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4回】

塩基配列データベースの現状とその有効活用方法

坊農　秀雅[1]・中村　保一[2]

はじめに：塩基配列データベースとは

1977年にサンガーらによってDNA配列決定法が発明され，さまざまな生物由来の塩基配列の配列決定が可能になった．1980年代に入ってそれらをデータバンクにアーカイブする動きが日米欧で起こり，それぞれDDBJ（DNA Data Bank of Japan），GenBank，EMBL-Bank（のちにENA（European Nucleotide Archive）に改名）という名前のデータベース（DB）が維持されるようになった．これらのDBはお互いにデータを交換しており，中身は同じデータをそれぞれ独自のフォーマットで掲載している．したがって，登録データとしては，たとえばDDBJとGenBank間の違いはない．塩基配列DBはこのどこか1か所に登録すれば残りの二つにも反映されるということになる．この枠組はINSDC（International Nucleotide Sequence Database Collaboration）と呼ばれ，約30年来続けられているものである[1]．塩基配列はこれらのDBに登録して，その登録番号（accession number）がないと論文は出版されないという論文誌側の強制もあり，生命科学分野において最大のDBとなっている．本稿ではそれに端を発する塩基配列DBに関して，データ登録とデータ利用の両側面から現状を紹介する．

塩基配列データベースの現状

塩基配列DBといえば，nr（やnt）を思い浮かべる方も多いのではないだろうか．これは配列類似性検索（BLAST: Basic Local Alignment Search Tool[2]）用に重複した配列を除いた，non-redundantと呼ばれてきたデータセットである．かつてはゲノム配列が決定されていない生物種がほとんどで，配列類似性検索対象にどのDBを使うかとくに考えず，このnrを使えば特に問題なかった．そのため，配列類似性検索の際にDBとして何を使っていたか，意識していなかった利用者が多いと思われる．しかしながら，多くの生物種のゲノム配列が解読された結果，現在は状況が変わっている．プライマー作成などの目的でゲノム上の特定の場所への配列一致検索をすることも多いのではないだろうか．そのような場合には，ターゲットとする生物種のゲノム配列のみをDBとして選ぶなどの工夫をしたほうが検索コストも低く，早く結果が得られるわけである．

それでは，塩基配列DBは今一体どうなっているのか？かつては表1に示したAnnotated sequencesのみで，NCBI（National Center for Biotechnology Information），EMBL-EBI（European Bioinformatics Institute），DDBJの塩基配列DBとしてGenBank，EMBL-Bank，DDBJがあり，お互いにそのデータを交換していた．その状況が，20世紀の末にハイスループットなキャピラリーシーケンサー，そして21世紀にはいって次世代シーケンサー（NGS: Next Generation Sequencer）が開発され，それらから得られる塩基配列（それぞれCapillary reads，Next Generation reads）を収めるため拡充され，DBの種類が増えてきている（表1）．

その結果，DBの種類が多くなりすぎてデータを登録する必要がある際にどこに登録すべきか，また利用する際にもどれを使うべきかわかりにくくなっているのが現状である．図1にそれぞれの手順について大ざっぱに簡略化してフローチャートとしてまとめた．

データ登録，データ利用の両方において，それぞれの事情に応じて利用すべきサービスやDBが異なる．その詳細について以下に述べる．

表1．国際塩基配列データベースがカバーするデータとDB名．INSDCのウェブサイト[1]を基に改変．

	NCBI	EMBL-EBI	DDBJ
Annotated sequences	GenBank	European Nucleotide Archive (ENA)	DDBJ
Capillary reads	Trace Archive		Trace Archive
Next Generation reads	Sequence Read Archive (SRA)		Sequence Read Archive (SRA)
Samples	BioSample		BioSample
Studies	BioProject		BioProject

著者紹介　[1]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
[2]国立遺伝学研究所　大量遺伝情報研究室（教授）　E-mail: yn@nig.ac.jp

