次世代シーケンサーデータの解析手法 第9回
ゲノムアノテーションとその可視化、DDBJへの登録

谷澤 靖洋、真島 淳、藤澤 貴智、李 慶範、
中村 保一、清水 謙多郎、門田 幸二

日本乳酸菌学会誌
Vol. 28　No. 1　3 ～ 11（2017）

Japanese Journal of Lactic Acid Bacteria

解　説

次世代シーケンサーデータの解析手法 第9回 ゲノムアノテーションとその可視化、DDBJへの登録

谷澤 靖洋[1]、真島 淳[2]、藤澤 貴智[1]、李 慶範[2]、
中村 保一[1*]、清水 謙多郎[3]、門田 幸二[3*]

[1] 国立遺伝学研究所生命情報研究センター
[2] 国立遺伝学研究所 DDBJ センター
[3] 東京大学大学院農学生命科学研究科

論文発表は研究の一部である。そして多くの学術雑誌は、論文発表される塩基配列を国際塩基配列データベース（International Nucleotide Sequence Database Collaboration; INSDC）に登録することを義務付けている。この意味において、これまで本連載で述べてきた解析手法に関する事柄とは若干スタンスが異なるが、ゲノム配列の登録作業もまた研究の一環といえる。第9回は、前回までに構築したゲノム配列 *Lactobacillus hokkaidonensis* LOOC260^T の総仕上げとして、染色体複製開始点の同定と回転、ゲノムアノテーション、INSDC を構成する日米欧三極の一つである DNA Data Bank of Japan（DDBJ）への登録、そして DNAPlotter を用いたアノテーション結果の描画について解説する。ウェブサイト（Rで）塩基配列解析（URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html）中に本連載をまとめた項目（URL: http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_book_JSLAB）が存在する。ウェブ資料（以下、W）や関連ウェブサイトなどを効率的に活用してほしい。

Key words：replication origin, annotation, INSDC, DDBJ

はじめに

第9回は、前回最後に作成した計3配列からなる LH_draft2.fa（2,400,619 bytes； 約 2.4MB； 第8回 W28）を利用する[1-2]。Bio-Linux[3] 環境下での主な作業は、最初の1項目（複製開始点の同定）のみであり、LH_draft2.fa ファイルのみ手元にあればよい。また、必ずしも第8回終了時点の解析環境をスタート地点とする必要はないため、wget コマンドを用いて LH_draft2.fa を任意の作業ディレクトリ上にダウンロードしておけば、例えば連載第4回終了時点の解析環境下でもよい。本稿では、共有フォルダ「~/Desktop/mac_share」上で作業を行う［W1-1］。

*To whom correspondence should be addressed.
Phone ： +81-3-5841-2395
Fax ： +81-3-5841-1136
E-mail ： yn@nig.ac.jp or kadota@bi.a.u-tokyo.ac.jp

複製開始点の同定

DNA 二重らせんモデルの提唱へとつながったシャルガフの法則[4]は、「細胞内における A と T および C と G の塩基の量比がそれぞれほぼ等しい（%A＝%T および %C＝%G）」というものである。この法則は、後に DNA 1本鎖に対しても適用可能であることがわかり、第2項（Chargaff's second parity rule[5]）［W1-2］として付け加えられている。しかし乳酸菌を含む多くのバクテリアにおいて、DNA の複製が始まる点（origin of replication；以下、*oriC*）の近傍では局所的に偏りが見られ、C が多い領域から G が多い領域に切り替わることが知られている[6-11]。GC skew は、この偏りを（G－C）/（G＋C）で表したものであり[7]、値の正負が入れ替わるポイントから *oriC* を

検出する指標として用いることができる。また、*oriC* 領域近辺には複製開始に関与する DnaA タンパク質をコードする遺伝子が存在し、その上流には DnaA box と呼ばれる DnaA の結合領域となる反復配列が存在することも *oriC* 同定の手がかりとなる[12-14]。GC skew は（C－G）/（G＋C）として定義される場合もあるが[6]、符号が逆転するポイントという意味では同じであり、どちらの数式が正統であるかといった議論は本質的ではない[11]。

環状ゲノムの場合、どこを配列の起点として表記するかは明確にルールが決まっているわけではないが、慣例として***dnaA* 遺伝子が配列の先頭となるように“回転”させておく**ことが多い。前回予備的に行った LH_hgap.fa の DFAST[15] アノテーション結果を眺めると、染色体に相当する sequence1 上のみに *dnaA* がコードされていた［W2-1］。LH_draft2.fa を入力として再度予備的に行った DFAST アノテーション結果においては、chromosome 上の**相補鎖側**の［1436009, 1437325 bp］に *dnaA* がコードされていた［W2-2］。これは、*oriC* が *dnaA* 遺伝子の翻訳開始位置である 1,437,325 番目の塩基周辺となるであろうことを意味する。

ウェブツール Ori-Finder[16] を用いて *oriC* の同定を行う［W3-1］。Ori-Finder は、① GC skew を効率的に計算する方法の 1 つである Z-curve 法[17]、② DnaA box の分布、そして③ *dnaA* 遺伝子の予測結果を総合的に評価している。Ori-Finder の入力は 1 配列のみからなる single-FASTA 形式ファイルである（プログラム開発者との確認済み）。ここでは multi-FASTA ファイル（LH_draft2.fa）から、染色体配列に相当する最初の 1 件分（つまり 2 行分）のみを抽出した single-FASTA ファイル（seq1_draft.fa；［W3-2］）を入力として Ori-Finder を実行した［W3-3］。Ori-Finder 実行結果として、3 つの *oriC* 候補領域が得られた［W3-4］：①［1396338, 1396990 bp］、②［1435834, 1436008 bp］、③［1437326, 1438021 bp］。我々は、3 番目の *oriC* 候補領域の先頭部分（1,437,326 番目の塩基）が前述の DFAST アノテーション結果から得られた *dnaA* 遺伝子の翻訳開始位置（1,437,325 番目の塩基）と近接していたことから、**基本的には 1,437,325 番目の塩基を起点として回転させればよい**と判断した。

