「塩基配列アーカイブのデータベース構築と統合への貢献」

２０年度科学技術試験研究委託事業
成果報告書

平成２１年３月

大学共同利用機関法人
情報・システム研究機構

五條掘　孝

本報告書は、文部科学省の科学技術試験研究委託事業による委託業務として、大学共同利用機関法人情報・システム研究機構が実施した平成２０年度「塩基配列アーカイブのデータベース構築と統合への貢献」の成果を取りまとめたものです。

１．委託業務の目的

各種生物の遺伝子やゲノムの塩基配列を決定するいわゆるシーケンシングセンターにおいてはその配列決定の原データになる波形データなどのいわゆるアーカイブ(Trace archive; 以下「Trace データ」という。)を有している。この Trace データは、品質管理やそれを基にして行う配列決定アルゴリズムの改良ならびに配列断片を連結するアセンブリにおいて大変重要で貴重な情報である。そして、これらの Trace データは、シーケンシングセンターの活動がその支持母体のプロジェクトの完了とともに終了すると、原則的には完全に消滅してしまう可能性が極めて高い状況にある。また、454 や Solexa 又は ABI-SOLiD といった次世代の超高速の塩基配列決定装置の登場により、その Trace データの量は飛躍的に巨大化してきており、シーケンシングセンター自身でもそのデータハンドリングを含めて保存はもちろんのこと対処が非常に困難な状況になっている。

この状況の理解の下、わが国における塩基配列決定における Trace データの保存と有効利用を目的として、当機関である情報・システム研究機構国立遺伝学研究所生命情報・ＤＤＢＪ研究センターのＤＤＢＪが、Trace データのデータベース構築とデータ提供の業務を実施し、情報システム研究機構ライフサイエンス統合データベースセンターへのデータ提供や連携を可能にする。

具体的には、同機関の同研究所の生物遺伝資源情報総合センターアカデミアＤＮＡシーケンシングセンターが保有する Trace データを当初の対象データとして、データベース構築を行なうことを目標とする。さらに、先行して Trace データを蓄積してきた米国 NCBI 及び欧州 EBI における Ensembl とはデータを共有して、ＤＤＢＪからの国際貢献とするとともに、ライフサイエンス統合データベースセンターの成果に寄与するものである。

そして最も重要なことに、このデータベースを統合データベースセンターに提供するに際し、統合データベースが必要とする基本的なアノテーションの実施やソフトウエアの開発を連携して行うものである。

２．平成２０年度（報告年度）の実施内容

２.１ 実施計画

統合データベースにおける情報資源のひとつとして利用できるような基礎データアーカイブ作成を目指して、昨年度に引き続き以下のような計画で本年度の作業を進める。

① 公開 FTP サイトと WWW サイトに関する開発

本年度は、情報・システム研究機構国立遺伝学研究所の生物遺伝資源情報総合センターアカデミア DNA シーケンシングセンターだけでなく、東京大学、独立行政法人理化学研究所、財団法人かずさディー・エヌ・エー研究所、独立行政法人製品評価技術基盤機構、独立行政法人農業生物資源研究所などの塩基配列シーケンシングセンターである国内主要機関を中核とするコミュニティーとの連携を深めながら、Trace データの実データの登録開始を目標にして、昨年度調査を進めた必要となる基盤技術（大規模 Trace データのデータベース化に関わるデータ登録・アーカイブ処理・データ提供に関する技術）の整備を行い、上記の基盤技術を用いて、大規模データ転送（数十ＧＢから数百ＧＢ）に必要なＦＴＰサーバの構築と、個別データの閲覧、波形データの表示、統計処理などを行うためのWWWシステムの開発を進め、データの受け入れを試験する。

② 登録処理及び波形表示システムに関する開発

平成１９年度に引き続き、登録処理及び波形表示システムのプロトタイプの構築を進め、情報・システム研究機構国立遺伝学研究所の生物遺伝資源情報総合センターアカデミア DNA シーケンシングセンターの Trace データを対象としたデータベース構築を進める。

２.２ 実施内容(成果)