塩基配列データを登録する

表1のAnnotated sequences（DDBJやGenBank）がこれまで言及してきた狭義の塩基配列データベースで，かつては塩基配列データを登録するというと，ここにデータを収めることであった．しかしながら，上記のように塩基配列の種類が増え，塩基配列情報に付随するメタデータもきっちり管理する必要が出てきた．そこで，BioSampleとBioProjectというデータアーカイブが創設され，それぞれ配列解読に用いたサンプルと研究のメタデータが収められるようになった．現在，次世代シーケンサーからの塩基配列データ（以下NGSデータ）を登録するという際には，表1のNext Generation readsに加えて，Samples，Studiesに関する情報入力も必須となっている．これらのデータはすべて，DDBJのウェブサイトから登録可能である[3]．

データ入力ウェブインターフェースは，データ入力の種類が多いことやさまざまなルールがあるため煩雑で，残念ながらお世辞にも使いやすいという状態にはなっていない．また，論文数に代表される業績の減少を招く原因となっていると指摘されている，国立大学・研究機関に対する運営費の継続的な削減はDDBJにも影を落としており，とくに近年，ユーザ登録の効率化や支援のための開発費の確保が困難になってきている．しかしながら，ユーザの指摘を受けた改善を可能な範囲で日々進めてはきている．ここでその詳細は述べないが，詳しくはDDBJ handbooks[4]やDDBJ YouTube Channel[5]など，ウェブ上の最新のリソースを参照されたい．

DDBJに入力された情報は「一次」データソースであり，その後若干の修正が入る場合があるものの，基本的には登録者が入力したデータそのものが共有され，直接のユーザや二次データベースに取り込まれて使われていくことになる．再利用されることを考えて，実験条件などのメタデータをしっかり入力していただきたい．そこに書ききれないというのであれば，原著論文のデータとして公開したデータエントリをデータジャーナルにより詳しい記述とともに公表するということも可能となっている．

注意してもらいたいのは，表1にはRefSeqが入っていないということである[6]．それは，RefSeqは一次データソースではなく，NCBIが作っている二次的なデータベースだからである．このRefSeqをはじめとして，リファレンスとなるようアノテーションされたデータベースが作成され，利用されている．その中でもゲノム配列ごとに整理されたデータは有用で使われることも多いので以下で述べる．

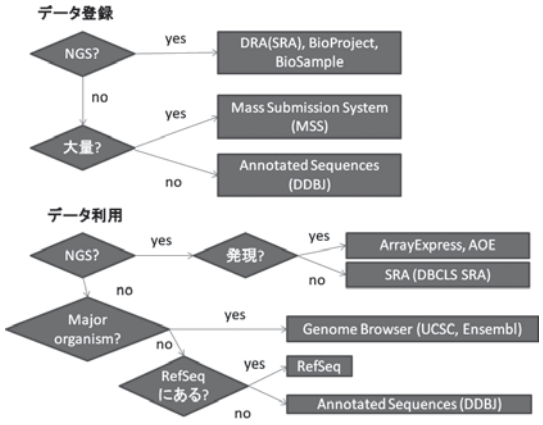

図1．塩基配列データ登録・利用の際のフローチャート

塩基配列データを利用する

冒頭でも紹介したように，かつてはBLASTによる配列類似性検索の対象として塩基配列DBは特に意識されることなく使われてきた．しかしながら，Capillary reads，Next Generation readsにより膨大な塩基配列データが蓄積されるようになり，2016年11月末の執筆時点で，Sequence Read Archive（SRA）から公開されている塩基数は4ペタバイトを超えた[7]．最近，数エントリのSRAレコードに対してはNCBI BLAST[2]のウェブインターフェースからBLAST検索が利用可能となっているものの（Nucleotide BLASTでDatabaseとしてSRAを選択する），すべての塩基配列に対するこれまでと同様に配列類似性検索を実施して利用することは事実上不可能である．

