特集

国立遺伝学研究所における DNAデータバンク：DDBJ

Introduction of the DNA Data Bank of Japan (DDBJ)

あらまし

日本DNAデータバンク（DDBJ：DNA Data Bank of Japan）は，1986年に国立遺伝学研究所の遺伝情報研究センター内に設立され，DNA配列の登録，研究活動を始めた。DDBJは活動開始時から米国のGenBank，欧州のEMBLとの3者で協力しながら，今日の国際塩基配列データベースを構築してきた。

1984年設置された遺伝情報研究センターは1995年に生命情報研究センターとなり，2001年4月には生命情報・DDBJ研究センターと改称され，DDBJの名前が公的センターの名前として取り入れられた。これは日本を代表する国際的な公的データバンクとしての位置付が明確にされたものであり，今後のDDBJのデータバンクとしての活動が日本のバイオインフォマティクスの飛躍的発展のかぎを握っていると言える。

また，国立遺伝学研究所のスーパーコンピュータシステムは国の共同利用機関に指定されており，VPP5000をはじめとするHPCシステムの利用を所外に公開するとともに，DDBJのDNAデータ登録，解析，検索サービスを全世界に向けて提供している。富士通は，このスーパーコンピュータシステムの導入から構築，運用までの業務を支援しており，また，DDBJのサービスの開発運用についてSE支援を行っている。

本稿ではこの国立遺伝学研究所のDDBJとサービスの概要と今後の展開について紹介する。

Abstract

The DNA Data Bank of Japan (DDBJ) was established at the Genetic Code Research Center of the National Institute of Genetics in 1986 to promote DNA sequence registration and research activities. The DDBJ, together with the GenBank of the US and the European EMBL, forms the international base sequence database. The Genetic Code Research Center was established in 1984 and was changed to the Center for Information Biology in 1995. Then, in April 2001, the center was changed again to The Center for Information Biology and DNA Data Bank of Japan (CIB/DDBJ). The DDBJ is a public research center and is the main international and public data bank in Japan. Because of its primary position, the DDBJ will play a very important role in the future development of bioinformatics in Japan. The supercomputer system at the National Institute of Genetics is designated as a cooperative facility of Japan. Its HPC system, which includes a VPP5000, is open to external researchers, and the DNA data registration, analysis, and search services of the DDBJ are made available worldwide. Fujitsu has helped the DDBJ to introduce, construct, and operate this supercomputer system and has supplied system engineers to develop and operate the DDBJ services. This paper introduces the DDBJ system and services of the National Institute of Genetics and describes its future development.

山口政仁（やまぐち　まさひと）

ライフサイエンス推進室 所属
現在，国立遺伝学研究所のスーパーコンピュータシステム，DDBJ運用業務のサポートに従事。

国立遺伝学研究所におけるDNAデータバンク：DDBJ

ま え が き

国際プロジェクトとして推進されてきた人の全塩基配列を決めるヒトゲノムプロジェクトは，30億塩基から成るヒトゲノムのドラフト配列を2001年2月に公表した。この配列情報はDDBJ/EMBL/GenBankの国際塩基配列データベースに登録された。このヒトゲノム配列の公開を契機にバイオインフォマティクス研究は，ゲノム上の遺伝子の決定や，遺伝子から創生される蛋白質の機能解析，ヒトの遺伝子の違いを調べる多型解析などのポストゲノムシーケンシングに研究はシフトしてきたと言える。遺伝子情報の原点は配列情報であり，現在DNAデータバンクには1,700万エントリ，200億塩基対以上の配列情報が登録されており，年率1.5から2倍の膨大な勢いで増加している。この配列情報を登録・維持し，全世界の研究者に公開していく重大な任務を背負っているのが国立遺伝学研究所生命情報・DDBJ研究センターである。

本稿では，このDDBJの概要，サービスの概要，今後の取組みについて紹介する。[(1)-(3)]

DDBJの概要

DDBJとはDNA Data Bank of Japanの略称であり，欧州のEBI/EMBLおよび米国のNCBI/GenBankとともに密接な連携の下「国際塩基配列データベース」を構築している三大国際DNAデータバンクの一つで，静岡県三島市にある国立遺伝学研究所の生命情報・DDBJ研究センター（Center for Information Biology and DNA Data Bank of Japan：CIB-DDBJ）が運営している。

DDBJの主な活動は以下のとおりである。

・国際塩基配列データベースの共同構築と運用
・様々な生命情報データベースの運営
・塩基配列データベースのオンライン利用の管理・運営
・解析ソフトウェアの開発
・印刷物発行や講習会開催などの利用者サービス
・国立遺伝学研究所コンピュータシステムの管理・運営

