平成１９年度
委託業務成果報告書

「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」

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

目　　次

「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」

頁

業務題目：「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」

担当者：（所　属）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

（氏　名）五條堀　孝

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事業とデータ提供の事業を実施する。

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。

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に提供し、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１９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

平成２０年度以降の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、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）の整備を行う。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年度に登録処理及び波形表示システムのプロトタイプを構築し、このプロトタイプをもとにして情報・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のための基礎とする。

２．２．実施内容（成果）

①公開 FTP サイトと WWW サイトに関する開発

統合データベースへの Trace データ統合のために必要となる機能に関して、独立行政法人理化学研究所、東京大学、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、シーケンサーによる塩基配列決定の際、波形情報などの一次情報として得られる Trace データの実データの登録開始を目標にして、必要となる基盤技術（大規模トレース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）の整備を進めることを目的として、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た。現在、データの転送（長時間が必要）を半自律化（エラー時再試行、完了後レポート送付、など）する仕組みはプロトタイプが完成し、試験中である。

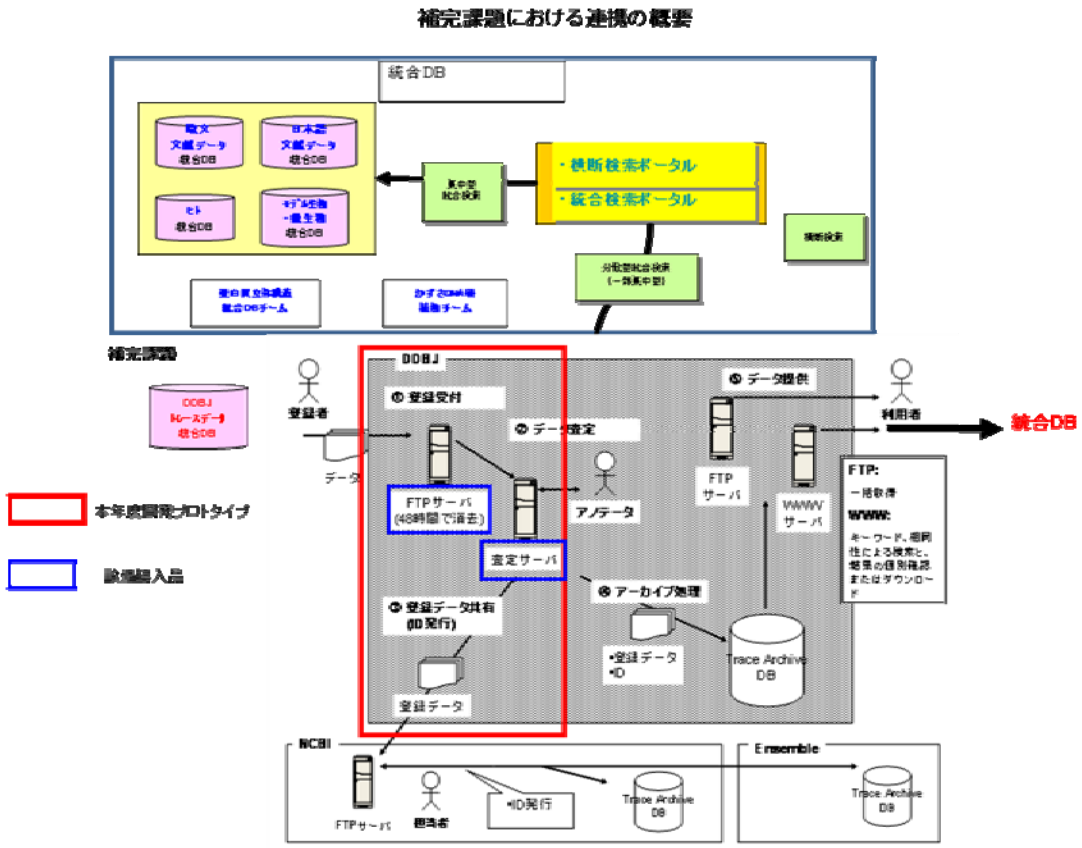

図１　統合との連携における DDBJ における Trace データの流れ

② ゲノムネットワーク利用システムの開発

統合データベースからの Trace データの利用実現のために、平成１９年度にプロトタイプ（NCBI / Ensembl の構成に基本的には準拠しながら、ＤＤＢＪ特有のデータベース管理システムやパイプ

