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

菅原秀明
大学共同利用機関法人情報・システム研究機構
国立遺伝学研究所
生命情報・DDBJ 研究センター
hsugawar@genes.nig.ac.jp

国際塩基配列データベース共同事業（International Nucleotide Sequence Databases Collaboration, INSDC）におけるデータ管理，データ項目，データの値ならびにデータ交換の実質標準化を紹介する．データ交換については，XML 技術，SOAP 技術および Web サービス技術の導入を紹介する．さらに，国際塩基配列データベースが，データ量が伸び続け，データ構造も変化し続けるダイナミックなデータベースであり，したがって，バイオの世界だけでなく，情報の世界にも挑戦的課題であることを指摘する．

生命現象の多様性と共通性

生命の誕生以来地球上には，微生物，植物そして動物と推定 1 千万種を超える多様な生物が生まれてきた．この多様な生物は分子，細胞，組織，個体，集団といった階層を内包している．塩基配列データベースは，多様で多層な生物を対象とする研究のバックボーンとなるデータベースである．なぜならば，世代から世代へ受け継がれていく遺伝情報によって生物の振る舞いの枠組みが決まっており，この遺伝情報は 4 種類の塩基（アデニン（A）・チミン（T）・グアニン（G）・シトシン（C））を要素とする文字列で記述されているからである．したがって，生物のあらゆる振る舞いを直接間接に，遺伝情報の総体であるゲノム配列上へマッピング可能と考えられるからである．

DDBJ（DNA Data Bank of Japan, http://www.ddbj.nig.ac.jp/）は[1]，欧州の The European Molecular Biology Laboratory（EMBL）ならびに米国の National Center for Biotechnology Information（NCBI）と協力しながら（**図 -1**），塩基配列データとその生物学的意味を網羅した国際塩基配列データベース（International Nucleotide Sequence Databases, INSD）を構築して，研究社会から一般社会まで広く提供している．この国際協力を国際塩基配列データベース共同事業（INSD Collaboration, INSDC（http://www.insdc.org/））という．

本稿では，DDBJ を例に INSDC におけるデータの登録，査定，管理，交換および提供の概要を紹介する．

データエントリー

DDBJ には，日々研究者や技術者から，また，毎月特許庁から，データが送付されてくる（図 -1 の左端の右向き矢印）．また，EMBL ならびに NCBI とそれぞれが蓄

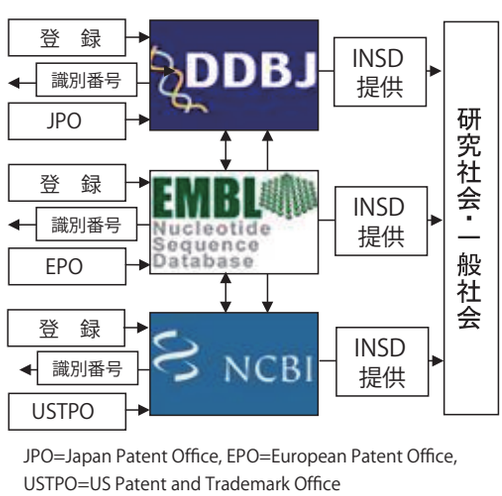

図 -1　国際塩基配列データベース共同事業の枠組み

積したデータを日々交換している．INSDC の 3 極はそれぞれが受領したデータ 1 件をエントリーと呼び，各エントリーに固有の識別番号を付与して，データを査定した後，登録者に通知する（図 -1 の左端の左向き矢印）．DDBJ ではデータ送付の方法として，対話形式の Web ブラウザ経由（http://sakura.ddbj.nig.ac.jp/Welcome-j.html）と，あらかじめ書式を定めた表形式ファイルの転送（http://www.ddbj.nig.ac.jp/sub/masssub-j.html）を用意している．

DDBJ は受領したデータを，INSDC が統一編集している規程集 Features Definition Table（FTD）に基づいて査定し，問題がある場合は登録者に連絡して解決を図る．査定は，バイオの専門知識を有するアノテーターが，DDBJ が独自に開発したワークベンチを使ってプログラム処理の出力結果を参照しながら行う．