若干本題からずれるが、予測された *dnaA* 遺伝子の開始コドンは、一般的によく知られている ATG ではなく GTG になっている［W4-1］。バクテリアの場合は GTG（や TTG）も開始コドンをコードしているため、この結果は間違いではない［W4-2］。但し、使用する遺伝子領域予測ソフト次第で開始コドンが異なる位置になることは珍しくない。実際、我々以外にも *L. hokkaidonensis* の基準株のゲノムを解読し公開しているグループが存在するが、彼らの結果（アクセッション番号：JQCH01000005.1）では、今回の結果よりも上流側に存在する ATG を翻訳開始点として予測している［W4-3］。どちらが真の翻訳開始点であるかの議論は行わないが、異なって予測されうるという点には留意するべきであろう。ここでは将来的に訂正を行う可能性を踏まえて、*dnaA* 遺伝子の翻訳開始位置（1,437,325 番目）よりも多少上流側（例えば 100 塩基上流側の 1,437,425 番目）に起点を設定する方針にした。

具体的な手順としては、まず *dnaA* 遺伝子を含む領域［1, 1437425 bp］とそれ以外の領域［1437426, 2277983 bp］を入れ替えた（回転させた）後に、相補鎖変換を行えばよい［W4-4］。ここでは、アドホックな Python スクリプトを実行するやり方［W4-5］、および Linux コマンドと EMBOSS[18] の revseq プログラムを併用するやり方［W4-8］を示した。

ゲノムアノテーション

ゲノム分野におけるアノテーションとは、塩基配列に対する生物学的意味を注釈付けすることである。大まかには塩基配列から遺伝子をコードする領域を見つけ出す構造アノテーション[19]、そしてその領域が果たす役割に関する情報を付加する機能アノテーション[20] に分けられる。構造アノテーションでは、アミノ酸に翻訳される領域（coding sequence；CDS）や、tRNA、rRNA などの遺伝子をコードする領域をはじめ、リピート領域やオペロン構造、真核生物であればエクソン構造といった様々な構造情報の推定も行われることがある。機能アノテーションでは、BLAST 検索による配列類似性や、Pfam・Rfam などによる配列モチーフ、タンパク質ドメイン構造予測結果を基にした遺伝子産物名の推定が中心的な作業となる。また、その推定がどのような根拠に基づいているかといった信頼度に関する情報や、外部のデータベース（以下、DB）への参照情報の追加も含まれる。

一般に、あるデータに対して注釈付けされた（アノテーションされた）付随情報のことを、データについてのデータという意味で**メタデータ**と呼ぶ[21]。INSDC を構成する DDBJ[21]・GenBank[22]・ENA[23] に登録された配列データにおいては、次のような配列自体に対する付随情報を指して特にメタデータと呼ぶ場合が多い。これらも広義のアノテーションといえる。

- 登録者情報（誰がどのような研究プロジェクトによって登録したか）
- 生物種に関する情報（どのような生物に由来する配列のものであるか）
- 実験条件（どのような方法で決定された配列であるか）
- 文献情報（その配列がどのような文献で述べられているか）

尚、INSDC 設立当初は現在の正式な DB 名である ENA ではなく、EMBL（European Molecular Biology Laboratory）という組織が提供する（塩基配列 DB の通称としての）EMBL であった。これは DDBJ という組

織（DDBJ センター）が運営する（塩基配列 DB としての）DDBJ と同じような位置づけである。ENA は DB 名であり、運営主体は EMBL 傘下の EBI（The European Bioinformatics Institute）である。GenBank が DB 名であることは広く知られているが、その運営主体は NCBI（National Center for Biotechnology Information）である。これらの理由により、運営主体と DB 名を連結した NCBI GenBank、EMBL-EBI ENA といった表現もよくなされる。連載第 3 回[24]で紹介した次世代シークエンサデータリポジトリ[25]もまた、INSDC を構成する三極（DDBJ・NCBI・EMBL-EBI）で運用されている。特に説明もなく DDBJ SRA（通称 DRA）・EMBL-EBI ENA（通称 ENA）・NCBI SRA（通称 SRA）と記載したが、運営主体と DB 名を連結した表記であったことがわかる。INSDC の三極はデータの共有を行っているため、1 つの機関で登録されたデータは他の機関にも同期される。しかし、次節でも述べるようにデータの書式は DB ごとに異なっている。

INSDC フラットファイルの構成

INSDC に登録されたゲノムとアノテーション情報の記載内容について述べる。INSDC のデータ公開形式はフラットファイルと呼ばれ、**entry**・**feature**・**qualifier** の 3 つの階層構造からなる。GenBank のフラットファイル形式（GenBank 形式と呼ばれる）は、多くのソフトウェアが対応している標準的な形式である。DDBJ は、GenBank と似た形式である DDBJ 形式を採用している。ENA のフラットファイル形式は、他とはやや異なっており独自の情報も追加されている。また、歴史的経緯から ENA 形式ではなく EMBL 形式と呼ばれる。この階層構造はいずれの形式においても共通であり、自作プログラムでフラットファイルから情報を抽出する場合や、INSDC へ配列を登録する場合にもこの構造を理解しておくことが重要である。もちろんこれらのファイル形式を取り扱うためのプログラム群は整備されている。それが EMBOSS[18]（の seqret プログラム）や、連載第 1 回[26]でも紹介した BioPerl[27]・Biopython[28]・BioRuby[29]などである。