CIB-DDBJは，発展しつつある生命情報学の日本における研究拠点として，1995年4月に設立された生命情報研究センター（CIB）を，2001年4月に改組してできたセンターである。今日の生物学の研究では，コンピュータが実験器具と同等に必要になってきており，とくに，大変なスピードで蓄積しつつあるDNA配列データなどの生命情報の処理や解析にはコンピュータ科学・技術が必須であり，これが生命情報学（バイオインフォマティクス）を誕生させ，発展させている原動力となっている。

CIB-DDBJは，コンピュータを使った生命情報の解析・処理のための環境と人材を整備しつつ，国内外における生命情報学の分野で主要な役割を果たすことを目的とした活動を行っている組織である。DDBJはCIBが行う事業の一つとして運営されていたが，2001年4月にセンターが改組されてからは，国立遺伝学研究所の公式な組織として活動を行っている。DDBJの沿革と国際協力への取組みを**図**-1に示す。

生命科学の目覚しい発展の基盤として，DNA塩基配列から得られる知識は欠かすことのできないものとなっている。日本においてもDNA塩基配列データを収集・

1980.8	EMBLデータライブラリー（欧州）設立，日本へ国際協力の要請
1982.8	EMBL，GenBank（米国）が国際協力事業 日本の参加を要請
1983.8	DNAデータバンク運営委員会設置
1984.4	大学共同利用機関に改組した国立遺伝学研究所に遺伝情報研究センター設置
1985.4	遺伝情報研究センターに遺伝情報分析研究室設置
1986.1	DNAデータ研究利用委員会設置
1987.1	遺伝情報研究センター棟竣工
7	DDBJリリースの配布開始
9	DDBJオンライン利用開始
1992.1	DDBJリリースにEMBL/GenBankデータを加える
1993.1	DDBJリリースを年4回配布へ
1994.10	遺伝情報研究センターに遺伝子機能研究室新設
1995.4	遺伝情報研究センターに大量遺伝情報研究室・分子分類研究室新設 新設の2研究室と遺伝情報分析研究室・遺伝子機能研究室からなる生命情報研究センター設置
1998.3	生命情報研究センター棟竣工
1999.4	国立遺伝学研究所にて国際実務者会議・国際諮問委員会開催
2001.4	生命情報研究センターの名称が生命情報・DDBJ研究センターに改称

図-1　DDBJの沿革と国際協力
Fig.1-History of DDBJ, and international cooperation.

評価・提供するデータバンク活動で国際的に貢献するため，1983年に試験的なデータ入力が始まった。翌年国立遺伝学研究所に遺伝情報研究センターが設置されたのに伴い，その中でDDBJが活動を始めた。現在はCIB-DDBJ体制の下，前述のような活動を行っている。DDBJは「DDBJ/EMBL/GenBank国際塩基配列データバンク」の一つとして欧州・米国のデータバンクと密接な連携の下に「DDBJデータベース」の構築を進めている。三大国際DNAデータバンクによる国際協調の結果，日本の研究者はデータの登録をDDBJを通じて行うことができるようになっている。DDBJへのデータの登録のほとんどは，インターネット上でWebを利用し，DDBJのデータ登録システム“SAKURA”を利用して行われている。データは生物学の専門知識を持つDDBJ要員の査定を受け「DDBJデータベース」に登録され，公開されている。また，公開と同時にEMBL，GenBankへ送られる。DDBJ，EMBL，GenBankの間で毎日データを交換することにより，最大1日のずれはあるが3者で同時に，質・量ともに同じデータを提供している。

「DDBJ/EMBL/GenBank国際塩基配列データベース」（**図**-2）を運用する3者の実務者が年1回一堂に会して，データベース上の問題点，情報項目の追加，削除についての会議を開催している。

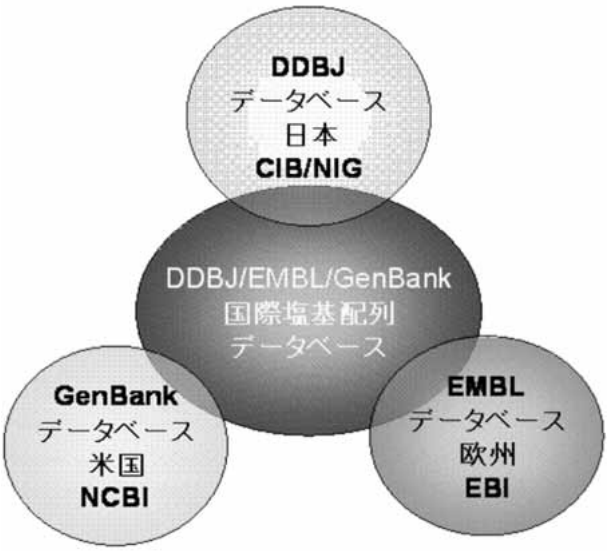

図-2　DDBJ/EMBL/GenBank国際塩基配列データベース
Fig.2-DDBJ/EMBL/GenBank International Nucleotide Sequence Database.

