谷澤　靖洋 助教

TANIZAWA, Yasuhiro

大量遺伝情報研究室

『誰もが正確にアノテーションできるパイプラインDFASTを開発！
-DDBJへの登録フォーマットも自動で作成-』

修士課程までは無機化学を専攻するが、社会人を経験した後に博士後期課程に進んでバイオインフォマティクスを専攻した谷澤靖洋助教。初心者から情報解析に慣れた研究者まで、誰もが平易かつ高精度に微生物をゲノム解析できるパイプラインDFASTを開発した。自らも、乳酸菌、ゼニゴケ、新型コロナウイルスなどの比較ゲノム解析を進めている。

DFASTに関する詳細はこちら

紆余曲折の末、情報学の道へ

大学の学部から大学院の修士課程までは、無機化学を専攻し、触媒の表面構造などについて研究していました。その後10年ほどは実家の事業を手伝い、独学で社内データベースづくりなどに携わりました。ゲームやコンピュータが好きで、修士時代には自らプログラミングして独自の解析ツールを作る、といった経験があったからです。2012年、偶然にも、遺伝研の情報研究系大量遺伝情報研究室教授の中村保一先生が、研究室のテクニカルスタッフを募集していると知り、興味が湧きました。応募してみたところ採用に至り、中村研究室でウェブサイトやデータベース作りに携わることになりました。このことが生物情報学（バイオインフォマティクス）への道を切り開くことになりました。2013年に東京大学新領域創成科学研究科メディカル情報生命専攻の博士後期課程に進み博士号を取得しました。

博士課程では、乳酸菌の全ゲノムの高精度解析をテーマにしました。指導教官は、当時は東大の教授で、現在は遺伝研教授（情報研究系生命ネットワーク研究室）とDDBJセンター長を兼務している有田正規先生です。乳酸菌は非常に多様性に富んだ微生物で、大きく分けても200種ほどあります。私は、それぞれの全ゲノムを解読したうえで比較し、「どのようなグループの乳酸菌にどのような特徴的な遺伝子があるのか」といったことを解析しました。当時もゲノム解析用のツールは色々ありましたが、使い勝手が悪かったり、同じ遺伝子なのに遺伝子として同定されてこないものがあったりと、問題が山積でした。また、解析したゲノムデータはDDBJなどの公共塩基配列データベースに登録する必要があるのですが、そのプロトコールは煩雑で大変苦労しました。

このような理由から、自分が使えるよう、乳酸菌に特化した「扱いやすく、高精度の解析パイプライン」を作ることを目指して開発したのがDDBJ Fast Annotation and Submission Tool（DFAST）というパイプラインです。初期版は、既存のツールに乳酸菌用の参照データ（既知の遺伝子配列や機能など）とDDBJのデータベース登録用ファイルの作成機能を組み合わせたシンプルなものでした。その後、解析できる微生

物種を増やし、独自の要素技術を開発するなど、少しずつ拡張していきました。ユーザー数が増えるにつれ、「初心者であっても簡単にDDBJへのデータ登録ができる」と評判が高まり、DDBJも高く評価してくれました。このような経緯があって、現在ではDDBJの公式ツールとして採用されるまでになりました。