階層の最上位である **entry** は、塩基配列の登録単位を表す。単一の遺伝子の塩基配列が 1 つの entry を構成することもあれば、コンプリートゲノムのように 1 本の染色体で 1 つの entry を構成することもある。ドラフトゲノムでは 1 本のコンティグやスキャッフォールドが 1 つの entry となる。entry 中の注釈付けされた各領域のことを **feature** と呼ぶ。feature には、ゲノム中での位置を示すための location 情報、および feature の内容をさらに細かく記述するためのいくつかの **qualifier** が含まれている。

DDBJ 形式の **entry** の具体例として、大腸菌 K-12 株の染色体配列の登録内容を示す（**図 1**；http://getentry.ddbj.nig.ac.jp/getentry/na/U00096.3）。U00096 はこの entry の識別子（アクセッション番号）を表し、ピリオドの後の 3 はバージョン番号を示す［W5-1］。これは、配列が最初の登録後から 2 回更新されていることを意味する。登録直後は U00096.1、1 回目の更新後に U00096.2、そして 2 回目の更新後に U00096.3 となる。上から順に見ていくと、生物種名とともに登録者情報や文献情報といったメタデータが記載されている［W5-2］。U00096 は、大腸菌の中でもモデル株といえる K-12 株のものである。この entry には計 18 件の REFERENCE が記載されており、1 件目のものは本菌株のゲノムが最初に解読された 1997 年の原著論文[30]である。

続いてこの entry にアノテーションされた **feature** が並ぶ［W5-3］。source feature は、配列全体に対する記述を行う特別な feature であり、基本的には各 entry の最初に 1 つだけ存在する［W5-4］。gene feature は CDS・rRNA・tRNA などの遺伝子領域を記載する際に用いられ、次の CDS feature はタンパク質をコードする領域について記載している［W5-5］。バクテリアは通常 1 遺伝子 1 タンパク質であるため、わざわざ gene feature を記載する意味合いが感じられないかもしれない。しかしながら、真核生物でスプライスバリアントがある場合には、1 つの gene feature に対して複数の CDS feature が結びつけられることを想像してもらえれば納得できるであろう。

CDS feature の内容をさらに細かく記述したものが **qualifier** であり、その具体例が遺伝子シンボルを示す gene qualifier、遺伝子産物名を示す product qualifier、翻訳されたアミノ酸配列を表す translation qualifier などである［W5-6］。このエントリ中にはないものの、experiment qualifier によって実験的に裏付けがなされていることが示されているものもある［W5-7］。

gene や CDS 以外の feature の種類としては、tRNA feature、rRNA feature、repeat_region feature［W5-8］、ncRNA feature［W5-9］などが存在し、それぞれを特徴付ける qualifier によって詳細な情報が記述されている。各 feature の記載方法は、INSDC の Feature Table Definition で定められている［W5-10］。他にどのような種類の feature があるかや、feature ごとにどのような qualifier が使えるかに関する情報は、DDBJ が公開している対応表がわかりやすいだろう［W5-11］。

locus_tag について

locus_tag という qualifier は、アノテーションされた遺伝子領域を指し示すための一意な識別子である［W6-1］。大腸菌 K-12 株の染色体配列（U00096）中でも locus_tag qualifier が見られるが、このゲノムは古くに登録されたものであり現在の locus_tag の形式とは異なっている［W6-2］。現在は、3 文字以上の英数字