２０年度は、当初予定に従い、

１．トレースデータ登録処理システム、波形表示システム開発

２．トレースデータ用 FTP サーバ構築。キーワード検索、ダウンロード、統計情報閲覧サイト作成

以上の２点について活動を行った。具体的には、

プロトタイプ開発をもとに、DB 構築、DB へのデータ登録システム、Web 検索システム、 波形

表示システムの開発（図１，２）を進め、並行してデータベース開発としてデータベースのスキーマを仮に構築し、実データを投入して検証を進めた。この検証に基づき、

１に関しては、今後多様なサービスの拡充を図るに当たり、Trace データを格納し、様々な検索を可能にするためのデータベースシステムと、そのデータベースにデータを投入するためのデータ登録システム（図１の①-②）との開発を行い、ＩＤ発行（③）とアーカイブデータベースの作成（④）、および外部公開データベース作成（⑤）を進めた（図１）。Trace Archive の代表的運営母体である NCBI と協調し、統一された ID の発行を可能にするため、メタデータの情報、アーカイブ形式は NCBI のガイドラインに準拠した内容を採用し、DDBJ が NCBI で公開されている RFC に準拠した内容や形式で登録データを準備できるよう、データ登録ユーザに対するコンサルティングやデータ作成の補助等の支援を行い、また後段階での処理エラーを防ぐため、NCBI にデータを送付する前に、事前のデータチェックを行う等の処理も行えるようにした。

また２に関しては、web 検索システムとして、検索・結果表示・個別データダウンロード機能は実装し、データベース構築と共に性能評価やテストを元に開発を継続した。波形表示システムに関しては、既存プログラムをベースに修正を進め、より良い表示方法がないか引き続き検討した。一方、統計情報表示ページについても表示項目の詳細の検討をすすめた。その結果、FTP サーバ構築を終了し、公開用ＦＴＰサイトの整備とｗｗｗサイトを公開するとともに(図２)、今年度８月には ddbj に登録された 2 件の ftp 公開を開始し、同時に試験的に受付も開始した。登録処理システムについては、データチェック、DB 投入部分について、プロトタイプの検証を進めた。機能としては、以下のように整備を進めた。
公開システムは、ddbjが保持するトレースデータを、WWW によって利用者に公開するものである。公開ステータスにあるデータ全件を対象とし、キーワード検索機能、個別データダウンロード機能、波形表示システム等を有するものとする。以下に各機能の詳細を示す。