FTD は，毎年 1 回 3 極の実務者が一堂に会する 3 日間の国際実務者会議での議論と電子メールによる調整によって 1990 年代から継続的に拡充されている．最新の版は 2005 年 10 月付けの Version 6.4 である（http://www.ddbj.nig.ac.jp/FT/full_index.html）．このバージョンの変遷は，生命現象の理解が深まるとともに新しい概念ひいては新しいタイプのデータが生まれてくることに対応している．FTD の具体についてはデータ標準化の章で紹介する．

INSDC において，データの質の変化に対して量の変化はどうか．1995 年から 2005 年までの膨張を**図 -2** に示す．右側の軸と線グラフがエントリー件数に対応し，左側の軸とヒストグラムが塩基対の数（図中では bp（=base pairs））に対応する．INSD はこの 5 年間（2001 年から 2005 年）平均して毎年 1.5 倍の割合で増大してきた．この伸びにはヒトゲノム配列[2]が大きく貢献したが，高効率の配列決定装置の開発と普及が見えてきていることから，今後もこの程度の伸びを覚悟しておく必要がある．たとえば，2005 年半ばの 4,500 万エントリー／ 500 億塩基対の規模から外挿すると 2006 年半ばには 6,750 万エントリー／ 750 億塩基対に達することになる．また，短期間に 100 万件規模の大量データを受付・査定・公開する機会も増えてきている．

DDBJ はエントリーをリレーショナルデータベースで管理しているが，そのデータ構造やアプリケーションプログラムは，1990 年代初めに設計されたものである．その後，ゲノムプロジェクトや多数のサンプルの一括解析が盛んに行われるようになり，当時は予想もしていなかった大量かつ質が異なるデータを高速に処理する必要が高まってきた．このため，ソフトの改訂とハードの増強が懸案となっている．しかし，日々のデータ処理と並行して新システムを開発し，旧システムから円滑に移行することは，現実には難しい．また，データベースの膨張に応じて必要な時に必要な計算機資源を導入することも，社会経済の観点から困難である．

エントリーの具体例をここで紹介する．ヒトの網膜異常にかかわると思われる遺伝子をマッピングした成果であるエントリーを Flat File（FF）形式で**図 -3** に示した．FF 形式は INSD の構築が始まって以来広く親しまれてきた書式である．FF の書式は，3 極の間で見た目が異なっているが，基本的には行頭にデータの大きな分類（大項目）を示す文字列があり，必要に応じて，小項目が段下げや“/”を目印に表示される書式になっている．ただし，図 -3 の中段にある FEATURES は，塩基配列データの由来や生物学的意味を記述する枠であり，段下げされた位置にある“source”が大項目に，次のカラムの“organism”が小項目に，“Homo sapiens”がその値に相当する．