国際塩基配列データベース

「DDBJ/EMBL/GenBank国際塩基配列データベース」とは，全世界の研究者が実験によって決定したDNA（またはRNA）の塩基配列データを「DDBJ/EMBL/GenBank国際DNAデータバンク」のそれぞれが定めたデータ構築規範に沿って収集・編集し，それを3者間で交換し合うことで国際的な共同構築が行われているデータベースである。DDBJ/EMBL/GenBank国際塩基配列データベースは，研究者から直接送付されたDNA塩基配列データと配列付加情報（アノテーション情報）から構成され，DNAデータベースは，データの単位である「エントリ」の集合として構成されている。それぞれのエントリは，塩基配列のほか，配列を決定した研究者，関連文献，生物種，遺伝子の機能・特性に関する情報（配列付加情報）を含んでいる。また，日本の特許庁，欧州の欧州特許局，米国の米国特許局が処理した特許DNAデータも含んでいる。塩基配列データは，久遠の時間をかけて生物が進化してきたことを直接示す記録であり，これらが人類共通の財産であるという認識のもとに，各データバンクでは，研究者が自由に利用できるようにオンラインでデータを公開している。

現在全世界で登録されている配列情報のエントリは1,700万件を超え，塩基対としては200億を超えている。ここ数年，年1.5～2.0倍のスピードで登録データが増えており，塩基配列情報の爆発を危惧する声も出始めている。データベース収集件数の推移を**図**-3に示す。

ゲノムデータとDDBJ

米欧日の公的研究機関が共同で進めていた「ヒトゲノ

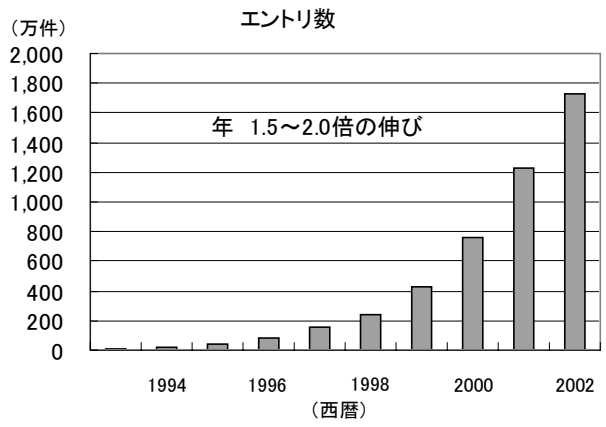

図-3　国際塩基配列データベース件数の推移
Fig.3-International Nucleotide Sequence Database Growth.

ム計画」をはじめ，国内では理化学研究所のマウスゲノムプロジェクト，かずさDNA研究所のアラビドプシスゲノムプロジェクトなどのゲノムプロジェクトチームが決定したデータについてもDBJ/EMBL/GenBank国際塩基配列データベースに登録されている。このうち国内のゲノムプロジェクトデータはDDBJに登録され，世界の研究者に向けて公開している。このほかにショウジョウバエ，線虫，酵母，ウィルスや様々なバクテリア，古細菌など次々とゲノムデータが公開されている。登録塩基数の多い生物種上位15種について**図**-4に紹介する。

DDBJが提供しているサービス

DDBJはインターネットを通して様々なサービスを提供している。日本DNAデータバンク（DDBJ）のホームページ（top）を**図**-5に示す。

DDBJで提供しているサービスの概要について紹介する。

(1)　塩基配列の登録

DNA-データ登録システム“ SAKURA ”を通して研究者は塩基配列データおよび配列付加情報（アノテーション情報）を登録する。一度に大量のエントリを登録する場合は別途大量登録システムが用意されている。