キーワード検索機能
指定のメタ情報で、登録されたデータを網羅的に検索し、ヒットしたデータを web ページで閲覧できるようにした。

個別データダウンロード機能
キーワード検索機能でマッチしたデータを一括でダウンロードできる機能を実装した。

波形データ表示
指定したデータの波形情報をグラフィカルに表示し、ATGC 各塩基をあらわすパターンは別の色で表現するなどして判別が容易なようにした。

統計情報表示
その時点で ddbj 経由で登録されているデータについての統計情報を記載し、データの公開や更新が発生したら速やかに更新できるようにした。

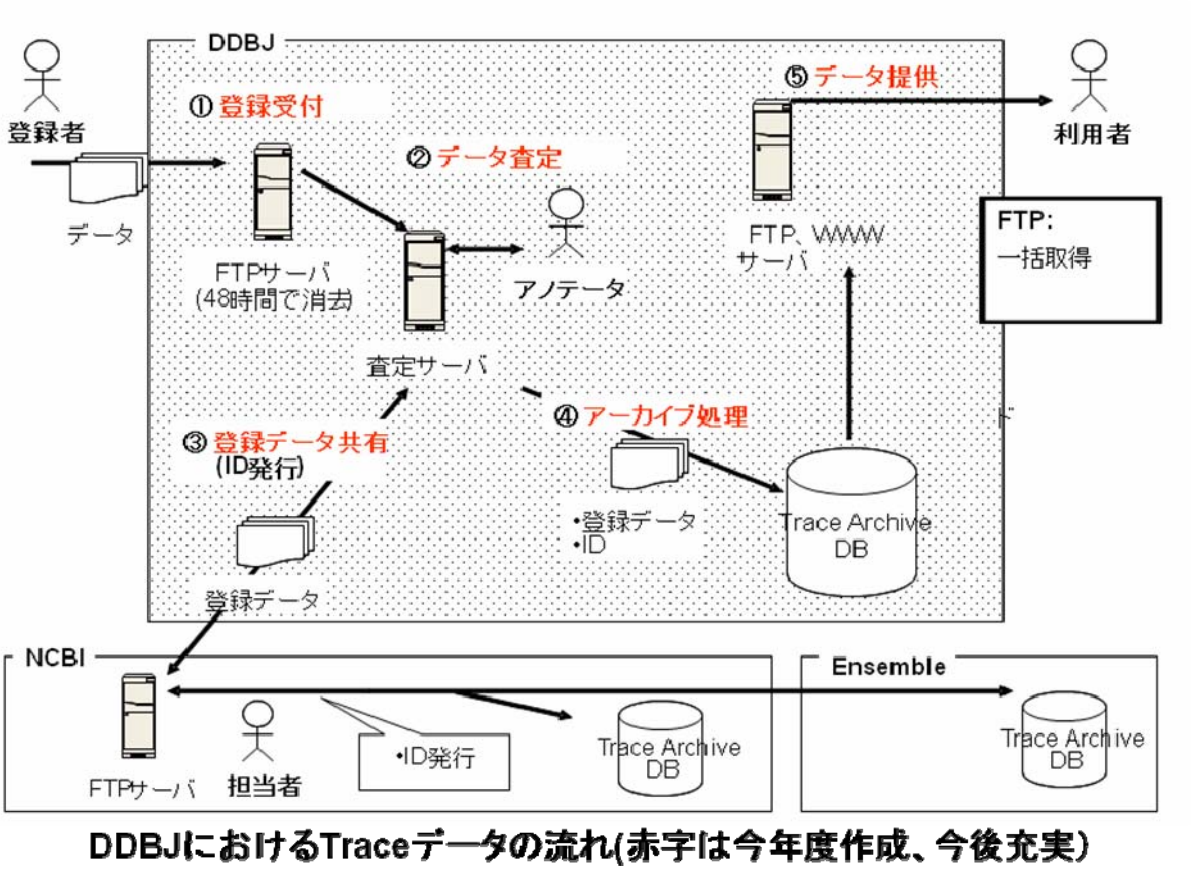


図１トレースアーカイブデータベースシステム概要

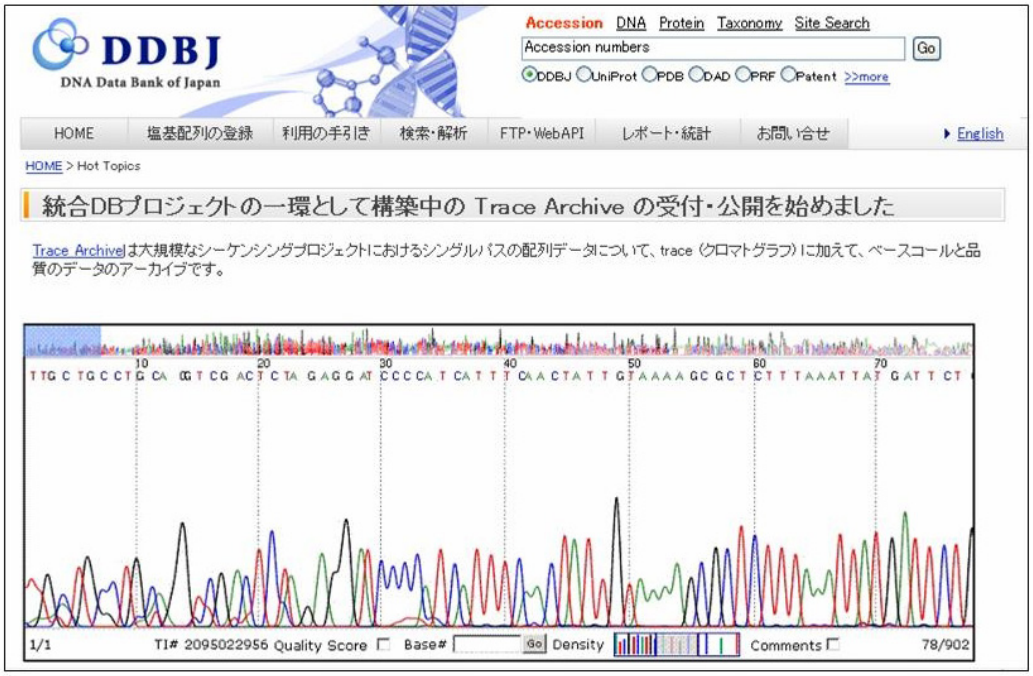


図２データ提供サイト

２．３ 成果の外部への発表

１．トレースデータベースのサービスの開始に関して、統合データベースセンターと共同でアナウンス（日経ＪＴＢ０８年１０月）を行うとともに、ＤＤＢＪのホームページを通じてアナウンスした。

２．学会等における口頭・ポスター発表

発表した成果（発表題目、口頭・ポスター発表の別）	発表者氏名	発表した場所（学会等名）	発表した時期	国内・外の別
Development of a database of complex disease by using it. 口頭発表	舘野義男	HGM2008 Satellite Meeting、カルカッタ、」インド	2008 年 9 月 25 日	国外
Construction of a tree of life、ポスター	舘野義男	国際遺伝大会、ベルリン、ドイツ	2008 年 7 月 15 日	国外
Genome and Database	五條堀　孝	The 18th CODATA-DSAO Task Group Conference、三島	2008 年 8 月 31 日	国内