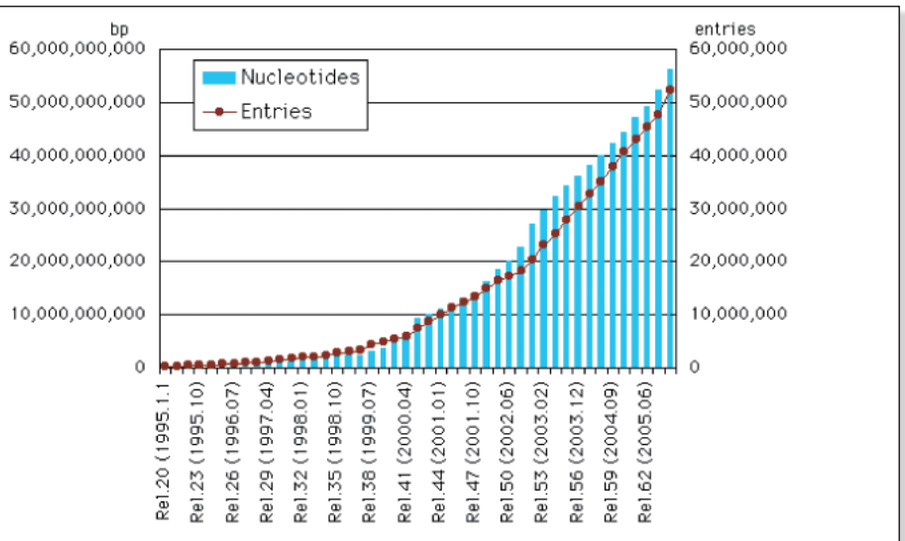

図 -2　国際塩基配列データベースの継続的増大

```
LOCUS      AI444820              241 bp    mRNA    linear   EST 02-AUG-1999
DEFINITION  RET3F3M13 subtract    ed retina cDNA library --------
ACCESSION   AI444820
VERSION     AI444820.1
SOURCE      Homo sapiens
          ------
REFERENCE   1
  AUTHORS   den Hollander,A.I., --------
  TITLE     Isolation    and mapping of novel candidate genes for retinal -----
  JOURNAL   Genomics 58    (3), 240-249 (1999)  ------
FEATURES              Location/Qualifiers
     source          1..241
                     /organism="Homo sapiens"
                       ----------
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="RET3F3"
                      ----------
BASE COUNT         72 a             53 c           69 g         46 t
ORIGIN
        1 gtacagcatg  ggcaaagagc agctttagtg gaatcacat aacaatgtgg gaggaaaaac
       61 aaggcaggaa tcagaaccct gtgggtgaaa acattgcaca ggagcccagg gccagagcac
      121 agcgtgtcag  tgaacagaag ----------
```

図 -3　INSD のエントリー例（Flat File 形式）．図中の ---- は一部省略を示す

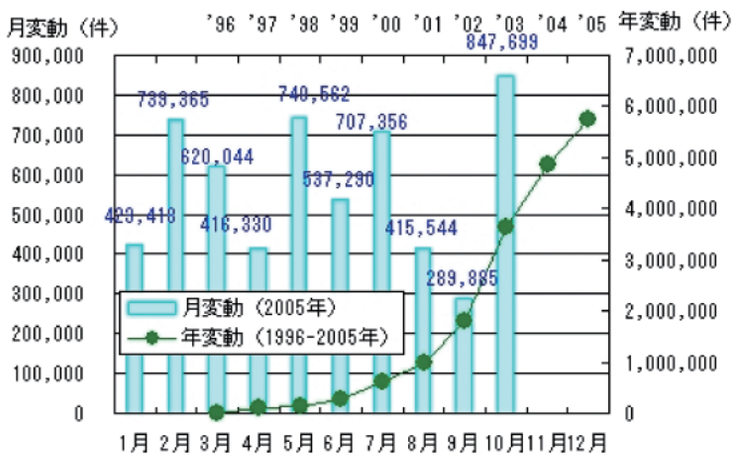

図 -4　アクセション番号による高速検索システム getentry の利用の伸び

データ管理

アクセション番号

「データエントリー」の章で紹介したように，エントリーにはすべて固有の識別番号が振られる．アクセション番号である．図 -3 の ACCESSION の行に書き込まれている“AI444820”が，それである．アクセション番号の重複割り当ては，EMBL が必要に応じて 3 極に割り当てるアルファベット（図 -3 の例では AI）の範囲内で，3 極がそれぞれ受け付け順に番号を振ることによって，回避している．

アクセション番号はバイオ分野において見かけ以上に重要な機能を果たしている．論文にはアクション番号が掲載されるので，論文の読者はアクセション番号を手がかりに実データを INSD から取得する．また，このアクセション番号は遺伝子配列に応じて作り出されるタンパク質の構造のデータベース，生体内で起きている生体分子の相互作用のデータベースといった分子生物学のデータベースから地球規模の生物多様性のデータベース（Global Biodiversity Information Facility, http://www.gbif.org/）まで，生物にかかわるあらゆるデータベースをつなげる鍵ともなっている．したがって，アクセション番号はエントリーの管理に加えて，効率のよい検索と連携にとっても主要なキーとなっている．

事実，アクセション番号は活用されている．アクセション番号の指定によって，4,500 万エントリーを超える規模のデータベースから該当するエントリーを瞬時に返す検索システム getentry の利用動向を図 -4 に示す．図

の左軸とヒストグラムが 2005 年の月変動に，右軸と折れ線グラフが 1996 ～ 2005 年の間の年変動に対応しているが，getentry の利用件数が毎年伸び続けていることを見て取れる（図中の 2005 年の年間利用数には 10 月分までしか計上していない）．この伸びがアクセション番号の重要性を実証している．したがって，アクセション番号の整合的付与は INSDC の最も重要な任務である．