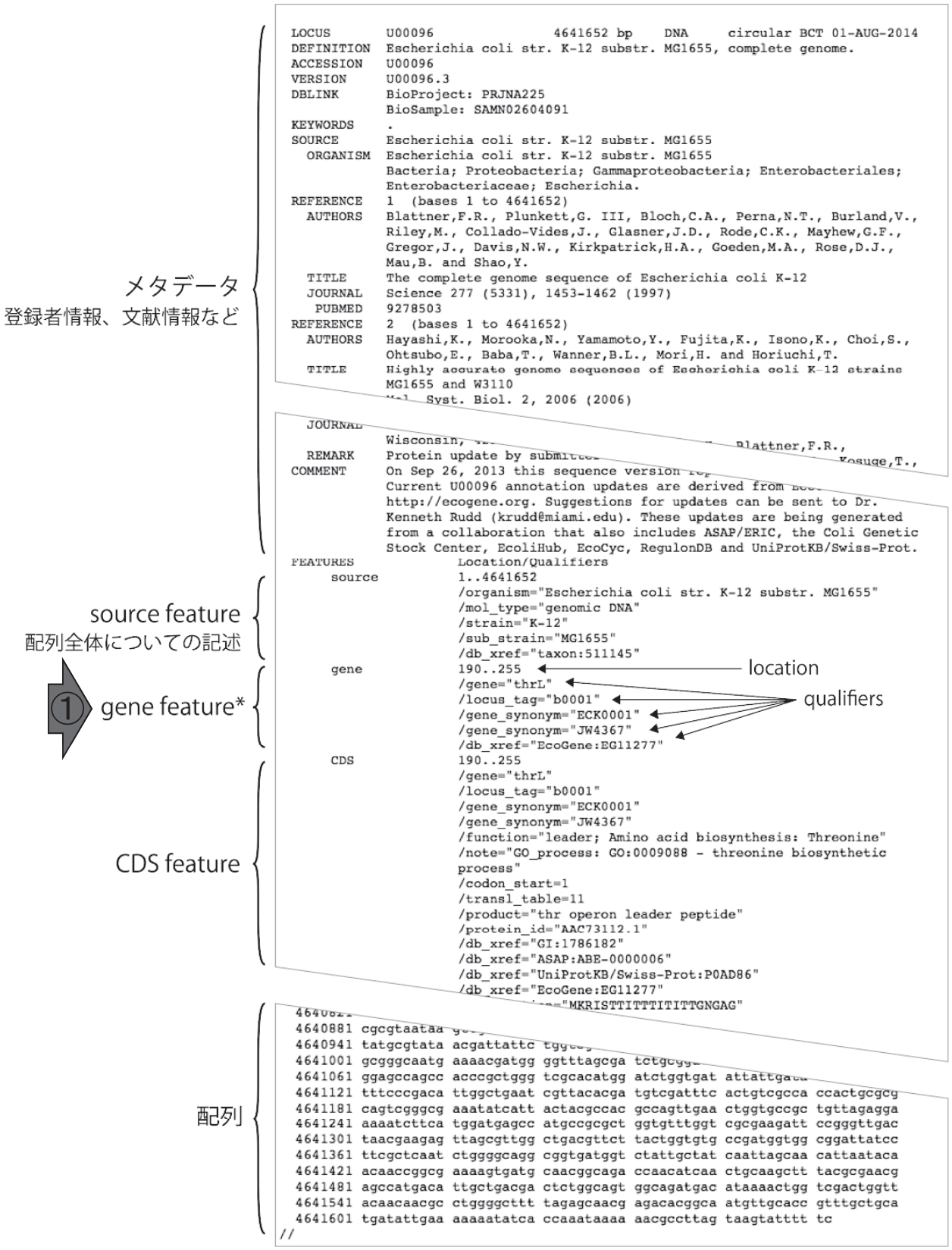

図 1．DDBJ フラットファイル形式の例
大腸菌 K-12 株の染色体配列。① gene feature* は、CDS、rRNA、tRNA などの遺伝子領域に対してつけられる。真核生物でスプライスバリアントがある場合には、1 つの gene feature に対して複数の CDS feature が結びつけられることがある。

からなる locus_tag の prefix と、登録者が任意に決められる英数字（tag）とをアンダースコアでつないだ形式（prefix_tag）に決められている。例えば、2014 年に登録された *L. hokkaidonensis* [1] の GenBank accession 番号 AP014680 を眺めると、1 つめの feature の locus_tag が LOOC260_100010 となっており、現在の形式と合致している［W6-3］。prefix は、同一ゲノム内では同じものを用いる必要がある。また、アノテーションされたゲノムを INSDC に登録する際には、他のゲノムで使用されていない locus_tag prefix を予め取得しておかねばならない。AP014680 の場合は、菌株名と同じ LOOC260 がそれに相当する。アノテーションをつけずに配列のみを登録する場合には、locus_tag prefix の取得は不必要である。tag はゲノム中での出現順に通し番号を使用することが多いが、後から追加・挿入されることを想定して 10 飛びの値を使うこともある［W6-4］。また、tRNA には t0001、rRNA には r0001 など区別して割り当てることもある。

他には、gene feature とそれに対応する CDS feature などの間では同じ locus_tag を割り当てなければならないなどの決まりがある。locus_tag の用法に関する公式情報にも目を通しておくとよいだろう［W6-5］。尚、DDBJ や ENA では登録時に gene feature を要求していないため、これらの 2 機関で登録されたデータには gene feature が存在しないものがある。しかし、GenBank にデータが同期される際には GenBank 上で自動的に gene feature が付加される［W6-6］。逆に、GenBank で登録されたデータについては、DDBJ や ENA 上で眺めても gene feature が存在する。

DFAST を用いたゲノムアノテーション

dnaA 遺伝子（の転写開始点上流 100 塩基）が先頭となるように変換した乳酸菌コンプリートゲノムファイル（LH_complete.fa;［W4-6］）を入力として、DFAST [15] を用いた本番のゲノムアノテーションを行う［W7-1］。DFAST は、バクテリア用ゲノムアノテーションパイプラインである Prokka [31] をベースとして、乳酸菌用に整備された参照データベースを組み合わせたものである [2]。また、アノテーションされたゲノム配列を DDBJ に登録するための支援機能をもつ（正確には登録に必要なファイルを半自動生成できる）のが特徴である。我々はこれまで DFAST 経由で 10 件以上の *Lactobacillus* 属ゲノム配列の DDBJ への登録を行ってきたが、いずれも手作業での修正をほとんど必要とせずに完了できている。これまでの予備的なアノテーション（W2-2；第 8 回の W4）作業の手順からもわかるように、ログインなど事前登録の必要なく無償で利用可能であり、DDBJ への登録を行わずにアノテーションのみを行うこともできる。ゲノムサイズ 2 ～ 3Mbp 程度の典型的なサイズの乳酸菌であれば、5 分ほどで結果を得ることができる。

