積極的に提供していく．

登録者の多くが DDBJ に塩基配列データを登録する際に難しく感じるのは，遺伝子等のアノテーション情報を DDBJ のルールに則って記載し，登録用ファイルを作成することである．バクテリアは遺伝子構造がシンプルであることからプログラムによる機能予測精度が高い．さらに，データ登録量もヒトなどの真核生物と比較して圧倒的に多い．そのため，DDBJ センターでは主にバクテリアを対象とした自動アノテーションサービスである DFAST を提供しており，バクテリアゲノム配列をアップロードすると自動でアノテーションが付与され，結果は登録用ファイルとして出力されるので，そのまま DDBJ に登録することができる[15]. DFAST を使うとゲノム登録にかかる時間が大幅に短縮されるため，バクテリアゲノム登録における DFAST 利用率は 9 割以上に達している．

図 4 DDBJ フラットファイルの例

最後にビッグデータの活用事例としてバクテリアゲノムを使った生物種同定サービスを紹介したい．以前からバクテリアゲノムの塩基配列を既知のゲノム配列と比較して生物種を同定する ANI（Average Nucleotide Identity）解析[16]が行われてきたが，既知ゲノムが少なかった時代には同定率が低かった．しかし，次世代シークエンスの時代になり INSDC にバクテリアゲノムが急速に集積してほぼ全てのバクテリアの基準株が網羅されるようになった．NCBI GenBank ではカバーしている全ての基準株について同種と判断する ANI の閾値のリストを公開しており，大腸菌は 97.7%，ピロリ菌は 94.7%といった株ごとの閾値が確認できる[17]. GenBank が 2018 年に公開されている全てのバクテリアゲノム 14 万件に対して ANI 解析を実施したところ，66.8%は生物名が正しく同定されているが，3.6%は間違っていることが示唆された（残りは対応する基準株のゲノムデータが無く評価できなかった）．GenBank では新規登録されるバクテリアゲノムに対して ANI チェックを自動で実施しており，推定される生物種と登録者が記載したそれとが異なっている場合には警告を提示し，間違った生物名がデータベースに入らないようにしている．この ANI による生物名チェックは単にゲノムデータが INSDC に集積することで可能になったわけではなく，NCBI が RefSeq というデータベースで GenBank に登録された一次データから生物種を代表する塩基配列とアノテーションを地道に構築してきたこと，及び，NCBI Taxonomy がバクテリア分類群に対して基準株の情報を付加してきたことの上に成り立っている．この GenBank の ANI リストによる生物名チェックは DFAST にも最近導入され，INSDC では集積したゲノムデータによって登録されるゲノムデータの品質が高まる，という好循環に入

っている．この事例はビッグデータを有効活用するためには単にデータが大量にあるだけでは不十分であり，それらが活用できるように加工・整理された高品質で網羅的なリファレンスが必要であることを示している．

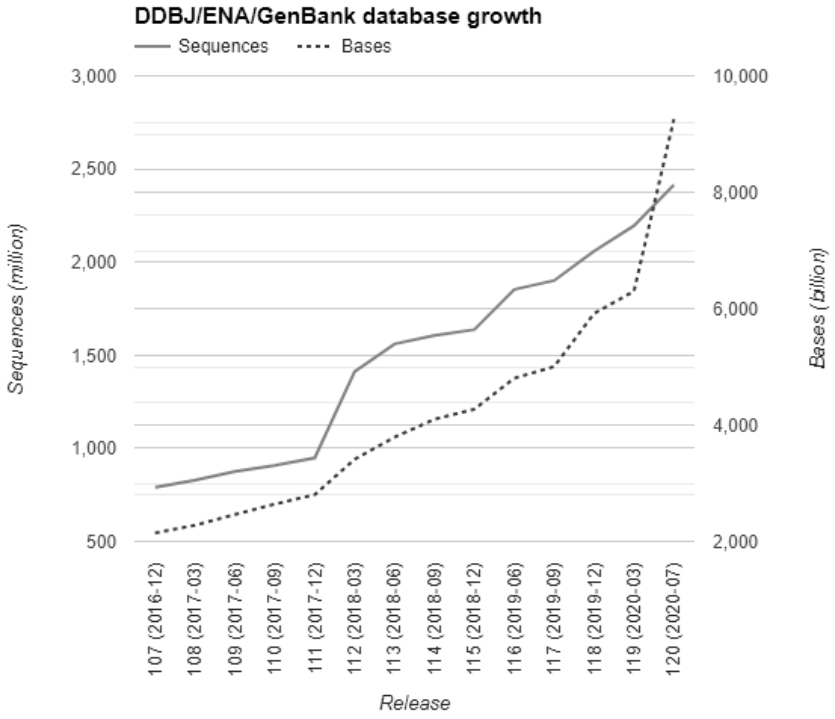

図 5　塩基配列データベースのデータ量 (配列数/塩基数)