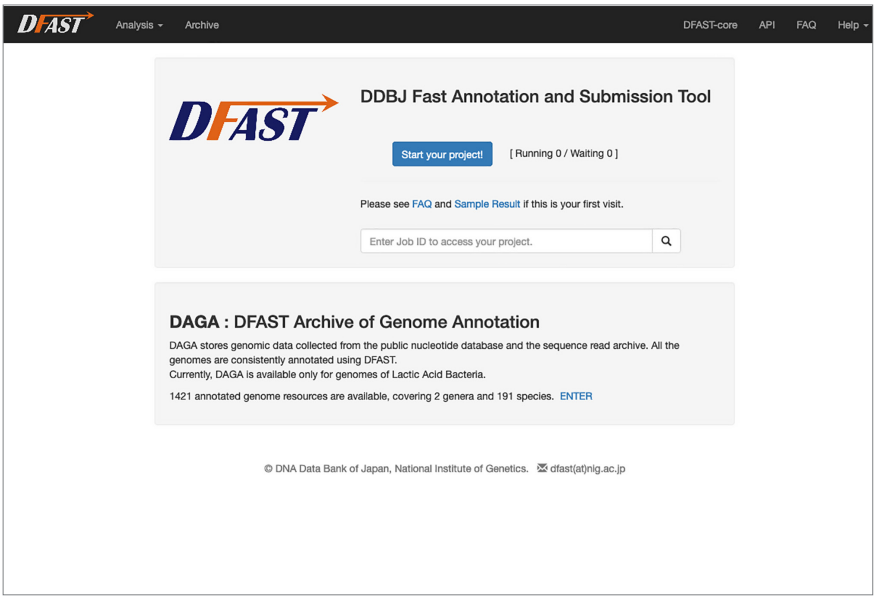
構造と機能のアノテーションを、平易かつ高精度に行えるように

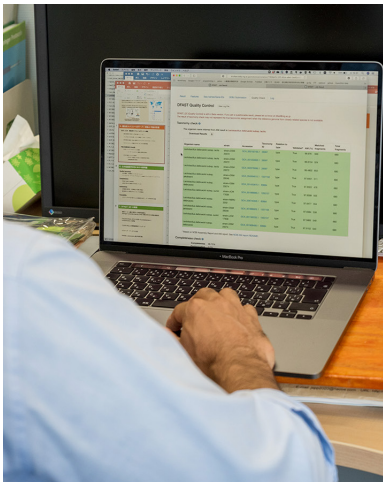
ゲノム解析は、DNAの塩基配列を読むところから始まります。その後、一般的なバクテリアゲノム解析では「解読された塩基配列の断片をつなぎ合わせて全ゲノム配列を再構築する作業(ゲノムアセンブル)」を行います。高速に配列を読める次世代シーケンサーの普及とアセンブル技術の向上により、高精度なゲノム配列が容易に得られるようになりましたが、ここまでの作業で得られるデータは「膨大な文字の羅列」に過ぎません。ここから生物学的な知見を取り出すための作業(アノテーション)も非常に重要です。アノテーションは大きく2つに分けられます。「どこからどこまでの文字列が遺伝子なのかを予測する構造アノテーション」と、「遺伝子と思われる配列が、どのような機能を担っているのかを予測する機能アノテーション」です。

地球上の生物は、細胞の中に核をもつ真核生物と、核の構造をもたない原核生物に分けられます。人間を含む動物や植物は真核生物で、乳酸菌などのバクテリア、高熱菌・メタン菌に代表されるアーキア(古細菌)などは、原核生物です。原核生物はゲノムサイズが小さく遺伝子の構造も単純なため、構造アノテーションについては、かなり高精度に予測できるようになっています。一方、機能アノテーションの基本的な考え方は、DDBJやUniProtKB(タンパク質配列データベースのひとつ)などが公開している既知の配列を参照して「似ているもの」を選び出し、その遺伝子機能を類推するというものです。こちらの方は、参照データにスペルミスがある、「同じ機能をもつ遺伝子なのに、異なる遺伝子名で重複して登録されている」といったノイズやバラツキがあり、得られた結果をそのまま採用することができないという状況でした。そのため、DFASTの開発では、解析結果をそのままDDBJに登録し、公開しても問題ないレベルの「信頼性の高い参照データ」を用意することも課題でした。

はじめに開発した乳酸菌用の参照データは、公開されているデータベースから乳酸菌ゲノムのものを集めて「すでに同定されている遺伝子配列」を取り出したものをソースにしました。これらの中から、似たもの同士を集めてグループ分けを行い(クラスタリングという)、各グループごとに「遺伝子から作られるタンパク質(遺伝子産物)」の名称チェックや、誤って記述されているものの修正を行いました。かなり泥臭い作業で、納得できるものを得るのに2か月あまりかかりました。同様の手法で、当時、大腸菌、ビフィズス菌、ピロリ菌などの参照データも用意しました。さらに、DFASTで解析できる微生物を原核生物全般に拡大しようと考え、研究が進んでいる微生物種から代表的な百数十件のゲノムデータを選び出し、それらに含まれるタンパク質配列も用意しました。広範な種を含んだ参照データを用意すればするほど解析範囲がより広がるのですが、その分、検索にかかる時間や実行に必要な計算資源が増えることになります。この辺りをどうするかは、バランス感覚が必要になります。