登録されたデータには世界でユニークな登録番号（アクセッション番号）が付与され，公開される。

(2)　データベース検索サービス

登録されたエントリ情報を検索するシステムや遺伝子情報独特の相同性を考慮した塩基配列，アミノ酸配列の多重整列に利用される検索システム（相同性検索システム）としてFASTA，BLAST，SSEARCHのサービスを提供しており，月間10万件以上の検索が実行されている。また，配列付加情報を検索するサービスとしてSRS（Sequence Retrieval System）がある。DDBJでも最近のXMLによる標準化を目指したデータ形式としてDDBJ-XMLのサービスも開始している。

(3)　データ解析サービス

検索した配列データを用いて配列の多重整列や生物の進化過程を解析する系統樹の作成ができる機能をサービスしている。

大規模な解析をサービスするためスーパーコンピュータVPP5000やPRIMEPOWER2000が導入され，強力な計算パワーを提供している。DDBJで稼働しているシステム構成図を**図**-6に示す。

(4)　ゲノム解析サービス

ヒトゲノムの全ゲノムが公開され，それぞれの染色体ごとに各遺伝子の位置やそれぞれの機能が明らかになってきている。また，微生物をはじめ多数の生物に関する全遺伝子配列が決定され公開されてきている。これらの情報を常にウォッチし，公開された情報を収集蓄積したデータベースとしてGIB（Genome Information Broker）のサービスを提供している（**図**-7）。

(5)　蛋白質のDBおよび構造解析サービス

PDB（蛋白質データベース）の検索サービスや蛋白質の二次構造予測，立体構造予測などのシミュレーションについてもDDBJのサービス機能として提供している。

Homo sapiens （ヒト）
Mus musculus （マウス）
Rattus norvegicus （ラット）
Drosophila melanogaster （キイロショウジョウバエ）
Arabidopsis thaliana （シロイロナズナ）
Oryza sativa (japonica cultivar-group)（イネ）
Caenorhabditis elegans （エレバンス線虫）
Oryza sativa （イネ）
Brassica oleracea （キャベツ）
Tetraodon nigroviridis （ミドリフグ）
Pan troglodytes （チンパンジー）
Danio rerio （ゼブラフィッシュ）
Zea mays （トウモロコシ）
Bos taurus （ウシ）
Glycine max （ダイズ）

図-4　登録塩基数の多い生物種上位15種
Fig.4-Top 15 organisms according to the total number of nucleotides.

図-5　DDBJのホームページ（Top）[4]
Fig.5-Top page of DDBJ website.

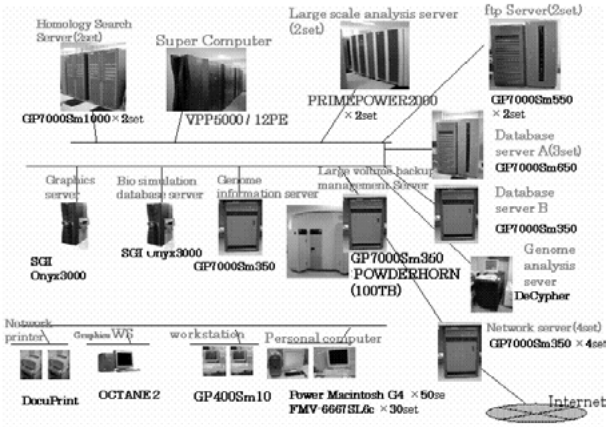

図-6　DDBJのスーパーコンピュータシステム構成
Fig.6-Supercomputer systems of DDBJ.

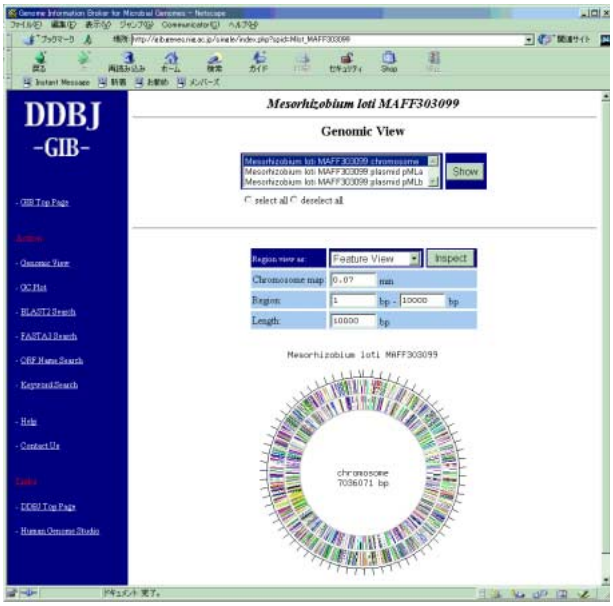

図-7　GIBサービスの一例
Fig.7-A sample of GIB (Genome Information Broker) services.