INSDC はエントリーの更新を反映するバージョンも管理している．すなわち，登録データに何らかの変更が加えられた場合にはアクセション番号の後のピリオドの後の数字がバージョンを表す．図 -3 のエントリーでは VERSION 行に“AI444820.1”とあるので，このエントリーには初回登録以後変更が加えられていないことが分かる．エントリーによってはバージョン番号が 2 桁に及ぶものもある．

区分（ディビジョン）

INSD のデータは均一ではなく，実験の観点や手法ひいては質が異なるデータが混在している．たとえば，実験で分離した特定の遺伝子の配列，細胞の中で働いている遺伝子群について網羅的に決定した配列，ゲノムの全配列，生物種を特定できないが環境から抽出した多数の遺伝子断片の配列，などである．INSD 全体を対象にした検索も必要な場合があるが，多様な配列データを区分しておいた方が，データを管理する側にとっても利用する側にとっても利便性が高い．図 -3 のエントリーでは，1 行目 LOCUS 行の日付の前にある“EST”がこの配列が属する区分を示している．EST は expressed sequence tag の略で，機能している遺伝子の部分配列を意味する．

区分は大きく，生物種による区分，配列の品質に基づく区分，その他の区分の 3 種類に分類できる．区分内ではデータが比較的均一になっているので，DDBJ における管理にとっても区分は有用である．たとえば，区分 EST に属するエントリーだけに共通な修正を，全データベースを前処理することなく容易に実行可能である．利用者にとっても区分を活用すれば，ヒトの配列だけ抽出したい，特許由来の配列だけ対象にしたい，EST だけ利用したい，というそれぞれの観点から INSD から効率よく必要な部分集合を入手することができる．

Global Unique Identifier（GUID）

GUID は文字通り「地球上で一意に決まる識別子」である．情報通信の世界ではたとえば MAC アドレスである．バイオの世界では INSDC のアクセション番号が GUID の一種であるが，近年，生物材料に付与する GUID が議論されている．ディジタルオブジェクト識別子（Digital Object Identifier, DOI）などが参考にされているが，INSDC では，生物材料の GUID として生物材料が由来する機関のコード，その機関における管理単位（例：生物分類群）およびその管理単位における管理番号の 3 つ組みを検討中である．GUID を運用することによって，塩基配列が由来する材料（図 -3 では /clone="RET3F3"）の異同を精密に判断することができる．このことは，流通の過程で変異が起きることがあり，また全体や一部を複製可能な生物材料を利用する実験研究の再現性の観点から肝要な点である．

データ標準化

「データエントリー」の章で触れた FTD は，塩基配列データとその管理情報とアノテーション情報の記述形式の規程集である．アノテーション情報の中でも最も興味が持たれていると思われる CDS の規程を**図 -5** によって一部紹介する．FTD では大項目を Feature key，小項目を qualifier と呼んでいるが，CDS は Feature key の 1 つである．FTD では，Feature key ごとに Definition（定義）と有効な qualifiers のリストが与えられている．CDS は図 -5 にあるように「タンパク質（ペプチド）のアミノ酸（終止コドンを含む）をコードする配列」と定義されている．CDS に属する qualifiers の中にはたとえば，CDS 部分の遺伝情報が指示する産物の名称を格納する product が設定されている．FTD で列挙されているすべての qualifiers に対して値を記述する必要はないが，新たな qualifier が必要になった場合は，国際実務者会議での議論と合意の手順を踏んで FTD に記述されて初めて使用可となる．

FTD では Feature key の記述部分に，有効な qualifier の値として許されるデータの型が示されているが，個々の qualifier の規程も別途用意されている．