培養できない微生物や極限環境に生息するアーキアなど、研究があまり進んでいない生物種を対象にする場合には、十分な解析結果が得られないこともありますが、現状の参照データは、実行時間とサイズのバランスが取れたコンパクトなものになっており、様々な微生物種でほぼ問題なく使っていただけると思います。一方で、公共の配列データベースには、日々、新たな知見が蓄積され続けていますので、より新しい情報を参照データとして取り込んでいくことが、今後の課題になると考えます。ちなみに、病原菌を研究対象にしているユーザー





アセンブル、アノテーションからDDBJ登録まで可能な微生物ゲノム解析パイプライン「DDBJ Fast Annotation and Submission Tool(DFAST)」

谷澤 靖洋 [たにざわ やすひろ]

東京大学理学部化学科卒。東京大学大学院 メディカル情報生命専攻 博士課程修了。博士（科学）。2012年より 国立遺伝学研究所 情報研究系大量遺伝情報研究室に所属。テクニカルスタッフ、特任研究員を経て、2018年より、同研究室助教。

日々の研究活動は全て PC 上でおこなわれる

には、薬剤耐性遺伝子などの情報をDFASTの参照データに付け加えて「より詳細なアノテーション」を行えるように独自にカスタマイズしている方もいるようです。

DFASTのパイプラインは、大きく2つの部分から構成されます。ユーザーがウェブブラウザを通して操作するウェブサービスの部分（ウェブ版）と、その内部で動くアノテーションエンジンの部分（DFAST_core）です。ウェブ版は、初心者であっても迷うことなく直感的に操作ができることを目指して設計しました。特にDDBJへの登録用データを作成する部分では、私自身がDDBJにデータ登録をしたときの苦労した経験をもとに、指示にしたがって空欄を埋めていくだけで自動的に必要なファイルを生成できるようにしました。また、ユーザーが実行する多数のジョブを管理する機能も作ってあります。これには、中村研究室でウェブサービスやデータベースの開発に携わった経験が役立ちました。

もう一つのDFAST_coreの名前には、内部で動くアノテーションエンジンとしての「DFASTのコア部分」との意味を込めています。DFAST_coreは、遺伝研先端ゲノミクス推進センターの野口英樹特任教授が開発した遺伝子領域の予測ツールMetaGeneAnnotatorや、東京工業大学で開発された高速アミノ酸配列検索ツールGHOSTXなどの国産ツールを多く取り入れました。GHOSTXは、広く使われているBLASTというツールとくらべて約10倍も早く計算でき、DFASTの実行時間の短縮に寄与しています。既存ツールや参照データベースを柔軟に組み合わせたDFAST_coreは、ユーザー自身でカスタマイズしやすく、それもDFASTの特徴のひとつといえます。ソースコードはGitHubというレポジトリで公開しているのですが、ユーザーから「×××というソフトをDFASTで使えるように改良したから採用して欲しい」といったリクエストが来るほどです。

微生物ゲノムアノテーションパイプラインDFASTとして一般公開

2016年に「微生物ゲノムアノテーションパイプラインDFAST」として最初の論文を発表し（*1）、ウェブ版を公開しました。初期のものは、乳酸菌を解析の対象にしたものです。その後、独自のアノテーションエンジンであるDFAST_coreを開発し、対象を原核生物全般に拡張することで現在のDFASTの原型ができあがりました（*2）。当初は自前のサーバーで運用していたのですが、2021年に遺伝研のスーパーコンピュータ上にサービスを移設しました。解析パワーが格段に改善され、数百万塩基対程度の典型的なサイズのバクテリアゲノムであれば2〜3分で結果が得られるようになりました。ゲノム塩基配列のファイルをアップロードするだけで実行でき、DFASTという名前の通り迅速に結果を得ることができますので、初心者にもおすすめです。DFAST以外の様々なパイプラインも開発されていますが、一連の解析の入り口として「何が出てくるかな？」という感じで、まずはDFASTを試していただくのが良いと思います。