Prokka は、CDS、rRNA、tRNA の予測といった基本的なアノテーションに加えて、CRISPR（Clustered regularly interspaced short palindromic repeats）[32] やシグナルペプチドの検出機能を備えている。Prokka の高速なアノテーションは、予測された CDS を複数の参照アミノ酸配列 DB に対して段階的に検索していくことで実現されている。はじめに近縁種から得られた配列を中心に構成された**より信頼できる**参照 DB に対して BLAST 検索を行い、そこでヒットしなかった遺伝子は**より包括的な**参照 DB を用いて検索される。そこでもヒットしなかった遺伝子については、最終的に隠れマルコフモデルを用いたモチーフ・ドメイン検索ソフト HMMER3 [33] を使って Pfam [34] や TIGRFAMs [35] などの DB に対して検索を行う。DFAST は、第 1 段階目の検索を乳酸菌（主に *Lactobacillus* 属および *Pediococcus* 属）用に独自に用意した参照 DB に対して行うことで、DDBJ にそのまま登録可能なレベルのアノテーションが可能となっている。現在の DFAST は、大腸菌やシアノバクテリアなどの生物種にも同様の参照 DB を拡張したほか、他の系統群のバクテリアにも使用できる汎用的な参照 DB を用意している。

DFAST 独自の機能として、CheckM [36] を用いたゲノムのクオリティチェックと average nucleotide identity（ANI）を使った系統名のチェックもオプションで行うことができる［W7-2］。CheckM は、系統群ごとにマーカーとして用いる遺伝子セットを定義しており、このマーカーの有無を調べることによってゲノムの completeness や contamination といった指標を算出している［W7-3］。マーカーには、通常 1 つのゲノムに 1 つのコピーのみ存在する遺伝子が選ばれている。このため、入力ゲノム配列中に同定できたマーカーの数が少ないと completeness が低く報告され、同じマーカーがゲノム中に複数コピー同定されると contamination の値が高く報告される傾向となる。以前我々が *Lactobacillaceae* 科乳酸菌ゲノム 743 件について CheckM を実行したところ、743 件中 654 件（約 88%）が CheckM 内で示された基準（95% 以上の completeness と 5% 未満の contamination）を満たした。高い completeness に越したことはないという一方で、ゲノムサイズが小さく縮退が進んでいるような菌種では低い傾向にあるという報告もある [37-38]。値の解釈に迷ったときは、同種のゲノムと比較検討するとよいだろう。

系統名のチェックに用いられる ANI は、全ゲノム情報を用いた種同定のための指標であり、2 つのゲノム間のアラインメントにおいて相同性が認められた領域の平均塩基一致度から算出される。ANI は同じ菌種に属するゲノムであれば、概ね 95% 以上の値を示すことが知られている [39]。DFAST では、基準株を中心に選定した 185 菌種（亜種も含む）の representative genomes を対象として ANI を計算することで、アップロードしたゲノムの系統的位置

付けのチェックが可能である［W7-4］。*Lactobacillus* 属には、*plantarum* グループや *casei* グループのように 16S rRNA の配列を用いた種同定が難しい菌種が含まれているが、ANI を用いることで十分な分解能で *Lactobacillus* 属のゲノムを同定できることが確かめられている[15]。DFAST では、誤った情報が公共 DB に登録されることを事前に防ぐ目的で、このような菌種同定ツールを提供している（**図 2**）。また、DFAST の関連 DB である DAGA（DFAST Archive of Genome Annotation）では、公開されている乳酸菌ゲノムデータに対し系統名や品質の査定を行った上で DFAST による再アノテーション情報を提供している。前述の representative genomes についても、DAGA で参照することができる。

DFAST 実行結果画面の "Features" タブをクリックすると、アノテーションされた feature の一覧が見られる［W8-1］。入力ファイル（LH_complete.fa）は、*dnaA* 遺伝子の翻訳開始位置よりも 100 塩基上流が配列の先頭になるように変換したものであった［W4］。アノテーション結果が実際にそうなっていることの確認は行っておくべきであろう［W8-2］。Feature table の View ボタンを押すと予測された遺伝子の塩基配列やアミノ酸に翻訳した配列が確認でき［W8-3］、そのまま NCBI の BLAST ウェブサービスに対して検索をかけることもできる［W8-4］。Edit ボタンを押すと、アノテーションされた遺伝子名、遺伝子シンボルを編集することができる。また、Note 欄に特記事項を加えることもできる［W8-5］。これらの編集した情報はダウンロードされる結果ファイルにも反映されるため、画面上でも正しく反映されているか確認しておくとよい。