4. 今後の展望

生命科学は数式で表現できる分野とは異なり，生物や環境に依存した多様なデータが存在し「記載」が中心の学問である．そのため生物分類であれば Taxonomy，サンプル情報は BioSample，アノテーション付き塩基配列は DDBJ/ENA/GenBank，次世代シークエンサからの生データは SRA といったように情報の種類に応じた複数のデータベースでデータを管理し，ID で相互に連携させることが必要になる．さらに「アノテーション付き塩基配列」といっても研究手法の進化によって様々な種類のデータが生み出されるため，新しいデータ種別や記載ルールを適宜追加していく必要がある．これらが生命科学データベースの難しいところであり，多種多様なデータを複数のデータベースとデータモデルで可能な限り正確に捉え，運用しながら日々更新し続ける必要がある．また，データの有効活用のためには研究者から登録された一次データを集積しているだけでは不十分であり，それらを整理加工した二次データベースの構築も不可欠である．次世代シークエンサの登場により，データの「多様性」に加え，ペタバイトスケールという「サイズ」の問題も加わっている．

DDBJ センターではデータベースのラインアップを増やし，また，INSDC メンバー間で話し合い，研究の発展に追随すべく新しいデータに対応したデータ種別を追加することで「多様性」に対応してきた（表１）．今後はセマンティック・ウェブテクノロジーを活用してデータモデルやルールの多様化に取り組んでいく．また，「サイズ」に対してはデータベースと遺伝研スパコンを一体として提供し，ビッグデータを直接解析できる環境を整備することで対応していく．DDBJ センタ

ーは今後もデータサイエンスを推進して生命科学の発展により一層貢献していく所存である.

表 1 DDBJ センターにおける新規データ種別とデータベースの年表

年	出来事*	対応するデータ
1993 年	EST 新設	転写産物配列
2002 年	WGS 新設	ゲノム配列
2005 年	ENV 新設	環境サンプル配列
2008 年	TSA 新設	転写産物アセンブリ配列
2008 年	SRA 稼働	次世代シークエンサからの生データ
2011 年	BioProject 稼働	研究プロジェクト
2013 年	JGA 稼働	アクセス制限が必要な個人ゲノムデータ
2014 年	BioSample 稼働	サンプル情報
2018 年	GEA 稼働	機能ゲノミクスデータ
2016 年	TLS 新設	特定遺伝子領域の配列
2020 年	MetaboBank 稼働	メタボロミクスデータ
2021 年	JVar 稼働予定	バリエーションデータ

*「新設」は塩基配列データベースへのデータ種別の追加,「稼働」はデータベースの運用開始を表す.

謝辞 本稿執筆にあたりご協力頂いた DDBJ センターの皆様に，謹んで感謝の意を表する.

参考文献

1) Kodama, Y et al.: DNA Data Bank of Japan: 30th anniversary, Nucleic Acids Res. 4;46(D1):D30-D35.. (2018)
2) “International Nucleotide Sequence Database Collaboration Policy” http://www.insdc.org/policy.html (参照 2020-07-30)
3) International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome, Nature. 431(7011):931-45 (2004)
4) Wheeler, D. L. et al.: Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 36(Database issue): D13–D21. (2008)
5) Barrett, T. et al.: BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata, Nucleic Acids Res. 40(Database issue):D57-63. (2012)
6) Kodama, Y et al.: The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data, Nucleic Acids Res. 43(Database issue):D18-22. (2015)
7) Kodama, Y et al.: DDBJ update: the Genomic Expression Archive (GEA) for functional genomics data, Nucleic Acids Res. 47:D69-D73. (2019)
8) 櫻井 望: 未知化合物を同定するためのメタボロームデータベースの開発と活用, 日本化学会情報化学部会誌, 2019, Vol.37, No.3, p. 68-71.
9) Saunders, G et al.: Leveraging European infrastructures to access 1 million human genomes by 2022, Nat Rev Genet. 20(11):693-701. (2019)
10) “STRIDE initiative”. https://datascience.nih.gov/strides
11) Insights : The entire corpus of the Sequence Read Archive (SRA) now live on two cloud platforms!NCBI Insights”. https://ncbiinsights.ncbi.nlm.nih.gov/2020/02/24/sra-cloud/

12) “Request for Information: Use of Cloud Resources and New File Formats for Sequence Read Archive Data”. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-108.html
13) “In Focus: Big data infrastructure”. https://www.youtube.com/watch?v=sTAhG9b_S4Y
14) Federhen, S.: The NCBI Taxonomy database, Nucleic Acids Res. 40(Database issue):D136-43. (2011)
15) Tanizawa, Y et al.: DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication, Bioinformatics. 34(6):1037-1039 (2018)
16) Goris, J. et al.: DNA–DNA hybridization values and their relationship to whole-genome sequence similarities, Int J Syst Evol Microbiol. 57(Pt 1):81-91. (2007)
17) Ciufo, S et al.: Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI, Int J Syst Evol Microbiol. 68(7): 2386–2392. (2018)

秦 千比呂（正会員）chata@nig.ac.jp
2020 年 3 月総合研究大学院大学生命科学研究科遺伝学専攻にて博士（理学）を取得。2019 年 4 月より国立遺伝学研究所 生命情報・DDBJ センターに勤務。

児玉 悠一（非会員）ykodama@nig.ac.jp
2007 年 3 月奈良先端科学技術大学院大学バイオサイエンス研究科にて博士（バイオサイエンス）を取得。2008 年 1 月より国立遺伝学研究所 生命情報・DDBJ センターに勤務。

投稿受付：2020 年 10 月 23 日
採録決定：2020 年 10 月 30 日
編集担当：藤原一毅（国立情報学研究所）