情報解析に慣れている研究者向けには、DFAST_coreをコマンド操作で直接実行できるよう公開しています。こちらは、パイプラインを柔軟にカスタマイズできるという点に加え、比較的大きな解析プロジェクトでデータを大量に扱う際に、コマンドでデータを一括して高速処理できるメリットもあります。DDBJへの登録用ファイルを自動で作ることができる点は、ウェブ版と共通です。

登録フォーマットの自動作成と高速動作がアピールし、すでに8万件の利用

ありがたいことに、国内外で多くの方に使っていただき、私のDFASTの論文は、近年（2016〜2021年）発表された遺伝研発の論文の中で、分野別の被引用数がトップになっていると聞いていま

す。ウェブ版ではこれまでに約8万件の利用があり、現在の利用件数は月に2000件ほどです。ユーザーの3~4割は海外の研究者で、国際的なゲノム解読プロジェクトであるThe Global Catalogue of Microorganisms (GCM) 10K type strain sequencing projectでも使っていただいています。このプロジェクトは、各国の生物資源保存機関に寄託されている微生物1万種の基準株(種分類において基準として用いられる菌株)のゲノムを決定することを目指したもので、日本からは理化学研究所の微生物材料開発室や製品評価技術基盤機構 (NITE)のNBRCが参画しています。初期の成果はすでに論文になっていますが、計5年のプロジェクトが現在も進行中で、決定されたゲノムの一部はDDBJを通して公開されています。

大規模なゲノム解析研究での利用事例には、私たちが九州大学と共同研究で行った大腸菌ゲノムの比較解析があります。この研究では詳細な遺伝子機能情報を得るために大腸菌用に参照遺伝子データを用意し、DFASTを用いてウシの腸管から分離された約800件の大腸菌ゲノムをアノテーションしました。その結果、ヒトに対して強い病原性を示す腸管出血性大腸菌がウシの常在性大腸菌を起源としていること、ウシ腸内で生存するために特定の病原因子を蓄積させていること、などがわかりました。私自身は、DFASTの開発を行う過程で、ヨーグルトで有名な乳酸菌のひとつであるガセリ菌が大きく2つのグループに分かれることを突き止め、新種を*Lactobacillus paragasseri*と提唱しました(既存グループは*Lactobacillus gasseri*)。通常、新種の提唱は、微生物研究者が屋外調査を行って分離した菌株を元に行うことが多いのですが、この研究は「データベースから得られる公開データを用いた情報解析を主体にした」という点がユニークだと思っています。当時、総合研究大学院大学(総研大)遺伝学専攻の学生だった多田一風太さんと共に進めて得た成果です(*3)。

繰り返しになりますが、DFASTのアピールポイントはなんといってもアノテーションとともにDDBJへの登録フォーマットを簡単に作成できる点にあります。米国の国際的な塩基配列データベースGenBankへの登録にはPGAPというパイプラインがありますが、DDBJへの登録については標準的なツールがありませんでしたので、DFASTによってデータ登録の負担を大きく軽減できたのではないかと自負しています。また、DFASTを使ってアノテーションされた結果は一定の品質基準を満たしているため、登録の際に内容のチェックを行うDDBJ側の査定者にとっても、負担軽減につながると思います。実際、2020年の1年間にDDBJに登録されたアノテーション付きのバクテリアゲノムの9割でDFASTが用いられるまでにシェアが広がっています。

現在、多くの学術誌において、新規決定の塩基配列に基づいた議論を行う論文を掲載する際には、データを公共塩基配列データベースに登録することが求められています。私は、このことは、単なる義務以上に研究者にとって意義があると考えています。論文の信頼度や再現性を担保することに加え、自身のデータが検索対象となって参照される、第三者によって再利用されることで自分の研究が他者の研究にも影響を及ぼす、といった効果があるからです。その意味で、データ登録は研究発表の一環と捉えることもできると考えています。