ラインの効率化に合わせて、より国内の状況に応じたシステム）を構築し、このプロトタイプを情報・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 データを対象として適応し、データベースの構築を開始した。（１）登録データの受け付けから、受付時の査定、NCBI へのデータ提供、本年度は、・登録データの機械的チェックのシステムや、トレースデータの形式の相互変換（AB1 - SCF）など、最低必要な部分のシステムは一通りプロトタイプを完成した。（２）現在、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より、最初のデータをテスト用に受領し、システムおよび運用の評価として、実際の処理を行っている。

また、データベース構築については、前述の通り、現段階では検索キーなどデータの論理設計の段階にある。

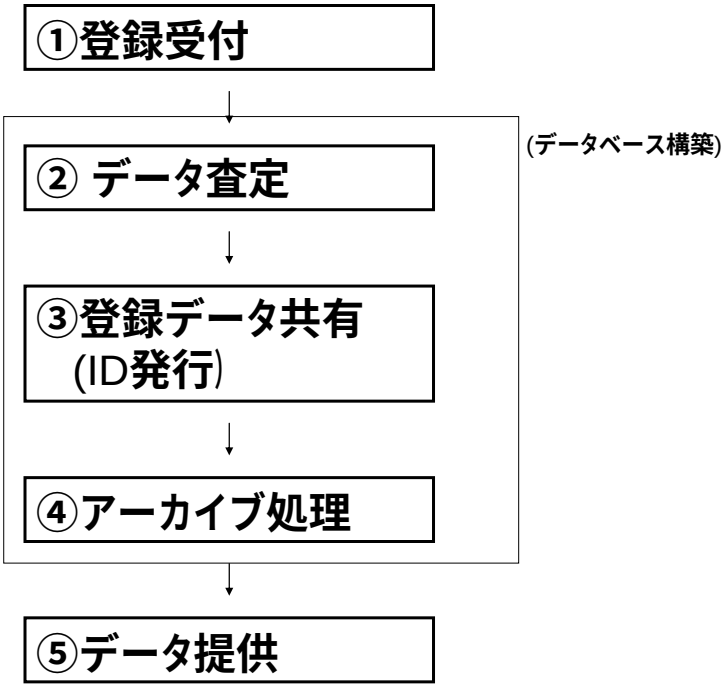

図２ データ受付のための業務フロー

２.３ 成果の外部への発表

論文寄稿

07.10.25(Epub)

DDBJ with New System and Face (Tateno, Y., Sugawara, H., Ogasawara, O., Okubo, K. and Gojobori, T.) Nucleic Acids Res. 36(Database issue), 22-24

２.４ 活動（運営委員会等の活動等）

２００８年３月　国立遺伝学研究所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へのヒアリング

２００８年３月　東京大学ヒトメタゲノムデータの調査

２.５ 実施体制

別表１の通り。

別表１　平成１８年度に於ける実施体制

研　究　項　目	担当機関等	研究担当者
全体構成及び総括	国立遺伝学研究所 生命情報・ＤＤＢＪ研究センター	◎○五條堀　孝
データ登録	国立遺伝学研究所 生命情報・ＤＤＢＪ研究センター	菅原秀明
提供サービス内容	国立遺伝学研究所 生命情報・ＤＤＢＪ研究センター 国立遺伝学研究所集団遺伝研究系	舘野義男 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研究センター	池尾一穂