それに代わって，それらをアッセンブルしてつなげられたゲノム配列や，機能アノテーションがなされた遺伝子配列といったデータが使われるようになっている．また，NGSの生データは，それぞれのデータエントリにアノテーションされた実験情報などのメタデータをたよりに検索し，再利用するのが一般的になっている．それらについて以下に紹介する．

ゲノム配列データベース　ゲノム配列解読された生物種ごとにデータが整理されたDBで，これまでよく使われてきた遺伝子ごとではなく，ゲノム上の位置情報に基づいてデータが統合されている．

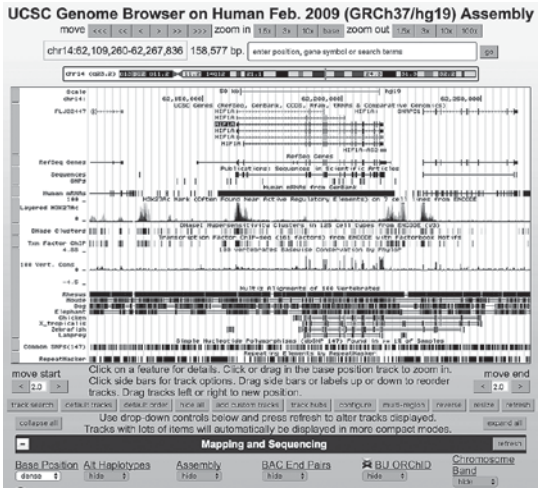

図2．UCSC Genome BrowserでHIF1A遺伝子がコードされたゲノム上の領域を表示した例

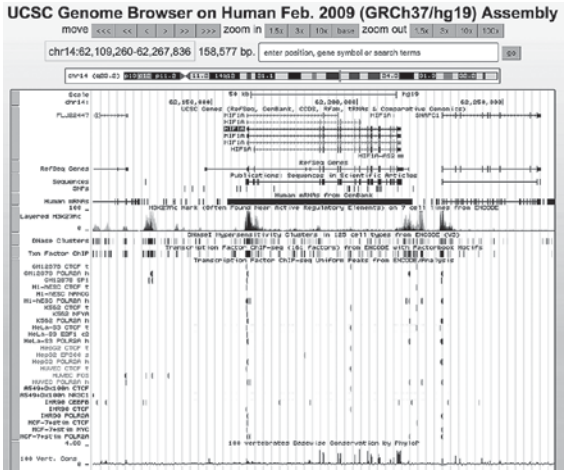

図3．ENCODE-ChIPのトラックがオンにされたゲノムブラウザ

カルフォルニア大学サンタクルーズ校のUCSC Genome Browser[8]とEBIのEnsembl Genome Browser[9]がよく使われている．図2に示したように，ゲノムブラウザでは現在見ている領域に付けられているさまざまな（ゲノム）アノテーションが表示される．アノテーションとは注釈情報のことで，この場合はゲノム上のその位置に関する情報のことである．たとえば，その場所で知られている多型情報にどういうものがあるか，生物種間でどれくらい保存されているか，などである．それらはTrackとよばれる．

隠れているTrackの活用　ヒトゲノム配列解読から10年以上経った現在では，非常に多様なアノテーション情報が利用可能となっているため，大部分のアノテーションのTrackはデフォルトでは非表示となっている．それは利用する研究者によって必要な情報が千差万別だからである．これまでに知られている多型情報を知りたい人もいれば，ChIP-seqにより得られた転写因子結合サイトが今見ている領域のどこにあるのか知りたい人もいるわけである．

したがって，自らの研究に有用なアノテーションを探し，情報を必要な粒度でゲノムブラウザ上に表示することが肝となってくる．たとえば，ENCODE Transcription Factor Binding Trackを‘show’にすると，ENCODEプロジェクトによって得られた転写因子結合サイトのTrackが追加され，現在見ているゲノム領域にどういった転写因子がどこに結合していたか，ひと目でわかる（図3）．図2と見比べると図3の下半分にその情報のTrackが追加されたことがわかるだろう．