DDBJ Webサービス

DDBJの検索や解析は研究者がブラウザや電子メールを利用してサービスを提供しており，それぞれのサービスは独立した形で利用できるようになっている。しかし，最近では配列の検索，それを用いた解析，あるいは検索結果からほかの情報の取得要求は多様性を増してきている。そこで，DDBJではXML技術とWebサービス技術を利用したサービス“XML Central of DDBJ”プロジェクトを立ち上げ，サービス機能の充実化に取り組んでいる。XML Central of DDBJのシステムイメージを**図**-8に示す。

このプロジェクトでサービスしている機能を次に示す。

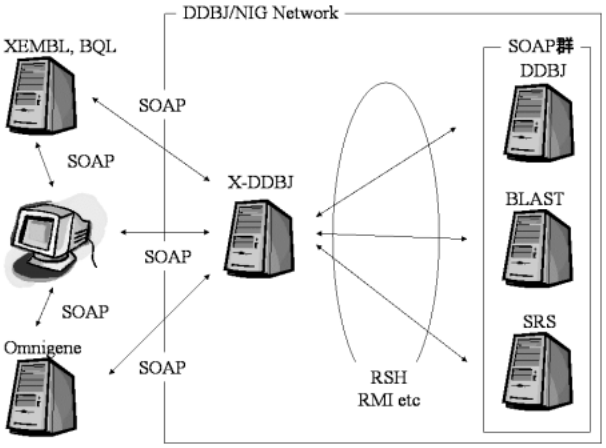

図-8　XML Central of DDBJのシステムイメージ
Fig.8-System image of XML Central of DDBJ.

（1）　Getentry機能

DDBJ，PIR，SwissProt，PDB，PRFからアクセッション番号やProteinIDなどを指定したエントリ取得サービス

（2）　BLAST

BLASTの実行および実行結果からのポジション情報抽出機能のサービス

（3）　FASTA

FASTA実行サービス

（4）　CLUSTALW

CLUSTALW実行サービス

（5）　SRS

20種類以上の公共DBに対するキーワード検索，エントリ取得サービス

（6）　TxSearch

Taxonomy Databaseの検索，詳細情報取得サービス

（7）　DDBJ

DDBJに特化したエントリ取得サービスであり，ロケーション情報から関連Featureや配列情報，配列付加情報の詳細内容の取得をサービス

（8）　サービスの簡易実行機能

DDBJが提供しているWebサービスの機能を試用したり，実行結果の内容などを確認するための簡易実行サービス

利用者はこれらのサービス機能を一連で呼び出すクライアントプログラムを開発することにより，複合的なDDBJのシステムを利用できるようになる。

富士通は，DDBJ設立当初より，これまで紹介してきたDDBJの各種サービスを提供するためのコンピュータ，

ネットワークシステムの構築・運用，サービスソフトウェアの開発を支援してきており，また，DDBJ業務運用についても支援を行うことにより，全世界のバイオ研究者を支援している。

むすび

バイオインフォマティクスの根幹であるゲノム配列の国際登録機関としてDDBJは今後も登録，検索，解析などの各サービスにおいて操作性，機能を充実化していくため，いろいろな取組みを始めている。現在登録できる配列情報の長さには制限が加えられているが，この制限の撤廃，また，これまで登録・修正できるのは登録者のみであったが，第3者が登録情報を修正したり配列付加情報を追加したりする第3者アノテーション（オープンアノテーション）への対応に対してもサービスの開始を予定している。

DDBJが提供するバイオインフォマティクス関連情報は全世界の研究者が利用するため，情報を有効かつ広範囲に流通させるためには標準化が必須であり，DDBJもこの標準化活動に取り組み始めている。われわれ富士通のシステムエンジニアの役割もますます重要度が増してきている。

年2倍の膨大なデータ量伸張への対応，サービスの充実化当に対して，先端技術を先取りし，世界に誇れるDDBJとなるよう支援していきたい。

参 考 文 献

(1)　国立遺伝学研究所　生命情報・DDBJ研究センター「DDBJ　日本DNA　データバンクの紹介」
(2)　国立遺伝学研究所　生命情報・DDBJ研究センター「DDBJ/CIB　レポート」
(3)　国立遺伝学研究所　生命情報・DDBJ研究センター「DDBJオフラインニュース」
(4)　DDBJホームページ
URL　http://www.ddbj.nig.ac.jp