DNA シーケンス革命とスーパーコンピューティング	五條堀　孝	生命体統合シミュレーション研究開発プロジェクトシンポジウム、東京	2008 年 12 月 26 日	国内

３．学会誌・雑誌等における論文掲載

掲載した論文（発表題目）	発表者氏名	発表した場所（学会誌・雑誌等名）	発表した時期	国内・外の別
DDBJ dealing with mass data produced by the second generation sequencer.	Sugawara, H., Ikeo, K., Fukuchi, S., Gojobori, T., Tateno, Y.	Nucleic Acids Research	2008 年 10 月 16 日	国外

２．４ 活動（運営委員会等の活動等）

２．５ 実施体制

別表１のとおり

２．６整備実績一覧

別表２のとおり

別表１　平成 20 年度に於ける実施体制

研　究　項　目	担当機関等	研究担当者
業務主任者 全体構成及び総括	国立遺伝学研究所 生命情報・DDBJ センター	◎○五條堀　孝
提供サービス内容	国立遺伝学研究所 集団遺伝研究系集団遺伝研究部門	斎藤成也
データ登録・ハードウエア	国立遺伝学研究所 生命情報・DDBJ センター	池尾一穂
データ産生サイトからの登録受付を担当する	国立遺伝学研究所 生命情報・DDBJ センター	舘野義男

注１．◎：課題代表者、○：サブテーマ代表者
注２.本業務に携わっている方はすべて記入。

【参考情報】

●公開から年度末までのアーカイブ（登録）されたデータの概要（＋規模）等

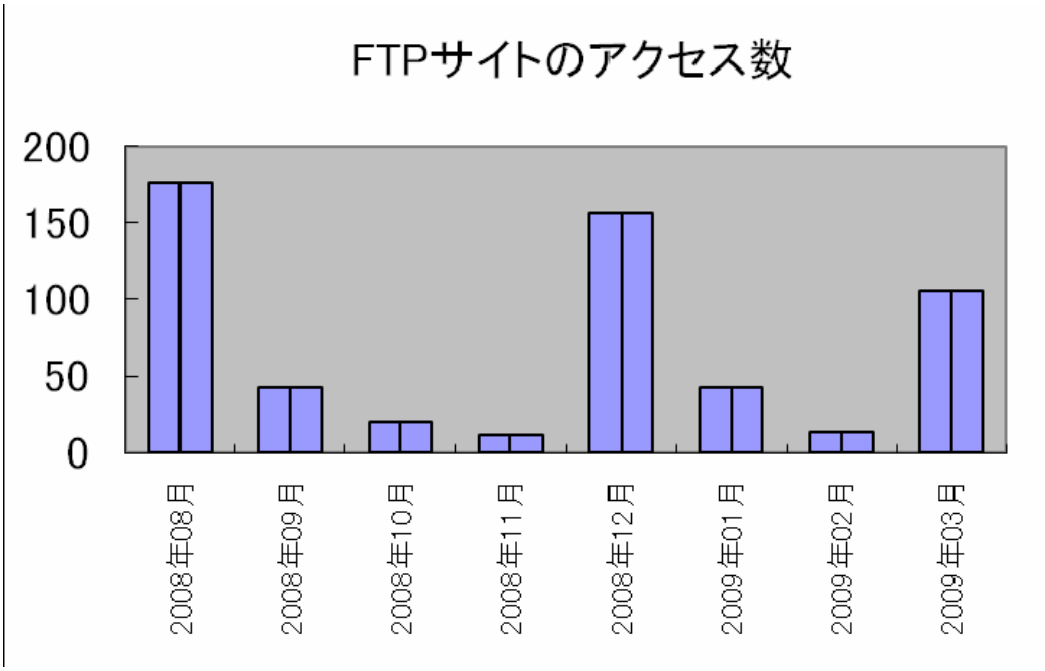
１．国立遺伝学研究所（NIG）が決定した Oryzias latipes（メダカ）の WGS データに対応するトレースデータ；1,481,568 件