INSDC の日米欧 3 極の間でのデータ交換には，1990 年代からの歴史的経緯があり，FTD の規程を守った FF 形式を使用してきた．また，3 極で FF の形式が微妙に異なっているので，お互いを検証して変換するプログラムをそれぞれが用意してきた．一方で，近年，大量にデータが測定・蓄積されることから，人間が理解しやすい FF 形式に加えて，コンピュータプログラムが理解しやすい XML 形式によるデータ提供を，INSDC の 3 極がそれぞれ独自の型式で始めた．しかし，3 極間のデータ交換と利用者の利便性の観点から 3 極が共通に対応する標準 XML 型式である INSDC-XML の検討が進んでいる．INSDC-XML が公開されれば，3 極相互ならびに第 3 者においても，プログラムから XML 文書の定義ファイルを参照できるようになり，データファイルを構造解析する過程を大幅に簡略化できるため，バイオインフォマテ

```
Definition：coding sequence; sequence of nucleotides that
            corresponds with the sequence of amino acids in a
            protein (location includes stop codon);
            feature includes amino acid conceptual translation.

Qualifiers （一部省略）
            /allele="text"
            /db_xref="<database>:<identifier>"
            /EC_number="text"
            /evidence=<evidence_value>
            /function="text"
            /gene="text"
            /product="text"
            /protein_id="<identifier>"
            /translation="text"
```

図 -5　大項目 CDS の規程

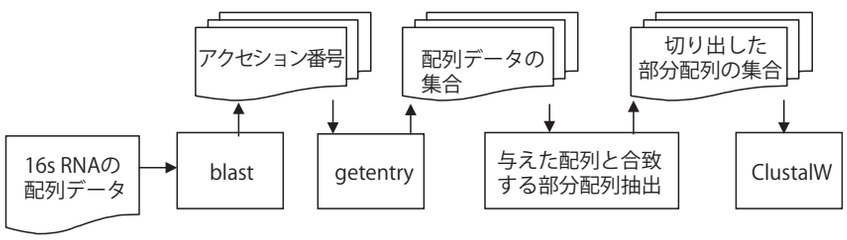

図 -6　相同性検索からアライメントまでのワークフロー

ィクスのアプリケーション開発が加速されるものと思われる．

データサービス

データベースの利用形態は多様である．電子メールによる要求と結果の受け取り，Web ブラウザを介してのインタラクティブな要求と結果の受け取り，ファイル転送（FTP）によるデータの一括ダウンロード，アプリケーションプログラムによる直接アクセスなどである．なかでも近年，分散したサイトにある多様なデータベースと解析ツールを組み合わせて大量のデータをプログラムで一括処理する傾向が強まっている．これに対応して，Web サービスを展開するサイトが増えている [3]．DDBJ においても，電子メールや Web ブラウザ経由で提供してきた検索や解析の機能を SOAP（Simple Object Access Protocol）を介して一通り Web サービスとしても提供している [4]．ここで SOAP を選択したのは，SOAP が通信内容の記述に XML を用い，言語やプラットフォームに依存せず，また，http 上で動作する点からである．なお，我々は独自に大量データ通信や非同期通信の機能を強化した SOAP サーバを運用している．また，Web サービスを組み合わせたワークフロー [5] も試作して公開している（http://www.xml.nig.ac.jp/）．

たとえば，第 1 段階で手持ちの塩基配列に類似した塩基配列の集合をデータベースから抽出し（相同性検索（「本特集 3. 配列データベース検索の現在 」参照）），第 2 段階でその集合に属する塩基配列を整列させる（アラインメント）作業が幅広く使われているが，部品となる Web サービスが整備されていれば，このデータ操作に対応する図 -6 のワークフローを効率よく実現できる．このワークフローでは，

①バクテリアの系統分類に定番として利用されている 16s rRNA の新規配列に対して類似の配列をデータベースから blast（相同性検索で最もよく使われるプログラム）で抽出し，

② blast から出力されたデータベース中のヒット配列のアクセション番号を使って getentry によって該当する全エントリーの配列を取得し，

③取得した配列から，検索に与えた配列と合致した部分を切り出し，

④項目③で取得した部分配列を ClustalW を用いて整列

化して比較
している．このワークフローが整備されていない場合は，Web ブラウザ経由で，blast 用のウィンドウ，getentry 用のウィンドウそして，ClustalW のウィンドウを開いて，各ウィンドウでの出力を次々にコピー＆ペーストしていく必要がある．あるいは，それぞれのプログラムの機能と入出形式を解析してスクラッチからプログラムパッケージを開発する必要がある．ワークフローが確立されている場合は，さらに，ワークフローを構成する部品の Web サービスをより高速なサイトのものへと変更することも容易である．