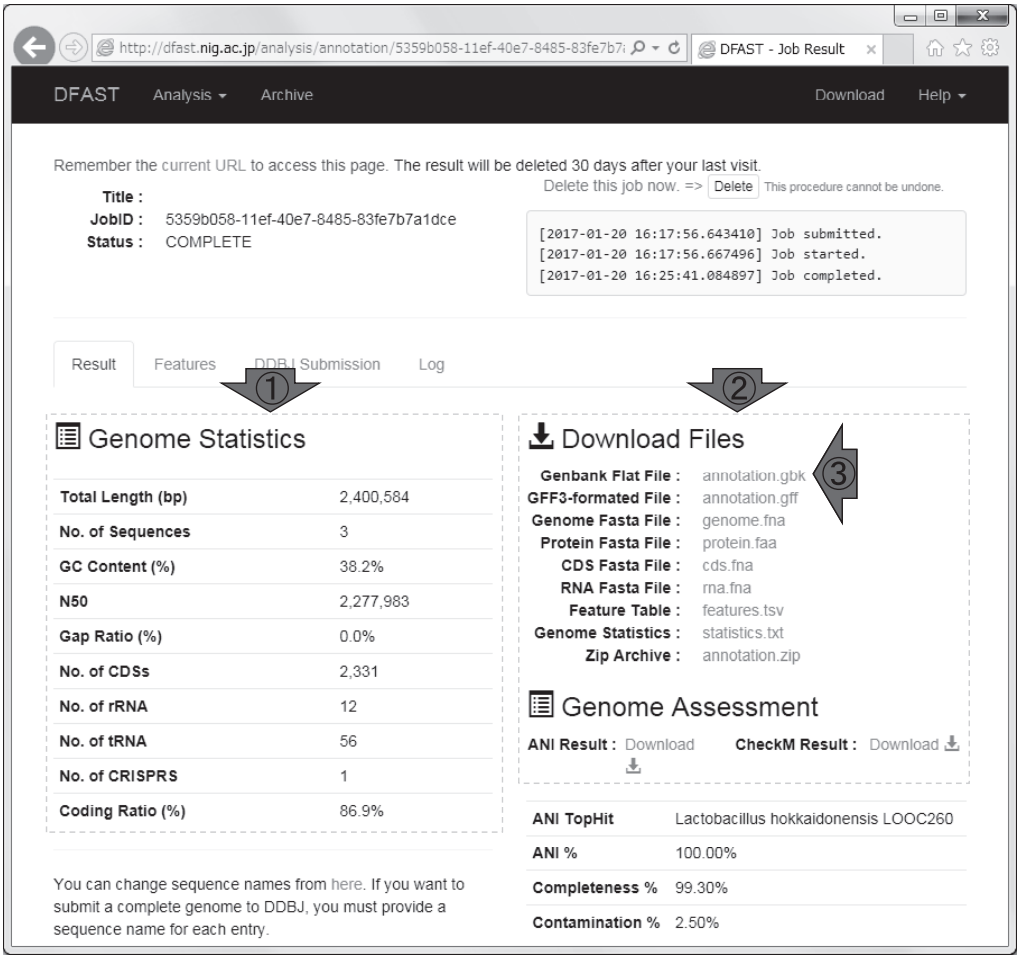

図 2. DFAST 実行結果画面

dnaA 遺伝子（の転写開始点上流 100 塩基）が先頭となるように変換した乳酸菌コンプリートゲノムファイル（LH_complete.fa; ［W4-6］）を入力として DFAST を実行した結果画面［W7］。① Genome Statistics のところで、入力配列の基本統計量などの情報が得られる。② Download Files および Genome Assessment のところで、アノテーション結果や ANI および CheckM 実行結果ファイルをダウンロードすることができる。③ Genbank 形式の DFAST アノテーション結果ファイル（annotation.gbk）は、**図 3** の入力として使われている。

DDBJ 登録用ファイルの作成

DFAST 実行結果画面の "DDBJ Submission" タブをクリックすると、DDBJ への登録ファイルを作成する画面が表示される［W9-1］。ここでは、ゲノムを登録するために必要な登録者情報や文献情報などのメタデータを入力することができる。画面の指示に従って入力項目を埋めていけば、登録用ファイルを作成できるようになっている。今回のようなコンプリートゲノムの登録の場合は、はじめに各配列の名称（entry 名）、種別（染色体またはプラスミドなど）、形状（直鎖または環状）の指定を行う必要がある［W9-2］。ドラフトゲノムの場合には、デフォルトでは sequence## となっている配列の接頭辞を、例えば contig## のように変更することができる。

以降の作業は通常、DDBJ への登録と並行して行う。具体的には、DDBJ の登録アカウントの取得、BioProject の登録、BioSample の登録などである。locus_tag prefix は、BioProject または BioSample の登録時に取得可能である［W6-3］。これらの情報を取得しておかないと、「2. Input Metadata」で行うメタデータ情報の実際の入力段階で行き詰まる［W9-4］。ここでは、作業後に得られる DDBJ 形式ファイルのイメージを掴むべく、仮の値を入力して最後まで作業を進めた［W10］。画面の指示に従って入力項目を埋めていくことで、実際に公開されているもの（AP014680）とほぼ同じ形式のものが得られることを実感できるであろう［W10-12］。W10 の内容については、DDBJ への登録の説明と合わせて第 10 回でも改めて解説する予定である。

アノテーション結果の描画

最後に、DNAPlotter [40] を用いた *L. hokkaidonensis* の染色体マップを示す（**図 3**；W12）。入力は、Genbank 形式の DFAST アノテーション結果ファイル（annotation.gbk）である。①時計の 0 時に相当する部分が複製開始点である。一番外側から、②順鎖上および相補鎖上の CDS、