Track Hubsによる外部データの統合　さらに，Track Hubs（Track Data Hubsとも呼ぶ）という機能を使うとUCSC Genome Browserのサイトが内部に保持しているデータだけでなく，外部のデータをゲノムブラウザ上に追加することができる[10]．たとえば，理化学研究所のFANTOM5プロジェクトによるCAGE（Cap Analysis of Gene Expression）のマッピングデータはこのTrack Hubsを利用すると，クリック操作だけでゲノムブラウザにそのデータを新たなTrackとして追加できる．

NGSデータ目次：DBCLS SRA　多くの有用なデータは上述のようにすでにゲノムブラウザに統合されている．もちろん，それ以外の有用なデータも数多くあるわけで，それらはNGSのデータベースであるSRAから探してくることになる．かつてのようにBLAST検索で引っ掛けてくることは無理なので，それぞれのデータに付与されたメタデータ（実験目的，使用した機器や試薬，サンプルの情報などの実験手法や実験条件のデータ）から検索してくることになる．その際に有用なサービスとしてDBCLS SRAがある[11,12]．定期的な更新がなされ誰でも無料で利用可能となっている．

さまざまな生物種のゲノムデータ　ヒトやマウスといったゲノムアノテーションがきちんとなされた生物種以外でも，ゲノム情報リソースはさまざまなレベルで情報がまとめられている．UCSC Genome Browserにもいくつか掲載されているが，EnsemblGenomesにはさらに多くの種類の生物のゲノム配列，予測遺伝子配列セッ

ト，予測アミノ酸配列セットがそこから利用可能となっている[13]．

またEnsemblGenomesにデータがまとめられていない生物であっても，そのゲノム配列解読の状況などがGOLD（Genomes Online Database）に公開されていることがある[14]．GOLDによると，2016年11月末の執筆時点で，ゲノム配列解読プロジェクトでCompleteになったものが約9千，Permanent draftsが約6万3千となっている．

最後に：データの共有

ただ配列を解読しただけの塩基配列情報では役に立たない．それに対してどういった生物の，どういう種類であるかといった「アノテーション」がないと使いものにならない．アノテーションは重要である．本稿が，再利用されやすいかたちでデータを公開することや，有効に公開データを活用されることの一助になれば幸いである．

文　献

1) International Nucleotide Sequence Database Collaboration: http://www.insdc.org/ (2016/11/29)
2) BLAST: Basic Local Alignment Search Tool: https://blast.ncbi.nlm.nih.gov/ (2016/11/29)
3) DDBJ | DNA Data Bank of Japan: http://www.ddbj.nig.ac.jp/ (2016/11/29)
4) DDBJ handbooks: http://trace.ddbj.nig.ac.jp/book/ (2016/11/29)
5) DDBJ YouTube Channel: https://www.youtube.com/user/DDBJvideo (2016/11/29)
6) RefSeq: http://www.ncbi.nlm.nih.gov/refseq/ (2016/11/29)
7) Overview: Main: Sequence Read Archive: NCBI/NLM/NIH: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi (2016/11/29)
8) UCSC Genome Browser: http://genome.ucsc.edu/ (2016/11/29)
9) Ensembl Genome Browser: http://www.ensembl.org/ (2016/11/29)
10) Using UCSC Genome Browser Track Hubs: https://genome.ucsc.edu/goldenpath/help/hgTrackHubHelp.html (2016/11/29)
11) DBCLS SRA: http://sra.dbcls.jp/ (2016/11/29)
12) 仲里猛留，坊農秀雅：化学と生物，**54**, 873 (2016).
13) Ensembl Genomes: http://ensemblgenomes.org/ (2016/11/29)
14) GOLD: Genomes Online Database: http://gold.jgi.doe.gov/ (2016/11/29)

（【第5回】は95巻5号に掲載予定です）