Web サービスとワークフローは，大量のデータや複数の機能を組み合わせた解析に対する強力な武器であるが，一方で，解決すべき問題も顕在化している．各サイトにそれぞれ導入された Web サービスに「方言」が生まれてしまっているという問題である．この問題を解決するためには，各サイトがそれぞれの Web サービスの機能と入出力形式を標準型式で記述するように誘導していくことが必要である．

バックボーンデータベースからの展開

INSD の塩基配列データは単純な文字列であるが，アノテーション情報を介して生命現象の多様な様相へと広がりを持っている．したがって，INSDC は，3 極間の整合性と相互運用性のための標準ばかりでなく，バイオの世界全般の標準に配慮しさらには提案をして，バイオ分野におけるデータ・情報・知識の共有に貢献していくことが期待される．

バイオ分野における標準の古典かつ典型が生物種の命名である．生物分類学においては，生物を観察し，その結果を比較し，相互によく似ているものをまとめて，他と識別可能なグループ（生物群）を定義し，一定のルールに従ってラベルをつける．これが学名である．膨大な情報が集約された学名によって我々は，特定の生物群の知識と概念を瞬時に共有することができる．塩基配列のデータについても，一定の基準が存在していることによって，データベースの相互運用性が保証され，ひいては，我々は塩基配列データとその生物学的意味（アノテーション情報）を共有することができる．

本稿では，INSDC における標準を，主として「外形」の観点からとりまとめたが，バイオ分野の研究開発には，「意味」の共有とそのための標準が必須である．しかし現実には，データ項目のラベルである gene がサイトによって異なる意味に使われていたり（したがって，対応する遺伝子配列が異なることになる），遺伝子記号が同じ塩基配列であっても生物によって異なる機能に対応するなどの例が多々ある．こうした「意味」にかかわる問題を解決するために，オントロジーの適用が試みられている．たとえば，遺伝子の記述を標準化することを目指した Gene Ontology（GO）が発展してきており（http://www.geneontology.org/），このほかにも生命現象のさまざまな局面を記述するオントロジーが提案されている（「本特集 4. バイオ知識の形成と表現」参照）．また，Web サービスの記述についても，サービス・オントロジーが提案されている [6]．これらのオントロジーをモノ，コト，プロセス，そして概念のグルーピングと捉えれば，古典的な生物分類・命名の考え方がバイオ分野で脈々と息づきかつ広がってきていると見立てることができる．

Web サービスの技術的側面については，従来の GRID 技術（「本特集 6. バイオデータサービス」参照）との融合が急速に進んでいるが，「意味」を上手に扱えるようになれば，さらにセマンティック Web への展開も期待できるであろう．

INSD は，10 年以上にわたりボトムアップで作り上げられてきたデータベースであるが，バイオ研究の進展とともに，質量ともにダイナミックに進化してきた．今後も中長期にわたり，データの構造などが毎年改訂され，規模も毎年大幅に増加し，新たな利用形態が生まれて，さらに進化を遂げるであろう．多様なデータベースの統合やデータからの意味抽出の仕組みとともに，変貌し続けるデータベースの安定運用を実現する仕組みもまた，バイオの分野においても情報の分野においても挑戦に足る課題ではなかろうか．

参考文献

1）五條堀孝，菅原秀明編著：DDBJ の利用法，共立出版，東京 (2005).
2）International Human Genome Sequencing Consortium: Initial Sequencing and Analysis of Human Genome, Nature, Vol.409, pp.860-921(2001).
3）Stein, L. D.: Integrating Biological Databases, Nature Revies Gentics, Vol.4, pp.337-345.
4）Sugawara, H. and Miyazaki, S.: Biological SOAP Servers and Web Services Provided by the Public Sequence Data Bank, Nucleic Acids Res., Vol.31, No.13, pp.3836-3839(2003).
5）Tom Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. and Li, P.: Taverna: A Tool for the Composition Andenactment of Bioinformatics Workflows, Bioinformatics, Vol.20, No.17, pp.3045-3054(2004).
6）Wroe, C., Stevens, R., Goble, C., Roberts, A. and Greenwoord, M.: International Journal of Cooperative Information Systems, Vol.12, No.2, pp.197-224(2003).

（平成 18 年 2 月 2 日受付）