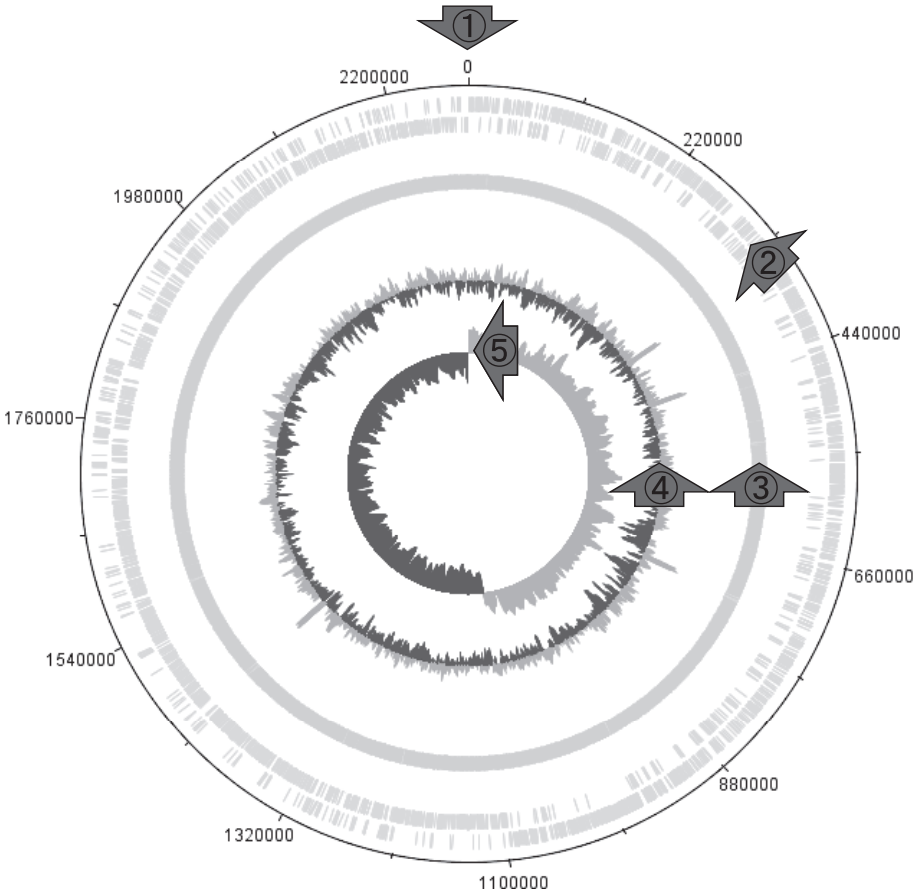

図 3. *L. hokkaidonensis* L00C260^{T} **の染色体マップ**
Genbank 形式の DFAST アノテーション結果ファイル（annotation.gbk）を入力として、DNAPlotter を用いて染色体配列に相当する sequence1 のアノテーション結果のみを表示［W12］。①複製開始点。②外側が順鎖上の CDS、内側が相補鎖上の CDS［W12-4］。③ tRNA［W12-6］。④ GC 含量［W12-11］。⑤ GC skew［W12-18］。①の複製開始点付近において、C が多い領域から G が多い領域に切り替わっているのがわかる。

③ tRNA、④ GC 含量、そして⑤（G－C）/（G＋C）で定義した GC skew が示されている。実際の複製では、①の複製開始点から時計回りに進む順鎖側、そして反時計回りに進む相補鎖側がリーディング鎖となる。**リーディング鎖により多くの遺伝子が存在**している傾向が、② CDS の密度の違いから確認できる。また、入力配列は *dnaA* 遺伝子が先頭に来るように配列を回転させたものであるため、①の**複製開始点付近において C が多い領域から G が多い領域に時計回り方向で切り替わっている**のがわかる。使用プログラムおよび色の違いはあるものの、原著論文[1]の Fig. 1A と本質的に同じであることもわかるであろう［W12-20］。

今回は、DNAPlotter が 3 配列からなる annotation.gbk を読み込んでも、染色体配列に相当する 1 番目の配列のみしか認識しないことを逆手にとって染色体マップを作成した。しかし、DNAPlotter のように 1 番目の配列しか認識しないことがわかっているようなプログラムを利用する場合には、描画したい単一の配列情報のみを抽出したファイルを作成・利用する。第一義的には想定外の挙動やバグを防ぐため、そしてプラスミド配列に相当する 2 番目のみ、あるいは 3 番目のみの配列を入力としたプラスミドマップを作成できるようにするためである。ここでは 1 番目のみの配列を抽出する目的で awk コマンドを［W14］、2 番目以降の配列を抽出する目的で grep・head・tail・パイプコマンドを併用する例[24]を示した。これらもまた、Linux コマンドを実際の現場で利用する好例であろう。

謝　辞

本連載の一部は、科学技術振興機構 バイオサイエンスデータベースセンター（JST-NBDC）、および情報・システム研究機構 国立遺伝学研究所（遺伝研）との共同研究（2012-2088, 2013-2070）の成果によるものです。また、JSPS 科研費 JP25712032, JP15K06919 の助成を受けたものです。農業・食品産業技術総合研究機構 畜産草地研究所の遠野雅徳先生には、本稿で用いた乳酸菌ゲノム配列解読論文の責任著者として、また編集委員会委員として本連載の円滑な執筆環境構築に尽力いただきました。