さらなる改良や新型コロナウイルスの変異解析も

昨年度は、DFASTに「ゲノムの生物種名や完成度をチェックする機能(DFAST_QC)」を追加しました。DDBJをはじめ各塩基配列データベースは、不適切なデータが登録されないための査定を行ってはいますが、誤った生物種名が記載されたデータや、低品質のゲノムデータが見逃されて登録されてしまうことがあります。新機能のDFAST_QCによって、この問題を解消することを目指しました。

さらに、DFASTとDDBJのデータ登録システムとの統合化も進めています。これまでは、DFASTを使ってアノテーションした後はDDBJの登録担当者とのメールでのやりとりによって登録手続きを行なっていました。統合化すれば、結果をワンクリックでDDBJに送信し、データ登録処理を行えるようになると思います。ゆくゆくは、今のところは人の手を介してやっているデータ査定作業も自動化され、アノテーションからデータ公開までワンストップで行えるようになれば良いと考えています。

一方、遺伝研として静岡県と連携し、新型コロナウイルスの変異株ゲノムの解析も始めています。静岡県内で採取された検体を対象に全ゲノムを読み、変異部位の同定やそのアノテーションを行い、データを順次DDBJに登録して公開するというものです。このなかで私はDDBJに登録を行う部分を担当し、すでに既存のパイプラインを応用してコロナウイルス用のDFASTを作成済みです。遺伝研には、ゲノム解読を行うシークエンス拠点、大規模解析が可能なスーパーコンピュータ、データ登録窓口であるDDBJが揃っていますので、迅速で正確なデータ公開を行うことで、感染経路の追跡や新規変異株の同定に貢献できると考えています。

バイオインフォマティシャンは、多様なバックグラウンドのるつぼ

研究室には、私以外に研究員が4人います。コロナ禍の前からラボではオンラインのチャットツールを使った情報共有が行われていましたので、パソコンさえあれば自宅でのテレワークには全く不便を感じません。各自が自分のプロジェクトを進めていますが、プロジェクトごとのチャットが常時オンライン状態になっているので、そこで進捗報告や相談、議論などをしながら進めています。

ボスの中村先生には自由にやらせていただいており、大変感謝しています。研究室として他の研究者のゲノム解析支援を目的とするプロジェクト(文部科学省新学術領域「先進ゲノム支援」)も進めており、ゼニゴケやケイソウなどの比較ゲノム解析を中村先生と共同で行なっています。

私は社会人を経験してから博士後期課程に進み、バイオインフォマティクスの研究者になったわけですが、自身の経験から「進路の変更はいつからでも可能」と申し上げたいです。求める者に対しては、門戸はいつでも開いていると思っています。特に、バイオインフォマティクスの研究者は、もとのバックグラウンドが非常に多様です。もちろん、もともと情報学を専門にしていた研究者もいますが、「実験をやっていたが、必要があって情報解析するようになり、そのまま情報学を専門にするようになった」というような人もいます。中村研究室のメンバーのバックグラウンドもバラバラです。

このような状況には、ゲノム解析用のツールが様々出てきたことで、以前よりも情報解析の垣根が低くなったことも寄与しているかもしれません。ツールの使い方や使い勝手について情報発信する研究者も、かなり増えています。私がそうだったのですが、生物学に詳しくないままバイオインフォマティクスを専門にするには、必要に応じて、共同研究者から一つずつ学んでいくことも重要だと思います。

聞き手:サイエンスライター　西村 尚子
写真撮影:遺伝研ORD　来栖 光彦
デザイン:遺伝研ORD　金澤 奈穂

2021年7月

引用論文等

*1----------
DFAST and DAGA: web-based integrated genome annotation tools and resources
Tanizawa, et al. 2016 BMFH.

*2----------
DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication
Tanizawa, et al. 2018 Bioinformatics.

*3----------
Revealing the genomic differences between two subgroups in *Lactobacillus gasseri*
Tada et al. 2017 BMFH.

Lactobacillus paragasseri sp. nov., a sister taxon of *Lactobacillus gasseri*, based on whole-genome sequence analyses
Tanizawa et al. 2018 IJSEM

その他関連リンク

- 生命情報・DDBJセンター
- 先端ゲノミクス推進センター
- 総合研究大学院大学(総研大)遺伝学専攻
- 文部科学省新学術領域「先進ゲノム支援」