２．東京大学(UTCOB)が決定したヒト腸内微生物群の WGS データに対応するトレースデータ；1060139 件

●公開サイトの利用状況（アクセス推移等）

・FTP サイトのアクセス数（総ファイル・ページ取得数、ロボット等を含む）

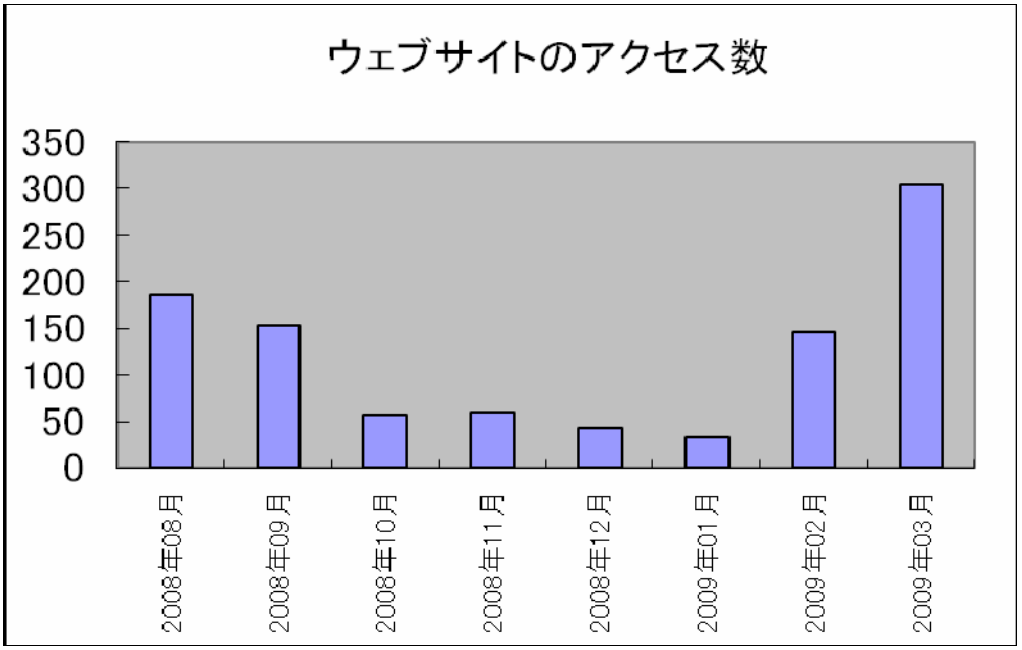
取得月	アクセス数
2008/8/1	176
2008/9/1	42
2008/10/1	20
2008/11/1	11
2008/12/1	157
2009/1/1	42
2009/2/1	13
2009/3/1	106



・WEB サイトのアクセス数

（総ページ閲覧数。ロボット除く。genes・ddbj ドメインからのアクセス除く）

取得月	アクセス数
2008/8/1	185
2008/9/1	153
2008/10/1	57
2008/11/1	60
2008/12/1	42
2009/1/1	32
2009/2/1	146
2009/3/1	305