参 考 文 献

1） Tanizawa Y, Tohno M, Kaminuma E, Nakamura Y, Arita M. (2015) Complete genome sequence and analysis of *Lactobacillus hokkaidonensis* LOOC260^{T}, a psychrotrophic lactic acid bacterium isolated from silage. BMC Genomics **16**: 240.
2） 谷澤靖洋，神沼英里，中村保一，遠野雅徳、寺田朋子、清水謙多郎，門田幸二（2016）次世代シーケンサーデータの解析手法：第 8 回アセンブリ後の解析．日本乳酸菌学会誌 **27**：187-195.
3） Field D, Tiwari B, Booth T, Houten S, Swan D, et al. (2006) Open software for biologists: from famine to feast. Nat Biotechnol **24**: 801-803.
4） Chargaff E. (1951) Structure and function of nucleic acids as cell constituents. Fed Proc **10**: 654-659.
5） Rudner R, Karkas JD, Chargaff E. (1968) Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci U S A **60**: 921-922.
6） Lobry JR. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol **13**: 660-665.
7） Rocha E. (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol **10**: 393-395.
8） Bird RE, Louarn J, Martuscelli J, Caro L. (1972) Origin and sequence of chromosome replication in *Escherichia coli*. J Mol Biol **70**: 549-566.
9） Prescott DM, Kuempel PL. (1972) Bidirectional replication of the chromosome in *Escherichia coli*. Proc Natl Acad Sci U S A **69**: 2842-2845.
10） Hiasa H, Marians KJ. (1994) Primase couples leading- and lagging-strand DNA synthesis from *oriC*. J Biol Chem **269**: 6058-6063.
11） Arakawa K, Tomita M. (2012) Measures of compositional strand bias related to replication machinery and its applications. Curr Genomics **13**: 4-15.
12） Meijer M, Beck E, Hansen FG, Bergmans HE, Messer W, et al. (1979) Nucleotide sequence of the origin of replication of the *Escherichia coli* K-12 chromosome. Proc Natl Acad Sci U S A **76**: 580-584.
13） Kimura M, Miki T, Hiraga S, Nagata T, Yura T. (1979) Conditionally lethal amber mutations in the *dnaA* region of the *Escherichia coli* chromosome that affect chromosome replication. J Bacteriol **140**: 825-834.
14） Ogasawara N, Moriya S, von Meyenburg K, Hansen FG, Yoshikawa H. (1985) Conservation of genes and their organization in the chromosomal replication origin region of *Bacillus subtilis* and *Escherichia coli*. EMBO J **4**: 3345-3350.
15） Tanizawa Y, Fujisawa T, Kaminuma E, Nakamura Y., Arita M. (2016) DFAST and DAGA: Web-based integrated genome annotation tools and resources. Biosci Microbiota Food Health **35**: 173-184.
16） Gao F, Zhang CT. (2008) Ori-Finder: a web-based system for finding *oriC*s in unannotated bacterial genomes. BMC Bioinformatics **9**: 79.
17） Zhang R, Zhang CT. (1994) Z curves, an intutive tool for visualizing and analyzing the DNA sequences. J Biomol Struct Dyn **11**: 767-782.
18） Rice P, Longden I, Bleasby A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16**: 276-277.
19） De Bodt S1, Raes J, Florquin K, Rombauts S, Rouzé P, et al. (2003) Genomewide structural annotation and evolutionary analysis of the type I MADS-box genes in plants. J Mol Evol **56**: 573-586.
20） Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, et al. (2001) Functional annotation of a full-length mouse cDNA collection. Nature **409**: 685-690.
21） Mashima J, Kodama Y, Fujisawa T, Katayama T, Okuda Y, et al. (2017) DNA Data Bank of Japan. Nucleic Acids Res **45**: D25-D31.
22） Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ. (2017) GenBank. Nucleic Acids Res **45**: D37-D42.
23） Toribio AL, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L. (2017) European Nucleotide Archive in 2016. Nucleic Acids

Res **45**: D32–D36.
24) 孫建強，三浦文，清水謙多郎，門田幸二（2015）次世代シーケンサーデータの解析手法：第3回 Linux 環境構築から NGS データ取得まで．日本乳酸菌学会誌 **26**：32–41.
25) Kodama Y, Shumway M, Leinonen R; International Nucleotide Sequence Database Collaboration. (2012) The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res **40**: D54–56.
26) 門田幸二，孫建強，湯敏，西岡輔，清水謙多郎（2014）次世代シーケンサーデータの解析手法：第1回イントロダクション．日本乳酸菌学会誌 **25**：87–94.
27) Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res **12**: 1611–1618.
28) Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**: 1422–1423.
29) Goto N, Prins P, Nakao M, Bonnal R, Aerts J, et al. (2010) BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics **26**: 2617–2619.
30) Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science **277**: 1453–1462.
31) Seemann T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics **30**: 2068–2069.
32) Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. J Bacteriol **169**: 5429–5433.
33) Eddy SR. (2011) Accelerated Profile HMM Searches. PLoS Comput Biol **7**: e1002195.
34) Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, et al. (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res **44**: D279–285.
35) Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, et al. (2013) TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res **41**: D387–395.
36) Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res **25**: 1043–1055.
37) Endo A, Tanizawa Y, Tanaka N, Maeno S, Kumar H, et al. (2015) Comparative genomics of *Fructobacillus* spp. and *Leuconostoc* spp. reveals niche-specific evolution of *Fructobacillus* spp. BMC Genomics **16**: 1117.
38) Maeno S, Tanizawa Y, Kanesaki Y, Kubota E, Kumar H, et al. (2016) Genomic characterization of a fructophilic bee symbiont *Lactobacillus kunkeei* reveals its niche-specific adaptation. Syst Appl Microbiol **39**: 516–526.
39) Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, et al. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol **57**: 81–91.
40) Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. (2009) DNAPlotter: circular and linear interactive genome visualization. Bioinformatics **25**: 119–120.

Methods for analyzing next-generation sequencing data IX. Genome annotation, visualization, and registration to DDBJ

Yasuhiro Tanizawa[1], Jun Mashima[2], Takatomo Fujisawa[1], Kyungbum Lee[2], Yasukazu Nakamura[1], Kentaro Shimizu[3], and Koji Kadota[3]

[1]Center for Information Biology, National Institute of Genetics.
[2]DDBJ Center, National Institute of Genetics.
[3]Graduate School of Agricultural and Life Sciences, The University of Tokyo.

Genome annotation is a fundamental process in the sequence analysis, through which biological knowledge is generated from sequenced genomic data. Good annotation not only enhances our own downstream analyses but also promotes subsequent researches by others because it can propagate through public sequence databases. In this article, we will show how annotated genomic data are described in the public databases. And then, we will introduce how to use the bacterial annotation pipeline DFAST (DDBJ Fast Annotation and Submission Tool). DFAST is developed to facilitate quick and accurate genome annotation as well as data submission to DDBJ. It was originally designed specific for lactic acid bacteria and now extended to other organism groups. We finally present a method to visualize chromosome maps from DFAST annotation results. Supplementary materials are available at our web site, http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html#about_book_JSLAB.