別表２

報告日：平成２１年　５月　１日

整備実績一覧【代表機関名：国立遺伝学研究所】

（１）保有データ情報

※　貴機関グループ内で保有するデータに関して、以下の内容を記述して下さい。

（１－１）データの種類

①生物種	多様
②試料・ライブラリー等の種類、数	多様 （種に関して制限をしていない。広く集めて行くことを目指している。）
③測定方法	従来型シーケンサーからの出力。
④データの内容	塩基配列決定時の波形データ、結果の配列データ、その生物種、目的分類（WGS、ゲノム、等）、他 ）
⑤その他、特記事項	

（１－２）データソース

①現在のデータ量	2 生物種、2,542,476 件
②データ区分	□自前　■第三者　□文献データ　□計算結果等の二次データ □その他（下欄に詳細を記述）　　※複数選択可。二次データのみの保有は不可。
③将来の増加の見込み	サービス開始後、急速に増加の見通し。次世代シーケンサの普及により、そこからのデータが急速に増加していく見込み。
④権利関係	所有者（それぞれのデータ登録者に帰属） 公開（■可　□否　□その他［　　］）
⑤その他、特記事項	http://trace.ddbj.nig.ac.jp/Trace/search

（１－３）データの管理状況

①更新頻度等の管理状況、体制	登録要求に応じて随時更新。
②その他、特記事項	

※　既にデータベースを保有している場合は、以下についても記述して下さい。

（１－４）データベース関係

①DB 管理者数	２名
②キュレータ・アノテータ数	１名
③データ構造	RDB
④DB 管理ソフト	HiRDB
⑤サーバの OS	RedHat Linux
⑥サーバ規模	CPU: Xeon 5160 3.00GHz, Memory 8GB, HDD 約 4TB
⑦DB へのアクセス数	公開されて間もないため、未計測
⑧独立 IP 数	公開されて間もないため、未計測
⑨その他、特記事項	※DB への検索メニュー（画面コピーの別紙添付でも可）、使用しているオントロジーがあれば記述

（２）データ（又はＤＢ）の連結、統合化整備（※試験的、限定的公開済みのものも含む。）

通番	データ（又はＤＢ）の名称 ※URL があれば記述	公開／未公開	概要（データの種類（生物種）・数量（kB 等）、本プロジェクトで実施した特徴点、進捗状況、今後の計画・課題などを簡潔にわかりやすく記述） ※　公開している場合は、開始年月、利用状況（平均利用者数、アクセス数、ダウンロード数等の数値的指標で記述） ※　必要に応じて画面コピー等の図表添付可
1	トレースデータベース http://trace.ddbj.nig.ac.jp/Trace/search	公開	塩基配列決定の 1 次データとして産生されるトレースデータのデータベース。ウェブインタフェースの検索サービス。開始年月は 2009 年 4 月からのため、利用状況については未計測。
2	トレース付随データ ftp://ftp.ddbj.nig.ac.jp/ddbj_database/trace	公開	FTP による公開中。統合のためのメタデータについては中核機関と内容をすりあわせ中につき、未完了。

（３）ＤＢ基盤システム、ツール等開発成果物の整備（※試験的、限定的公開済みのものも含む。）

通番	ＤＢ基盤システム、ツール等の名称	公開／未公開	概要（主な機能・特徴点、進捗状況、今後の計画などを簡潔にわかりやすく記述） ※　プログラムプロダクトに限らず、データ形式共通化、標準化のための仕様書、共通規約等のドキュメントについてもリリースしているものは対象とする。 （リリース済みドキュメントは参考として目次一覧、抜粋を添付） ※　必要に応じて画面コピー等の図表添付可
1	トレース登録データチェック・ＤＢ投入処理システム	未公開	トレースアーカイブサービスへのデータ登録に際して、データの文法的・意味的なチェックを行い、ＤＢに投入するためのツール。現在内部で運用中。
2	波形表示プログラム	公開	トレースアーカイブサービスへの登録データについて、ウェブサービス上で波形表示を行うためのシステム。現在ウェブシステムの一部として公開中。

（４）その他の成果物（（２）、（３）に該当しないもの）

通番	名称	公 開 ／ 未公開	概要 ※　必要に応じて画面コピー等の図表添付可